Раздел 1. Задачи оптимизации и байесовские решающие правила.

Лекция 1. Байесовские решающие правила как решение оптимизационной задачи без ограничений.

Будем считать, что объект анализа характеризуется парой $(x, y) \in X \times Y$. Параметр x – наблюдаемый параметр, параметр y – скрытый параметр. Множества X и Y наделены структурами измеримых пространств. Относительно множеств X и Y более ничего не будем предполагать, поскольку широкий спектр приложений не позволяет заранее описать подробнее эти множества. Далее, определяется множество D – множество решений. Множество решений может не совпадать с множеством значений скрытого параметра. Если множество решений совпадает с множеством значений скрытого параметра, то решается задача оценки скрытого параметра по наблюдаемому параметру. Задается отображение множества значений наблюдаемого параметра во множество решений, а именно $l: X \to D$, которое называется детерминированной решающей функцией или байесовским решающим правилом. На декартовом произведении $Y \times D$ задается штрафная функция $W: Y \times D \to R$, значение которой $W(y, d)$ – величина штрафа за принятое решение d при значении скрытого параметра равном y. Например, при вычислении оценки скрытого параметра может быть использована норма, если множество значений скрытого параметра нормированное пространство, то есть $W(y, d) = \|y - d\|$.

Единственным источником информации о ненаблюдаемом параметре является наблюдаемый параметр, который является аргументом решающего правила. Для пары (x, y) локальное качество решающего правила измеряется величиной штрафа $W\big(y, l(x)\big)$. Естественно полагать, что в наблюдаемом параметре содержится информация о ненаблюдаемом параметре. Эту связь между параметрами объекта мы определим с помощью стохастической модели. Прежде всего, рассмотрим вероятностное пространство $\langle \Omega, F, P \rangle$ и будем считать, что отображения $x: \Omega \to X, y: \Omega \to Y$ являются измеримыми отображениями, то есть случайными величинами. Для оценки качества решающего правила определим средний риск следующим образом:

$$R(l) = EW\big(y, l(x)\big). \qquad (1.1)$$

Задача заключается в выборе решающего правила из заданного множества решающих правил L с минимальным средним риском. То есть требуется решить оптимизационную задачу:

$$\min_{l \in L} R(l)$$

Отметим одну особенность формулы (1.1). В этой формуле в результате усреднения отсутствует явная зависимость решающего правила от наблюдаемого параметра. Для того, чтобы проявить эту зависимость, рассмотрим следующий элемент байесовской конструкции.

Условный средний риск. Для определения условного среднего риска используем условное математическое и телескопическое свойство условного математического ожидания. Средний риск с использованием условного математического ожидания вычисляется следующим образом: $R(l) = EE\big(W\big(y, l(x)\big)/F_x\big)$, где F_x – минимальная под σ - алгебра, относительно которой случайная величина x является измеримой.

Назовем условным средним риском величину:

$$r(l, x) = E\big(W\big(y, l(x)\big)/F_x\big). \quad (1.2)$$

Пусть множество допустимых решающих правил совпадает со всеми отображениями множества X в множество D. Из монотонности математического ожидания следует, что оптимальное решающее правило (байесовское решающее правило) выглядит следующим образом:

$$l(x) = arg \min_{d \in D} r(d, x). \quad (1.3)$$

Мы предполагаем достижимость минимума, а также если минимум достигается на нескольких элементах множества, то дополнительно существует механизм разрешения конфликта. Вычисление по формуле (1.3) предполагает, что вначале наблюдается значение случайной величины x, а затем определяется значение решающего правила.

Пусть множества Y и D являются вещественными векторными пространствами со скалярным произведением. Штрафная функция $W(y, d) = (y - d, y - d)$. После несложных выкладок получаем байесовское решающее правило: $l(x) = E(y/F_x)$.

Рассмотрим ряд примеров.

Пример 1. В первой серии примеров положим, что множества X и Y содержат не более чем счетное число элементов. В этом случае вероятностная мера на декартовом произведении $X \times Y$ определяется распределением вероятностей $P(u, v)$, порождаемом случайными величинами x и y, и средний риск $R(l) = \sum_{v \in X} \sum_{u \in Y} W\big(u, l(v)\big)P(v, u)$. Определим на множестве X маргинальное распределение вероятностей $P(v) = \sum_{u \in Y} P(u, v)$. Далее на множестве Y определим условное распределение вероятностей $P(u/v) = \frac{P(u,v)}{P(v)}$. Здесь и далее будем считать, что при равенстве нулю знаменателя, равен нулю числитель и результат деления. Таким образом, частный риск $r(l, x) = \sum_{u \in Y} W\big(u, l(x)\big)P(u/x)$. В результате байесовское решающее правило вычисляется следующим образом:

$$l(x) = arg\min_{d\in D} \sum_{u\in Y} W(u,d)P(u/x) \tag{1.4}$$

Замечание. Легко установить параллель между матричной игрой и байесовской схемой. Игрок A применяет чистую стратегию, а игрок B – смешанную стратегию. Предположим, что множество D содержит не более чем счетное число элементов и предоставим игроку A возможность применять смешанную стратегию $l(x,D)$. Смешанная стратегия $l(x,D)$– это распределение вероятностей на множестве D. Оптимальная смешанная стратегия $l(x,D) = arg\min_{q\in S} \sum_{d\in D} q(d) \sum_{u\in Y} W(u,d)P(u/x)$, $S = \{q: q(d) \geq 0, \sum_{d\in D} q(d) = 1\}$. Определим $\bar{d}(x)$ следующим образом: $\sum_{u\in Y} W(u,\bar{d})P(u/x) \leq \sum_{u\in Y} W(u,d)P(u/x)$. Отсюда непосредственно следует, что $\sum_{u\in Y} W(u,\bar{d})P(u/x) \leq \sum_{d\in D} q(d) \sum_{u\in Y} W(u,d)P(u/x), \forall q \in S$. Следовательно, чистая стратегия $\bar{d}(x)$ – не хуже любой смешанной стратегии.

Продолжим пример. Пусть $D = Y \cup @$, символ $@$ обозначает отказ от принятия решения. Определим штрафную функцию следующим образом:

$W(u,@) = a, a > 0$, для $d \neq @$, $W(u,d) = 1 - \delta(u,d)$, где $\delta(u,d) = \begin{cases} 1, u = d \\ 0, u \neq d \end{cases}$. Подставим штрафную функцию в выражение (1.4) и в результате получим $\sum_{u\in Y} W(u,d)P(u/x) = \begin{cases} a, d = @ \\ 1 - P(d|x), d \neq @ \end{cases}$. Из этого равенства следует, что байесовское решающее правило определяется следующим образом. Сначала, определяется максимально правдоподобное значение $\bar{d} = arg\max_{d\in Y} P(d/x)$, затем вычисляется байесовское решающее правило $l(x) = \begin{cases} @, a < 1 - P(\bar{d}/x) \\ \bar{d}, a \geq 1 - P(\bar{d}/x) \end{cases}$, то есть выбор делается между отказом от распознавания и максимально правдоподобным значением скрытого параметра. В данном примере задача заключается в оценке значения скрытого параметра. Штраф равен нулю, если оценка совпадает со значением скрытого параметра, иначе штраф равен единице. Положим $D = convex(Y) \cup @$, используем квадратичную штрафную функцию: $W(u,d) = \begin{cases} a, d = @ \\ (u-d)^2, d \neq @ \end{cases}$. Мы полагаем, что природа множества Y такова, что выражение $(u-d)^2$ обладает смыслом, например, множество Y – множество натуральных чисел. Решающее правило вычисляется следующим образом. Сначала вычисляется условное математическое ожидание $\bar{d} = \sum_{u\in Y} uP(u/x)$, затем условная дисперсия $\sigma^2(x) = \sum_{u\in Y}(u - \bar{d})^2 P(u/x)$, и решающее правило $l(x) = \begin{cases} @, a < \sigma^2(x) \\ \bar{d}, a \geq \sigma^2(x) \end{cases}$, то есть выбор делается между условным средним и отказом от распознавания. При определении множества D использована выпуклая оболочка $convex(Y)$ множества Y. В первом случае качество решения определяется вероятностью ошибки и выбирается решение, которое минимизирует вероятность ошибки. Во втором случается качество решения измеряется условным средним квадратичным разбросом. Разница

между максимально правдоподобной оценкой и условным средним проявляется в следующей задаче. Пусть $x = (x_i)_{i=1}^n$ – последовательность изображений десятичных цифр, соответствующая ей последовательность $y = (y_i)_{i=1}^n$ – последовательность цифр. По наблюдаемому параметру оценивается сумма $s = \sum_{i=1}^n y_i$. Естественно, использовать множество S из различных значений суммы. Множество $D = S$. Применим штрафную функцию $W(s,d) = 1 - \delta(s,d)$. Для данной штрафной функции решающее правило $l(x) = arg\max\limits_{s\in S} P(s/x)$. Это решающее правило является весьма трудоемким.

Теперь положим $D = convex(S)$ и $W(s,d) = (s-d)^2$. Решающее правило для этой постановки $l(x) = E(s/x) = \sum_{i=1}^n E(y_i/x) = \sum_{i=1}^n l_i(x)$ существенно проще максимального правдоподобия и сводится к вычислению условного математического ожидания для каждого элемента последовательности. Продолжим рассмотрение примеров после некоторой подготовки.

Пример 2. Вторую серию примеров начнем с формулы Байеса. В байесовском решающем правиле используется семейство условных распределений $P(u/v)$ на множестве Y, которое чаще всего неизвестно. Обычно известно семейство условных распределений $P(v/u)$на множестве X и иногда маргинальное распределение $P(u)$ на множестве Y. Формула Байеса позволяет выразить $P(u/v)$ через условное распределение $P(v/u)$ и маргинальное распределение $P(u)$следующим образом: $P(u/v) = \frac{P(v/u)P(u)}{P(v)}$, где $P(v) = \sum_{z\in Y} P(v/z)P(z)$. Подстановка в (1.4) приводит к следующей формуле для байесовского решающего правила:

$$l(x) = arg\min_{d\in D} \frac{1}{P(x)} \sum_{u\in Y} W(u,d)P(x/u)P(u) = \tag{1.5}$$
$$= arg\min_{d\in D} \sum_{u\in Y} W(u,d)P(x/u)P(u).$$

поскольку $P(x) > 0$. Продолжим рассмотрение предыдущего примера. Для максимально правдоподобной оценки необходимо вычислить условное распределение вероятностей $P(s/x) = \sum_{y_1+y_2+\ldots+y_n=s} P(y/x)$. Применим формулу Байеса $P(u/v) = \frac{P(v/u)P(u)}{P(v)}$. Естественно считать, что $P(u/v) = \frac{1}{P(v)} \prod_{i=1}^n P(v_i/u_i)P(u_i)$. Отсюда максимально правдоподобная оценка $l(x) = arg\max\limits_{s\in S} \sum_{u_1+u_2+\ldots+u_n=s} \prod_{i=1}^n P(x_i/u_i)P(u_i)$. Рассмотрим минимальную среднюю квадратичную оценку при тех же предположениях о независимости. Как нетрудно показать, входящие в оценку условные математические ожидания вычисляются следующим образом $E(y_i/x) = \frac{1}{P(x_i)} \sum_{u\in Y} uP(x_i/u)P(u)$. Средняя квадратичная оценка $l(x) = \sum_{i=1}^n \frac{1}{P(x_i)} \sum_{u\in Y} uP(x_i/u)P(u)$.

Часто встречается ситуация, когда множество X – вещественное пространство R^n, множества Yи D – конечные множества. Будем считать, что семейство мер на R^n, порождаемых x, абсолютно непрерывно по отношению к мере Лебега, следовательно, существует семейство условных плотностей $p(v/i)$

для условных законов распределений $Law(x/i)$. Условные плотности следует использовать вместо условных распределений в решающем правиле (1.5):

$$l(x) = arg\min_{j\in D}\sum_{i=1}^{|Y|} W(i,j)p(x/i)P(i). \quad (1.6)$$

Поскольку множество D – конечное множество, то есть смысл рассмотреть разбиение пространства $R^n = \bigcup_{i=1}^{|D|} Q_i, P(Q_i \cap Q_j \neq \emptyset) = 0, i \neq j$, для которого $l(x) = i, \forall x \in Q_i$. Каждое из множеств Q_i определяются системой неравенств:

$$\sum_{r=1}^{|Y|}(a_{i,r} - a_{j,r})p(x/r) \leq 0\ , \quad (1.7)$$

в которой $a_{i,j} = W(j,i)P(j)$. Рассмотрим отображение $\Phi: R^n \to R^{|Y|}, z_i = \Phi_i(x) = p(x/i)$. Каждое из множеств $\bar{Q}_i = \Phi(Q_i)$ определяется системой однородных линейных неравенств:

$$(a_i - a_j, z) \leq 0 \quad (1.8)$$

и является выпуклым конусом. Векторы $a_i = \{a_{i,j}\}$ Очевидно, что $l(x) = i, \forall x: \Phi(x) \in \bar{Q}_i$. Пространство $R^{|Y|}$ называется спрямляющим пространством, а отображение Q – спрямляющим отображением.

Отношение правдоподобия. Интересен случай, когда $|Y| = 2$. Система неравенств (1.8) преобразуется в систему неравенств:

$$(a_{i,1} - a_{j,1})\frac{p(x/1)}{p(x/2)} + (a_{i,2} - a_{j,2}) \leq 0 \quad (1.9)$$

Система (1.9) является линейной системой относительно отношения условных плотностей, которое называется отношением правдоподобия, поэтому множества

$$Q_i = \left\{x: \alpha_i \leq \frac{p(x/1)}{p(x/2)} \leq \beta_i\right\} \quad (1.10)$$

Ситуация еще больше упрощается, если $|D| = 2$. Множество $Q_1 = \left\{x: \frac{p(x/1)}{p(x/2)} \leq \theta\right\}$ и множество $Q_2 = \left\{x: \frac{p(x/1)}{p(x/2)} \geq \theta\right\}$, порог $\theta = -\frac{a_{i,2}-a_{j,2}}{a_{i,1}-a_{j,1}}$. Поскольку логарифм – возрастающая функция, то его можно использовать при определении множества $Q_1 = \left\{x: ln\frac{p(x/1)}{p(x/2)} \leq \bar{\theta}\right\}$ и множества $Q_2 = \left\{x: ln\frac{p(x/1)}{p(x/2)} > \bar{\theta}\right\}$, где порог $\bar{\theta} = ln\,\theta$.

Приведем пример расчета. Рассмотрим два многомерных нормальных закона с условными плотностями $p(x/1) = \frac{1}{\sqrt{(2\pi)^n|C|}} exp\left(-\frac{1}{2}(C^{-1}(x - m_1), x - m_1)\right)$, $p(x/2) = \frac{1}{\sqrt{(2\pi)^n|C|}} exp\left(-\frac{1}{2}(C^{-1}(x - m_2), x - m_2)\right)$. Здесь использовано скалярное произведение $(.,.)$. В выражениях для условных плотностей C – ковариационная матрица, положительно определенная и симметричная, m_1 и m_2 – условные математические ожидания. Вычислим логарифм отношения правдоподобия:

$\ln\frac{p(x/1)}{p(x/2)} = \left(C^{-1}(m_1 - m_2), x\right) + \frac{\left(C^1 m_2, m_2\right) - \left(C^1 m_1, m_1\right)}{2}$. Далее определим множество $Q_1 = \{x: (C^{-1}(m_1 - m_2), x) \le \bar{\theta}\}$ и множество $Q_2 = \{x: (C^{-1}(m_1 - m_2), x) > \bar{\theta}\}$. Порог $\bar{\theta} = \ln\theta - \frac{\left(C^1 m_2, m_2\right) - \left(C^1 m_1, m_1\right)}{2}$. В данном примере возникло **линейное пороговое решающее правило**:

$$l(x) = \begin{cases} 1, (g, x) \ge \bar{\theta} \\ 2, (g, x) \le \bar{\theta} \end{cases}. \quad (1.11)$$

В формуле (1.11) вектор $g = C^{-1}(m_1 - m_2)$. Напомним, что вероятность наступления случайного события $\{(A, x) = \bar{\theta}\}$ равна нулю. Линейное пороговое решающее правило играет особую роль в распознавании образов. Именно с линейного порогового решающего правила начинается распознавание образов. Поэтому далее линейному пороговому решающему правилу будет уделено достаточно внимания.

Задание. Данное решающее правило легко распространяется на случай нескольких классов. Пусть $p(x/j) = \frac{1}{\sqrt{(2\pi)^n |C|}} exp\left(-\frac{1}{2}\left(C^{-1}\left(x - m_j\right), x - m_j\right)\right)$ и $W(i,j) = 1 - \delta_{i,j}$. Доказать, что оптимальное решающее правило будет иметь вид: $l(x) = arg\max_j\left(\left(l_j, x\right) + \theta_j\right)$.

Еще один пример. Сглаживание изображений. В этом примере $X = Y = D = R^n$. Штрафная функция $W(y,d) = (y - d, y - d)$. Как уже отмечалось, байесовское решающее правило $l(x) = E(y/F_x)$. Условный закон распределения $Law(x/y)$–нормальный закон с условной плотностью $p(v/u) = \frac{1}{\sqrt{(2\pi)^n \left(\sigma_1^2\right)^n}} exp\left(-\frac{1}{2\sigma_1^2}(v - u, v - u)\right)$. Ненаблюдаемый параметр y является случайным блужданием: $y_i = m + \sum_{j=1}^{i}\xi_j, \quad \xi_j \in N(0, \sigma_2^2)$. Маргинальное распределение ненаблюдаемого параметра нормальный закон с математическим ожиданием m и ковариоционной матрицей $C_y = \sigma_2^2(min(i,j))$, то есть $p(u) = \frac{1}{\sqrt{(2\pi)^n |C_y|}} exp\left(-\left(C_y^{-1}(u - mI), u - mI\right)\right)$. Применение формулы Байеса позволяет получить плотность условного закона распределения $Law(y/x)$. Поскольку условный закон распределения является нормальным законом распределения, то условное математическое ожидание – это точка, в которой достигается максимум по u произведения $p(v/u)p(u)$. После несложных преобразований получим выражение для решающего правила:

$$l(x) = \left(C_y + \frac{\sigma_1^2}{\sigma_2^2}E\right)^{-1}\left[C_y x - \frac{\sigma_1^2}{\sigma_2^2}mI\right]. \quad (1.12)$$

На этом примере закончим описание байесовских решающих правил. Очевидно одно, байесовская конструкция является универсальным средством принятия решений и ее следует применять в тех случаях, когда достаточно информации для определения всех ее элементов. Наиболее уязвимым элементом

конструкции является маргинальный закон распределения на множестве значений скрытого параметра.

Ниже мы рассмотрим два варианта решения этой проблемы.

Лекция 2. Небайесовские решающие правила.

Минимаксное решающее правило. Линейное программирование

Мы оттолкнемся от решающего правила (1.5) или от решающего правила (1.6). Для определенности остановимся на решающем правиле (1.5), дополнительно будем считать, что множества Yи D содержат конечное число элементов. Итак, за основу возьмем решающее правило:

$$l(x) = arg \min_{j \in D} \sum_{i=1}^{n} W(i,j)P(x/j)P(j). \tag{1.13}$$

Далее мы допустим возможность применения смешанной стратегии, поскольку это необходимо при минимаксной постановке задачи. Мы считаем, что маргинальное распределение $P(j)$ – неизвестно.

Определим минимаксное решающее правило следующим образом:

$$l(x) = arg \min_{q \in S_m} \max_{p \in S_n} \sum_{j=1}^{m} \sum_{i=1}^{n} q_j W(i,j)P(x/i)p_i \tag{1.14}$$

Стратегия (1.14) является осторожной стратегией. Устройство, алгоритм или человек не знает маргинального распределения на множестве Y и рассчитывает на наихудший вариант. В формуле (1.14) нетрудно узнать матричную игру, поэтому далее будет использована теория матричных игр. Матрица проигрышей в этой матричной игре $R(x) = \big(W(i,j)P(x/i)\big)$ зависит от наблюдаемого параметра x. Рассмотрим функцию $F_x(q,p) = (R(x)q,p)$, определенную на декартовом произведении симплексов $S_m \times S_n$. Эта функция имеет седловую точку $\big(\bar{q}(x), \bar{p}(x)\big)$, то есть для этой точки выполняется двойное неравенство: $F_x(\bar{q}(x),p) \le F_x\big(\bar{q}(x),\bar{p}(x)\big) \le F_x\big(q,\bar{p}(x)\big)$. С помощью седловой точки вычисляется байесовское решающее правило

$$l(x) = \bar{q}(x). \tag{1.15}$$

При этом наихудшее маргинальное распределение вероятностей на множестве Y – это второй элемент седловой точки $\bar{p}_i(x)$. Для вычисления седловой точки рассматриваются двойственные задачи линейного программирования.

Прямая задача линейного программирования заключается в следующем:

$$\begin{aligned} &min\, y \\ &(R(x)q)_i \le y, \;\; i = 1,2,\dots,n \\ &\sum_{i=1}^{m} q_i = 1, q_i \ge 0 \end{aligned} \tag{1.16}$$

Допустим, что все элементы матрицы $R(x)$ больше нуля ($minR_{i,j} > 0$), тогда $y > 0$. Выполним замену переменных $q = yz$. Для новых переменных задача (1.16) будет выглядеть следующим образом:

$$max \sum_{i=1}^{m} z_i \tag{1.17}$$
$$(R(x)z)_i \leq 1, \;\; i = 1,2, \dots, n$$
$$z_i \geq 0$$

Множество допустимых решений задачи (1.17) непустое и ограниченное, поэтому задача имеет решение $z^*(x)$. Байесовское решающее правило выражается через это решение следующим образом:

$$l(x) = \frac{1}{\sum_1^m z_i^*(x)} z^*(x) \tag{1.18}$$

Двойственная задача заключается в вычислении

$$min \sum_{i=1}^{n} u_i \tag{1.19}$$

при ограничениях:

$$(R^T(x)u)_i \geq 1, \;\; i = 1,2, \dots, n$$
$$u_i \geq 0.$$

Двойственная задача также имеет решение u^*, через которое выражается наихудшее маргинальное распределение

$$p_i = \frac{1}{\sum_{j=1}^n u_j^*} u_i^*. \tag{1.20}$$

Предположим, что $minR_{i,j}(x) \leq 0$. Рассмотрим матрицу $\bar{R}(x)$с элементами $\bar{R}_{i,j}(x) = R_{i,j}(x) + d$. Причем d выберем таким образом, чтобы $\min\limits_{i,j} \bar{R}_{i,j}(x) > 0$. Рассмотрим задачу:

$$min(y + d)$$
$$(\bar{R}(x)q)_i \leq y + d, \;\; i = 1,2, \dots, n$$
$$\sum_{i=1}^m q_i = 1, q_i \geq 0,$$

которая эквивалентна задаче (1.16), причем элементы матрицы $\bar{R}(x)$ больше нуля.

Рассмотрим пример. Множество $Y = D$ и содержит два элемента. Множество X – вещественное конечномерное векторное пространство. Условные законы распределения определяются условными плотностями $p(v/1)$ и $p(v/2)$. Штрафная функция $W(u,d) = \begin{cases} 0, u = d \\ 1, u \neq d \end{cases}$. Матрица $R(x) = \begin{pmatrix} 0 & p(x/1) \\ p(x/2) & 0 \end{pmatrix}$. Задача (1.17) в этом примере заключается в следующем:

$$\begin{matrix} max(z_1 + z_2) \\ p(x/1)z_2 \leq 1 \\ p(x/2)z_1 \leq 1 \\ z_1 \geq 0, z_2 \geq 0 \end{matrix} \quad (1.19)$$

Решением задачи (1.19) будут $z_1^* = \frac{1}{p(x/2)}, z_2^* = \frac{1}{p(x/1)}$. Отсюда смешанное решение $q^*(x) = \begin{pmatrix} \frac{p(x/1)}{p(x/1)+p(x/2)} \\ \frac{p(x/2)}{p(x/1)+p(x/2)} \end{pmatrix}$. Решение принимается в пользу первого класса с вероятностью $\frac{p(x/1)}{p(x/1)+p(x/2)}$, решение принимается в пользу второго класса с вероятностью $\frac{p(x/2)}{p(x/1)+p(x/2)}$. Наихудшее маргинальное распределение $p^*(x) = \begin{pmatrix} \frac{p(x/2)}{p(x/1)+p(x/2)} \\ \frac{p(x/1)}{p(x/1)+p(x/2)} \end{pmatrix}$.

Полученное минимаксное решение выглядит естественным, хотя и осторожным.

Задание. Рассмотреть случай, когда множество D содержит n – элементов.

Эмпирическое байесовское решение. Нелинейная задача.

В этом разделе мы будем предполагать, что перед принятием решения производится серия независимых наблюдений, в результате которых формируется выборка наблюдений $\{x_1, \dots, x_k\}$. При этом предполагается, что маргинальное распределение вероятностей на конечном множестве Y остается неизменным. Выборка используется для оценки неизвестного маргинального распределения, чтобы использовать эту оценку в решающем правиле (1.5) или в решающем правиле (1.6). Возможны различные варианты вычисления оценки. Здесь будут рассмотрены три оценки.

Для первой оценки следует предположить, что множество $X = \{v_1, v_2, \dots, v_m\}$ – конечное множество и множество $Y = \{y_1, y_2, \dots, y_n\}$- конечное множество. Используя выборку, определим эмпирическую вероятностную меру на множестве X: $P_X(v_i) = \frac{1}{k}\sum_{j=1}^{k} I_{v_i}(x_j)$. Используем приближенное равенство $P_X(v_i) \approx \sum_{j=1}^{n} P(v_i/y_j)p_j$, которое приводит к следующей нелинейной оптимизационной задаче:

$$\begin{matrix} \min\limits_{p} \left\| P_X - P_{X/Y}p \right\| \\ \sum\limits_{i=1}^{n} p_i = 1, p_i \geq 0 \end{matrix} \quad (1.20)$$

В задаче (1.20) использованы обозначения: вектор $P_X = \left(P_X(v_i)\right)$ и матрица $P_{X/Y} = \left(P\left(v_i/y_j\right)\right)$. Задача (1.20) сводится к вычислению вектора с минимальной нормой в выпуклой оболочке, натянутой на множество векторов $g_i = P_X - P_{X/Y}^i$,

где $P^i_{X/Y}$- й столбец матрицы $P_{X/Y}$. Пусть $g^* = \sum_{i=1}^n g_i p_i^*$ - решение задачи. Алгоритм вычисления вектора с минимальной нормой будет рассмотрен позже. Маргинальное распределение определяется равенством

$$P(y_i) = p_i^*. \tag{1.21}$$

Во второй оценке будет использован метод максимального правдоподобия. Более подробно о методе максимального правдоподобия будет рассказано ниже. Множество X может быть конечным, счетным или конечномерным вещественным линейным пространством. В последнем случае мы будем предполагать существование условных плотностей относительно меры Лебега для условных законов распределения. Мы будем использовать одно обозначение для условных законов распределения вероятностей – $p(v/y)$. Для первого и второго случаев – это распределение вероятностей, для третьего – это условная плотность. Множество Y по-прежнему конечное множество. Для выборки вычисляется логарифм функции правдоподобия:

$$F(x_1, x_2, \dots, x_k) = \sum_{i=1}^{k} ln \sum_{j=1}^{n} p_j p\left(x_i/y_j\right). \tag{1.22}$$

Параметры p_jфункции правдоподобия удовлетворяют ограничениям: $p_j \geq 0, \sum_j p_j = 1$. Для получения максимально правдоподобных оценок необходимо вычислить максимум логарифма функции правдоподобия по допустимому набору параметров. Поскольку логарифм функции правдоподобия, как правило, не является вогнутой функцией, то можно говорить лишь о необходимых условиях максимума логарифма правдоподобия. Эти необходимые условия выглядят следующим образом:

$$p_k = \frac{1}{N} \sum_{i=1}^{N} p(y_k/x_i)\,, \tag{1.23}$$

$$p(y_k/x_i) = \frac{p(x_i/y_k)p_k}{\sum_{j=1}^{n} p\left(x_i/u_j\right)p_j}, k = 1,2,\dots,n.$$

Если для решения системы линейных алгебраических уравнений использовать метод Гаусса-Зейделя, то итерации будут выглядеть следующим образом:

$$p^t(y_k/x_i) = \frac{p(x_i/y_k)p_k^{t-1}}{\sum_{j=1}^{n} p\left(x_i/y_j\right)p_j^{t-1}}, \quad p_k^t = \frac{1}{N} \sum_{i=1}^{N} p^t(u_k/x_i)\,, k = 1,2,\dots,n.$$

Допустим p_k^* – решения системы уравнений (1.23). Маргинальное распределение на множестве $YP(u_k) = p_k^*$. Более подробно этот метод вычисления маргинального распределения на множестве Y будет рассмотрен позже в связи с задачей самообучения.

Задача Неймана-Пирсона.

Пусть множество X – вещественное линейное конечномерное пространство. Множество $Y = \{1,2\}$. Известны условные плотности $p(x/1)$ и $p(x/2)$. Маргинальное распределение отсутствует, и провести наблюдения перед

принятием решения невозможно. Например, невозможно произвести предварительное наблюдение в задачах обнаружения чужого объекта, при анализе нестационарных временных рядов, маргинальное распределение изменяется с течением времени, и предварительное наблюдение за временным рядом с целью оценки маргинального распределения приведет к ошибочному результату. Кроме маргинального распределения на множестве Y байесовская модель нуждается в функции штрафа. Функция штрафа оценивает качество принятого решения вещественным числом. Можно легко представить себе задачи, в которых решение не может оцениваться вещественным числом. Например, в семейной жизни бывают ситуации, когда герой стоит перед выбором: потерять ему жену или мать. Естественно, потери от такого выбора вряд ли можно оценить вещественным числом. При отсутствии маргинального распределения можно применить осторожную стратегию, которая была описана ранее. Отсутствие же функции штрафа делает невозможным применение байесовской модели.

В этом разделе мы опишем решающее правило, для которого достаточно знания условных распределений. Рассмотрим разбиение множества $X = D_1 \cup D_2$ и решающее правило $l(x)$, порождаемое этим разбиением:

$l(x) = \begin{cases} 1, x \in D_1 \\ 2, x \in D_2 \end{cases}$. Рассмотрим условные вероятности ошибки данного решающего правила: $P(x \in D_2/1)$ и $P(x \in D_1/2)$. Естественно выбрать такое решающее правило, чтобы эти условные вероятности были как можно меньше. В результате возникает оптимизационная задача с векторной целевой функцией. Существует много приемов для решения оптимизационной задачи с векторной целевой функцией. Один из приемов состоит в следующем. Один из критериев рассматривается как ограничение, а второй критерий рассматривается как скалярная целевая функция. Выбор критерия, который сформирует ограничение, зависит от содержания задачи. Например, в задаче обнаружения чужого объекта условная вероятность $P(x \in D_2/1)$ – это вероятность пропуска чужого объекта, а условная вероятность $P(x \in D_1/2)$– вероятность ложной тревоги. Последствия от пропуска чужого объекта значительно превосходят последствия от ложной тревоги, поэтому естественно ограничить условную вероятность ошибки $P(x \in D_2/1)$ близким к нулю числом, и при этом минимизировать вероятность ложной тревоги. Во многих реальных задачах такой выбор сделать несложно.

Таким образом, вычисление решающего правила связано с решением задачи оптимального выбора одного из множеств разбиения. Допустим, необходимо выбрать множество D_2. Для этого необходимо решить задачу:

$$\begin{matrix} \max\limits_{D_2} P(x \in D_2/2) \\ P(x \in D_2/1) \leq \alpha \end{matrix} \qquad (1.27)$$

Есть возможность найти решение задачи (1.27) при помощи леммы Неймана-Пирсона.

Лемма Неймана-Пирсона. Пусть P и Q – две меры на измеримом пространстве (Ω, F). Пусть мера P – абсолютно непрерывна по отношению к

мере Q. Тогда для фиксированного λ справедливо утверждение. Если $A \in F$ таково, что $Q(A) \le Q(A^\lambda)$, то $P(A) \le P(A^\lambda)$, где $A^\lambda = \left\{\omega \in \Omega: \frac{dP}{dQ} > \lambda\right\}$.

Применим лемму Неймана-Пирсона к решению задачи (1.27). Если положить $X = \Omega$, $P(A) = P(x \in A/2)$ и $Q(A) = P(x \in A/1)$, вычислить такое λ, что $Q(A^\lambda) = \alpha$ (мы предполагаем, что это уравнение имеет решение), то для любого допустимого множества A будет выполняться неравенство: $P(A) \le P(A^\lambda)$. Следовательно, оптимальное множество $D_2 = A^{\lambda^*}$.

Замечание. Если уравнение $Q(A^\lambda) = \alpha$ не имеет решения, то лемма Неймана-Пирсона для решения задачи (1.27) не применима.

Рассмотрим задачу распознавания двух нормальных законов с разными математическими ожиданиями и одинаковыми ковариационными матрицами. Эта задача уже упоминалась ранее. Далее будет продемонстрировано, что задача распознавания нормальных законов относится к многочисленным задачам, для которых лемма Неймана-Пирсона применима.

Схема применения леммы такова. Сначала описывается семейство множеств A^λ, решается уравнение $Q(A^\lambda) = \alpha$. Для описания множества A^λ используем формулу (1.11). Согласно этой формуле, множество $A^\lambda = \{x: (g,x) < \lambda\}, g = C^{-1}(m_1 - m_2)$. Напомним, что матрица C – ковариационная матрица, m_1, m_2 – математические ожидания первого и второго нормальных законов. Для решения уравнения мы воспользуемся тем, что условный закон распределения случайной величины $\xi = (g,x)$ – одномерный нормальный закон распределения с дисперсией $\sigma_1^2 = (Cg,g), \mu_1 = (g,m_1)$. Следовательно, уравнение $Q(A^\lambda) = \alpha$ сводится к уравнению $\Phi\left(\frac{\lambda-\mu_1}{\sigma_1}\right) = \alpha$, в котором $\Phi(.)$ – функция Лапласа. Поскольку функция Лапласа – монотонно возрастающая функция, то данное уравнение имеет единственное решение: $\lambda^* = \mu_1 + \sigma_1\Phi^{-1}(\alpha)$, где $\Phi^{-1}(.)$ – функция обратная к функции Лапласа. Таким образом, в данном примере оптимальным решающим правилом будет **линейное пороговое решающее правило**:

$$l(x) = \begin{cases} 1, (g,x) \ge \lambda^* \\ 2, (g,x) < \lambda^* \end{cases} \qquad (1.28)$$

Особенность решающего правила (1.28) по отношению к решающему правилу (1.11), с которым оно полностью совпадает по своей структуре, заключается в способе вычисления порога. При вычислении порога не используется штрафная функция и маргинальное распределение на множестве Y.

Задача Неймана- Пирсона и задача о рюкзаке.

Пусть множество $X = \{v_1, \dots, v_i, \dots, v_k\}$, множество $Y = \{1,2\}$. Определены два условных закона распределения: $P(x = v_i/1) = p_i^1$ и $P(x = v_i/2) = p_i^2$. Попытаемся применить лемму Неймана-Пирсона. Следуя уже описанной схеме, опишем множества A^λ, для этого определим плотность $\frac{dP}{dQ}(v_i) = \frac{p_i^2}{p_i^1}$. Не нарушая

общности, будем считать, $p_i^1 \neq 0$. Поскольку, если найдется такой элемент v_i из множества X, для которого $P(x = v_i/1) = 0$, то из абсолютной непрерывности следует, что и $P(x = v_i/2) = 0$. Такой элемент без последствий для дальнейшего можно удалить из множества X. Упорядочим элементы множества X по убыванию плотности. Чтобы не вводить дополнительных обозначений, будем считать, что элементы множества X были упорядочены изначально. Является справедливым следующее очевидное утверждение, характеризующее множество A^λ.

Характеристическое свойство множества A^λ.

Если $v_i \in A^\lambda$ и $i \geq 2$, то элементы $v_1, \ldots, v_{i-1}$ также принадлежат множеству A^λ. Используем это свойство в алгоритме, с помощью которого попытаемся решить задачу (1.27).

Алгоритм, использующий утверждение.

begin

if $p_1^1 > \alpha$ then begin $D_2 = \emptyset$; stop end;

add v_1 to D_2 ;

$s = p_1^1 + p_2^1$; $i = 2$;

do while $(s \leq \alpha)$

add v_i to D_2;

$i = i + 1$; $s = s + p_i^1$;

end while;

$s = s - p_i^1$;

if $s < \alpha$ then print ("Algorithm cannot solve problem") else print(D_2);

end.

Если алгоритм, использующий лемму Неймана-Пирсона, не решает задачу, то возможны два варианта поведения. Первый вариант – скорректировать α, а именно, добавить оператор $if(\alpha - s) > (s + p_i^1 - \alpha) then\ begin\ \alpha := s + p_i^1; Add\ v_i to D_2 end\ else\ \alpha := s$.

Второй вариант – применить другую математическую модель.

Рассмотрим второй вариант. Определим переменные $z_i = \begin{cases} 1, v_i \in D_2 \\ 0, v_i \in D_1 \end{cases}$, используя которые вычислим условные вероятности $P(x \in D_2/1) = \sum_{i=1}^k z_i p_i^1$ и $P(x \in D_2/2) = \sum_{i=1}^k z_i p_i^2$. Подставим эти вероятности в (1.27), в результате получим задачу для вычисления множества D_2.

$$max \sum_{i=1}^{k} z_i p_i^2 , \sum_{i=1}^{k} z_i p_i^1 \leq \alpha, z_i \in \{0,1\}. \qquad (1.29)$$

Задача (1.29) известна под названием «Задача о рюкзаке», в связи с одной из первых ее интерпретаций. Для ее решения используются различные алгоритмы, среди которых метод ветвей и границ, который является одним из самых популярных. Часто решение задачи сводится к полному перебору вариантов, число которых равно 2^k. Ясно, что это число может быть очень

большим. Сложность задачи порождается последним ограничением. Заменим его и рассмотрим другую задачу:

$$max \sum_{i=1}^{k} z_i p_i^2 , \sum_{i=1}^{k} z_i p_i^1 \le \alpha, 0 \le z_i \le 1. \quad (1.30)$$

Интерпретация решения задачи (1.29), следующая: $z_i = 1 \leftrightarrow z_i \in D_2$. При интерпретации решения задачи (1.30) множество D_2 удобно рассматривать как нечеткое множество с функцией принадлежности $\mu_{D_2}(v_i) = z_i$, тогда задача (1.30) заключается в том, чтобы вычислить оптимальную функцию принадлежности нечеткого множества D_2. Целевую функцию задачи можно рассматривать как условную вероятность принадлежности к нечеткому множеству $P(x \in D_2/1) = E(\mu_{D_2}/1) = \sum_{i=1}^{k} z_i p_i^2$. Аналогично интерпретируется левая часть второго неравенства.

При помощи функции Лагранжа запишем двойственную задачу:

$$\min_{\lambda \ge 0} \max_{0 \le z_i \le 1} \left[\sum_{i=1}^{k} z_i \left(p_i^2 - \lambda p_i^1\right) + \lambda \alpha \right]. \quad (1.31)$$

Обозначим через $f(\lambda) = \max\limits_{0 \le z_i \le 1} \left[\sum_{i=1}^{k} z_i \left(p_i^2 - \lambda p_i^1\right) + \lambda \alpha\right]$. Функция $f(\lambda)$ – кусочно-линейная выпуклая функция. Определим ее минимум. Для этого вычислим индекс $j = max\left\{l: \sum_{i=1}^{l} p_i^1 \le \alpha\right\}$. Пусть $1 \le j < k$. На интервале $\left(\frac{p_{j+1}^2}{p_{j+1}^1}, \infty\right)$ функция $f(\lambda)$– неубывающая функция. На интервале $\left(0, \frac{p_{j+1}^2}{p_{j+1}^1}\right)$ функция $f(\lambda)$– невозрастающая функция. Следовательно, минимум функции $f(\lambda)$ достигается при $\lambda = \frac{p_{j+1}^2}{p_{j+1}^1}$. Оптимальные значения z:

$$z_i = \begin{cases} 1, i \le j \\ \frac{\alpha - \sum_{i=1}^{j} p_i^1}{p_{j+1}^1}, i = j+1 \\ 0, i > j+1 \end{cases}. \quad (1.32)$$

Рассмотрим крайние ситуации. Пусть $j = k$, тогда $1 \le \alpha$, это противоречит тому, что $\alpha < 1$. Пусть $j = 0$, тогда все $z_i = 0$ и $D_2 = \emptyset$. Из формулы (1.32) следует, чтоv_{j+1} – единственный нечетко принадлежащий множеству Q_2 элемент множества X. Этот элемент можно отнести к множеству D_1, если z_{j+1} близок к нулю (принять решение в пользу первого класса), можно отнести к множеству D_2 (принять решение в пользу второго класса), если z_{j+1} близок к единице, или отказаться от распознавания.

Приведем алгоритм решения задачи (1.30), предполагая, что множество X упорядочено по убыванию отношения $\frac{p_i^2}{p_i^1}$.

begin
for$i = 1$to$k$
$z_i = 0$;
if $p_1^1 > \alpha$ then begin print z; stop end;

$z_1 = 1;$
$s = p_1^1 + p_2^1 \; ; i = 2;$
do while ($s \leq \alpha$).
$z_i = 1;$
$i = i + 1 \; ; s = s + p_i^1;$
end while;
$z_i = \frac{\alpha - s + p_i^1}{p_i^1} \; ;$
print z;
end.

Таким образом, задача (1.30) решена аналитически.

Отметим одно замечательное свойство задачи Неймана-Пирсона. И в первом, и во втором случае решающее правило строилось с использованием отношения правдоподобия.

Далее мы рассмотрим естественные обобщения задачи (1.30).

Задание. На счетном множестве $X = \{0,1, \dots, k, ..\}$ законы распределения вероятностей определяются формулой Пуассона $p(x/1) = \frac{\mu_1^x}{x!} e^{-\mu_1}$, $p(x/2) = \frac{\mu_2^x}{x!} e^{-\mu_1}$. Определить множество $A(\lambda)$. При заданном α найти решение задачи Неймана-Пирсона, точное, если это возможно, или приближенное.

Лекция 3. Продолжение.

Несколько опасных состояний.

Пусть множество $Y = \{u_1, u_2, \dots, u_r\}$ и состояния $u_2, \dots, u_r$ являются опасными состояниями. Множество решений по-прежнему будет содержать два элемента: «опасно» и «неопасно»,$D = \{d_1, d_2\}$. Решающее правило выглядит так же, как и раньше: $l(x) = \begin{cases} d_1, x \in D_1 \\ d_2, x \in D_2 \end{cases}$. Теперь у нас имеется несколько условных вероятностей пропуска целей: $P\left(x \in D_2 / u_j\right), j = 2, \dots, r$, которые необходимо ограничить и минимизировать при этом вероятность ложной тревоги. Поставленной цели соответствует следующая задача линейного программирования:

$$\begin{gathered} max \sum_{i=1}^{k} z_i p_i^1 \\ \sum_{i=1}^{k} z_i p_i^j \le \alpha, j = 2, \dots, r \\ 0 \le z_i \le 1 \end{gathered} \tag{1.33}$$

В задаче (1.33) использованы обозначения $p_i^j = P\left(x = v_i / u_j\right)$. Задача (1.33) имеет решение, поскольку область допустимых решений – ограниченная и замкнутая. К сожалению, переход к двойственной задаче не дает такого замечательного эффекта, как в задаче (1.33).

Одно опасное состояние.

Допустим, что у нас одно опасное состояние – u_1, и ограничить нужно только одну условную вероятность пропуска цели. Условных вероятностей ложных тревог несколько, поэтому требуется выбрать целевую функцию. Один из возможных вариантов выбора заключается в следующем: будем искать минимум максимально возможной условной вероятности ложной тревоги. В результате возникает другая задача линейного программирования:

$$\begin{gathered} max\, w \\ \sum_{i=1}^{k} z_i p_i^1 \le \alpha \\ \sum_{i=1}^{k} z_i p_i^j \ge w, j = 2, \dots, r \\ 0 \le z_i \le 1 \end{gathered} \tag{1.34}$$

Задача (1.34) также имеет решение.

Возможны и другие варианты обобщений задачи Неймана-Пирсона, но мы ограничимся приведенными вариантами.

Задача Вальда как минимаксная задача.

Задача, которую мы сейчас рассмотрим, является одной из многочисленных задач, которые объединяются в последовательный анализ Вальда. В последовательном анализе в каждый момент времени n принимается решение о продолжении наблюдений, решение в пользу первого класса или решение в пользу второго класса. Мы будем использовать следующие обозначения: $X^n = X \times X \ldots \times X$, x^n – случайный вектор со значениями в множестве наблюдений X^n. Множество наблюдений разбивается на три подмножества: $X^n = D_0^n \cup D_1^n \cup D_2^n$. В соответствии с этим разбиением принимаются три решения: отказ от распознавания и продолжение наблюдений, решение в пользу первого класса и решение в пользу второго класса. В задаче Неймана-Пирсона предпочтение отдается первому классу, который считается опасным состоянием, и вероятность пропуска этого опасного состояния ограничивается малой величиной α, вероятность пропуска другого состояния минимизируется и может оказаться довольно большой. Было бы замечательно, если бы удалось построить такое решающее правило, при котором удалось бы ограничить обе вероятности. Однако, такое одновременное ограничение двух вероятностей может привести к противоречию. Поэтому в последовательном анализе предусматривается возможность получения дополнительной информации перед распознаванием.

Приступим к формулировке задачи Вальда. Качество разбиения теперь характеризуется четырьмя условными вероятностями: вероятностью пропуска цели – $P(x^n \in D_2^n/1)$, вероятностью ложной тревоги – $P(x^n \in D_1^n/2)$, и условными вероятностями отказа от распознавания – $P(x \in D_0^n/1)$ и $P(x \in D_0^n/2)$. Первые две вероятности мы ограничим: $P(x^n \in D_2^n/1) \leq \alpha_1$,$P(x^n \in D_1^n/2) \leq \alpha_2$, и потребуем, чтобы система не слишком часто отказывалась от распознавания. Для этого будем искать минимум, например, от $max\{P(x^n \in D_0^n/1), P(x^n \in D_0^n/2)\}$. Теперь множество допустимых решений непустое. Чтобы в этом убедиться, достаточно положить $D_0^n = X^n$. Таким образом, для вычисления оптимального разбиения требуется решить задачу:

$$\begin{aligned} &minmax\{P(x^n \in D_0^n/1), P(x^n \in D_0^n/2)\} \\ &P(x^n \in D_2^n/1) \leq \alpha_1 \\ &P(x^n \in D_1^n/2) \leq \alpha_2 \end{aligned} \qquad (1.35)$$

Для решения задачи (1.35) рассмотрим две вспомогательные задачи:

$$\begin{aligned} &max\, P\,(x^n \in D_2^n/2),\ \ P(x^n \in D_2^n/1) \leq \alpha_1 \ \text{ и} \\ &max\, P\,(x^n \in D_1^n/1), P(x^n \in D_1^n/2) \leq \alpha_2. \end{aligned} \qquad (1.36)$$

Допустим, что для решения каждой из них применима лемма Неймана-Пирсона. Тогда решение первой задачи $D_2^n = A^n(\lambda_n^*)$, где $A^n(\lambda) = \left\{v^n : \frac{dP_2^n}{dP_1^n}(v^n) > \lambda\right\}$, λ_n^*

– решение уравнения $P(x^n \in A^n(\lambda)/1) = \alpha_1$. Решение второй задачи $D_1^n = B^n(\mu_n^*)$, где $B^n(\mu) = \left\{v^n: \frac{dP_2^n}{dP_1^n}(v^n) < \mu\right\}$, μ_n^* – решение уравнения $P(x^n \in B(\mu)/2) = \alpha_2$. Вернемся к исходной задаче (1.35). Допустим, что $\mu_n^* \le \lambda_n^*$, то есть $A^n(\lambda_n^*) \cap B^n(\mu_n^*) = \emptyset$. Для любого допустимого разбиения выполняются очевидные неравенства: $P(D_0^n/1) \ge P(C^n(\lambda_n^*, \mu_n^*)/1)$ и $P(D_0^n/2) \ge P(C^n(\lambda_n^*, \mu_n^*)/2)$, в которых $C^n(\lambda_n^*, \mu_n^*) = X \backslash (A^n(\lambda_n^*) \cup B^n(\mu_n^*))$. Следовательно, разбиение $D_0^n = C^n(\lambda_n^*, \mu_n^*), D_1^n = B^n(\mu_n^*), D_2^n = A^n(\lambda_n^*)$ является решением задачи (1.35). Допустим, что $\mu_n^* > \lambda_n^*$. Выберем порог $\theta \in (\lambda_n^*, \mu_n^*)$ и определим $D_0^n = \emptyset, D_1^n = \left\{v^n: \frac{dP_2^n}{dP_1^n}(v^n) < \theta\right\}, D_2^n = \left\{v^n: \frac{dP_2^n}{dP_1^n}(v^n) \ge \theta\right\}$. Множества D_1^n, D_2^n являются допустимыми множествами в задаче (1.35) а целевая функция для такого разбиения равна нулю. Следовательно, разбиение $D_0^n = \emptyset, D_1^n = \left\{v^n: \frac{dP_2^n}{dP_1^n}(v^n) < \theta\right\}, D_2^n = \left\{v^n: \frac{dP_2^n}{dP_1^n}(v^n) \ge \theta\right\}$ является решением задачи(1.35). Если лемма Неймана-Пирсона не применима, то следует использовать уже описанный ранее прием, использующий нечеткие множества. По понятным причинам мы не будем разбирать эту ситуацию.

Последовательное решающее правило определяется моментом остановки $\tau = min\{n: \neg(x^n \in D_0)\}$, далее $l(x^n) = \begin{cases} 1, x^n \in D_1^n \\ 2, x^n \in D_2^n \end{cases}$.

Задание. Рассмотрим пример, в котором требуется распознать два одномерных нормальных закона с математическими ожиданиями $a_1 < a_2$ и одинаковыми дисперсиями σ^2. Будем считать, что наблюдения – независимые. Определить $A^n(\lambda)$ и $B^n(\mu)$, вычислить λ_n^* и μ_n^*.

Задача о разладке. Последовательный анализ Вальда тесно связан с задачей о разладке. Рассматривается случайная последовательность $x_1, x_2, \dots x_n, \dots$ и случайная величина $y \in \{1,2, \dots\}$ - дискретное случайное время. Для произвольного момента времени n совместный закон распределения представлен равенством:

$$p(x_1, x_2, \dots x_n) = \sum_{i=1}^{n} p(x_1, x_2, \dots x_n/y = i)p(y = i)$$

Для условной вероятности будем использовать следующую формулу:

$$p(x_1, x_2, \dots x_n/y = i) = p_\infty(x_1, \dots, x_{i-1})p_0(x_i, x_{i+1}, \dots, x_n)$$

То есть последовательность $x_1, x_2, \dots x_n$ составляется из двух независимых подпоследовательностей. Кроме этого, $p_\infty(x_1, \dots, x_{i-1}) = \prod_{k=1}^{i-1} p_\infty(x_k)$, $p_0(x_i, x_{i+1}, \dots, x_n) = \prod_{j=i}^{n} p_0(x_j)$. В момент времени n требуется выбрать одно из двух решений $y \le n$ и $y > n$. Матрица штрафа $W = \begin{pmatrix} 0 & \alpha \\ \beta & 0 \end{pmatrix}$, α-штраф за ложную тревогу, β- штраф за пропуск цели . Оптимальное решение $d = \begin{cases} y \le n, \frac{P(y \le n/x_1, \dots x_n)}{P(y > n/x_1, \dots x_n)} > \frac{\alpha}{\beta} \\ y > n, \frac{P(y \le n/x_1, \dots x_n)}{P(y > n/x_1, \dots x_n)} \le \frac{\alpha}{\beta} \end{cases}$.

Отношение правдоподобия $\psi_n(x_1,\dots,x_n)=\frac{P(y\le n/x_1,\dots x_n)}{P(y>n/x_1,\dots x_n)}=\frac{1}{P(y>n)}\sum_{k=1}^{n}P(y=k)\prod_{j=k}^{n}\frac{p_0(x_j)}{p_\infty(x_j)}$.

Рекуррентное уравнение. Запишем $\psi_n(x_1,\dots,x_n)$ следующим образом $\psi_n(x_1,\dots,x_n)=\frac{1}{P(y>n)}\left(\sum_{k=1}^{n-1}P(y=k)\prod_{j=k}^{n}\frac{p_0(x_j)}{p_\infty(x_j)}+P(y=n)\frac{p_0(x_n)}{p_\infty(x_n)}\right)$. Далее $\psi_n(x_1,\dots,x_n)=\frac{1}{P(y>n)}\frac{p_0(x_n)}{p_\infty(x_n)}\left(\sum_{k=1}^{n-1}P(y=k)\prod_{j=k}^{n-1}\frac{p_0(x_j)}{p_\infty(x_j)}+P(y=n)\right)$ и окончательно $\psi_n(x_1,\dots,x_n)=\frac{1}{P(y>n)}\frac{p_0(x_n)}{p_\infty(x_n)}\big(P(y>n-1)\psi_{n-1}(x_1,\dots,x_{n-1})+P(y=n)\big)$. Для $n=1$ справедливо равенство: $\psi_1(x_1)=\frac{1}{P(y>1)}P(y=1)\frac{p_0(x_1)}{p_\infty(x_1)}$. Далее требуется определить оптимальную остановку для принятия решения о наступлении разладки. Оптимальная остановка $\tau=min\left\{n:\psi_n(x_1,\dots,x_n)>\frac{\alpha}{\beta}\right\}$.

Задание. Рассмотреть разностное уравнение $y_n=y_{n-1}\big(m_1I_{\{n<\theta\}}+m_2I_{\{n\ge\theta\}}+\sigma\varepsilon_n\big)$, в котором последовательность ε – последовательность независимых стандартных случайных величин. Вывести рекуррентную формулу для отношения правдоподобия.

Динамическое программирование в распознавании скрытых марковских цепочек.

Начнем с постановки задачи. Рассматривается прежняя задача вычисления ненаблюдаемой последовательности $y=\{y_0,\dots,y_n\}$ по наблюдаемой последовательности $x=\{x_1,\dots,x_n\}$. Связь между x – м и y – м определяется совместным законом распределения $p(x,y)=p(y_0)\prod_{i=1}^{n}p(x_i/y_i)p(y_i/y_{i-1})$. Максимально правдоподобная оценка — это решение следующей оптимизационной задачи: $\max\limits_{y}p(x,y)=\max\limits_{y}p(y_0)\prod_{i=1}^{n}p(x_i/y_i)p(y_i/y_{i-1})$. При фиксированном x . Вместо этой задачи мы будем решать задачу $\max\limits_{y}\{lnp(y_0)+\sum_{i=1}^{n}\big(lnp(x_i/y_i)+lnp(y_i/y_{i-1})\big)\}$, которая эквивалентна исходной. Структура задачи позволяет применить для ее решения динамическое программирование Р. Беллмана.

Динамическое программирование состоит из нескольких действий.

1. Действие первое. Определение семейства функций Беллмана.

Для рассматриваемой задачи это семейство имеет следующий вид:

$$V_k(u)=\max_{y_k^n}\sum_{i=k}^{n}\big(lnp(x_i/y_i)+lnp(y_i/y_{i-1})\big),\ y_{k-1}=u;\ k=1,2,\dots,n;$$
$$V_0=\max_{y}\{lnp(y_0)+\sum_{i=1}^{n}\big(lnp(x_i/y_i)+lnp(y_i/y_{i-1})\big)\}. \qquad (1.37)$$

Очевидно, что функция V_0-решение задачи.

2. Действие второе вывод уравнения Беллмана.

Представим функцию Беллмана $V_k(u)$ следующим образом $V_k(u) = \max\limits_{y_k^n} \sum_{i=k}^{n}\left(lnp(x_i/y_i) + lnp(y_i/y_{i-1})\right) = \max\limits_{y_k}\left[lnp(x_k/y_k) + lnp(y_k/u) + \max\limits_{y_{k+1}^n} \sum_{i=k+1}^{n}\left(lnp(x_i/y_i) + lnp(y_i/y_{i-1})\right)\right] = \max\limits_{y_k}[lnp(x_k/y_k) + lnp(y_k/u) + V_{k+1}(y_k)]$. Таким образом, уравнения Беллмана выглядит следующим образом:

$$V_k(u) = \max_{y_k}[lnp(x_k/y_k) + lnp(y_k/u) + V_{k+1}(y_k)],$$
$$W_k(u) = \operatorname{argmax}_{y_k}[lnp(x_k/y_k) + lnp(y_k/u) + V_{k+1}(y_k)]. \quad (1.37)$$

Краевым значением для этих уравнений является функция $V_n(u) = \max\limits_{y_n}[lnp(x_n/y_n) + lnp(y_n/u)]$. Уравнения Беллмана позволяют свести $n-$ мерную задачу к последовательности одномерных задач.

3. Решение основной задачи. Напомним, что задача распознавания заключалась в вычислении оценки последовательности y . С помощью последовательности уравнений Беллмана вычисляется функция $V_1(u)$, которая позволяет вычислить функцию $V_0 = \max\limits_{y_0}[lnp(y_0) + V_1(y_0)]$, $y_0^* = \operatorname{argmax}\limits_{y_0}[lnp(y_0) + V_1(y_0)]$.

Последовательность функций $W_k(u)$ позволяет вычислить оценки остальных членов ненаблюдаемой последовательности с помощью формулы: $y_k^* = W_k(y_{k-1}^*)$.

Пример. Вернемся к задаче сглаживания. В этой задаче $p(x_k/\,y_k) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x_k-y_k)^2}{2\sigma_1^2}}$, $p(x_k/\,y_k) = \frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y_k-y_{k-1})^2}{2\sigma_2^2}}$, $p(y_0) = \frac{1}{\sqrt{2\pi\sigma_3^2}} e^{-\frac{(y_0-m)^2}{2\sigma_3^2}}$.

Подставляем заданные плотности в исходную задачу: $\max\limits_{y}\left\{-\frac{1}{2}\left(ln2\pi\sigma_3^2 + \frac{(y_0-m)^2}{\sigma_3^2} + \sum_{i=1}^{n}\left(ln2\pi\sigma_1^2 + \frac{(x_i-y_i)^2}{\sigma_1^2} + ln2\pi\sigma_2^2 + \frac{(y_i-y_{i-1})^2}{\sigma_2^2}\right)\right)\right\}$. Эта задача эквивалентна задаче: $\min\limits_{y}\left\{\frac{(y_0-m)^2}{\sigma_3^2} + \sum_{i=1}^{n}\left(\frac{(x_i-y_i)^2}{\sigma_1^2} + \frac{(y_i-y_{i-1})^2}{\sigma_2^2}\right)\right\}$. Вычислим краевые функции $V_n(u) = \min\limits_{y_n}\left[\frac{(x_n-y_n)^2}{\sigma_1^2} + \frac{(y_n-u)^2}{\sigma_2^2}\right] = \frac{(x_n-u)^2}{\sigma_1^2+\sigma_2^2}$ и $W_n(u) = \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2}x_n + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2}u$. Далее предположим, что $V_k(u) = A_k u^2 + 2B_k u + C_k$, причем $A_k > 0$ и найдем рекуррентные уравнения. Для этого воспользуемся уравнением: $V_k(u) = \min\limits_{y_k}\left[\frac{(x_k-y_k)^2}{\sigma_1^2} + \frac{(y_k-u)^2}{\sigma_2^2} + A_{k+1}y_k^2 - 2B_{k+1}y_k + C_{k+1}\right]$. Несложные преобразования приводят к следующим равенствам: $A_k = \frac{A_{k+1}\sigma_1^2+1}{A_{k+1}\sigma_1^2\sigma_2^2+\sigma_1^2+\sigma_2^2}$, $B_k =$

$\frac{B_{k+1}\sigma_1^2+x_k}{A_{k+1}\sigma_1^2\sigma_2^2+\sigma_1^2+\sigma_2^2}$, $C_k = C_{k+1} + x_k^2\sigma_2^2$, $W_k(u) = \frac{B_{k+1}\sigma_1^2\sigma_2^2+x_k\sigma_2^2+u\sigma_1^2}{A_{k+1}\sigma_1^2\sigma_2^2+\sigma_1^2+\sigma_2^2}$. Очевидно следствие: $A_{k+1} > 0 \to A_k > 0$.

На рисунке представлена зашумленная последовательность x (непрерывная траектория) и сглаженная последовательность y (траектория из точек)

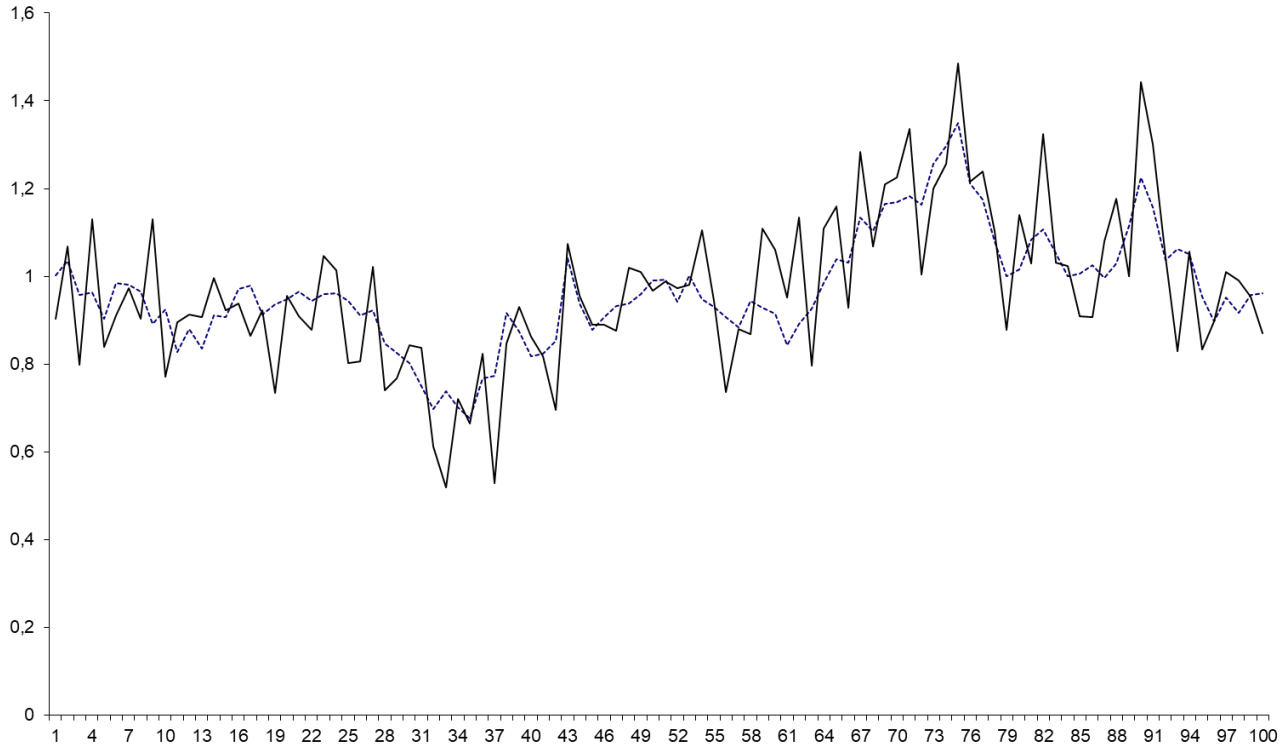


Рис. 1. Сглаживание

На оси абсцисс отложено дискретное время, на оси ординат значения.

Задание. Рассмотрим распознавание бинарных последовательностей, то есть $y_i \in \{0,1\}$. Условные распределения вероятностей имеют вид: $p(x_i/y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x_i-y_i)^2}{2\sigma^2}}$, $p(y_i/y_{i-1}) = p^{y_i y_{i-1}}(1-p)^{(1-y_i)y_{i-1}} q^{y_i(1-y_{i-1})}(1-q)^{(1-y_i)(1-y_{i-1})}$, начальное значение $y_0 = a$. Предложить алгоритм вычисления последовательности y.

1.1. Исторический обзор

Приведенная модель названа байесовской моделью в честь монаха Томаса Байеса, который в начале 18-го века предложил формулу для оперирования условными частотами. Литература по математической статистике насыщена байесовскими методами принятия решений и байесовскими оценками. Можно порекомендовать монографию М. Де Грота [1], которая полностью посвящена байесовскому подходу к принятию решений. В литературе по распознаванию

образов также много внимания уделяется байесовской модели принятия решений. Из наиболее ранних работ – это работы В.А. Ковалевского [2] и Чау [3]. Следует упомянуть более современную монографию Дуды и Харта [4] и, конечно, работу М.И. Шлезингера и В. Главача [5], в которой очень подробно с разных позиций анализируются байесовские решающие правила.

В нашем понимании к небайесовским решающим правилам приводит, прежде всего, отсутствие штрафных функций, а не отсутствие маргинального или априорного распределения на множестве значений оцениваемого параметра. Поскольку всегда есть возможность наделить произвольное множество вероятностной структурой. В этой связи можно порекомендовать монографию [1]. Основу составляет лемма Неймана-Пирсона в обычном и нечетком представлениях. Знаменитая лемма появилась впервые в работе ее авторов [6], а ее рандомизированный или нечеткий вариант в работе [7]. Представленная задача Вальда и ее решение является небольшим фрагментом последовательного анализа, изложение которого частично представлено в уже упомянутой монографии [1] и более подробно в монографии [8]. Задача Неймана-Пирсона и последовательный анализ Вальда не так уж часто появлялись в распознавальческой литературе. Приятным исключением является уже упомянутая здесь работа [5], в которой и задача Неймана-Пирсона и задача Вальда рассматриваются как задачи линейного программирования и для конечного случая дается исчерпывающий ответ, что делать при отсутствии штрафной функции.

Оказывается, нужно по-прежнему вычислять отношение правдоподобия и сравнивать его либо с одним порогом, либо с двумя порогами. Но чтобы прийти к такому выводу, потребовались определенные математические усилия.

Применение байесовской процедуры при анализе изображений появилось в работе [9].

Часть вторая. Обучение и оптимизация.

Лекция 4. Задача обучения как задача оценки параметра по выборке

Задача обучения возникает тогда, когда недостаточно сведений для построения решающего правила. Прежде всего, это относится к условным законам распределения на множестве наблюдений X, которые в данном разделе мы будем считать абсолютно непрерывными по отношению к мере Лебега, тем самым мы будем предполагать, что множество $X = R^d$. Далее мы будем считать, что множество значений ненаблюдаемого параметра Y – конечно, и условные плотности зависят от набора параметров $a_1, a_2, \dots, a_{|Y|}$. Мы предварительно не будем заострять внимание на природе параметров, поскольку характер параметров зависит от конкретной задачи распознавания. Итак, имеется семейство плотностей распределений $p(v/a_k)$, $k = 1,2,\dots,|Y|$. Требуется найти оценки неизвестных параметров $a_1, a_2, \dots, a_{|Y|}$.

В начале мы рассмотрим класс задач, которые называются задачами обучения с учителем. Для того, чтобы найти оценки, производятся наблюдения за случайным вектором x. В результате возникает выборка наблюдений $V = (x_1, \dots, x_{|V|})$, относительно которой делается стандартное предположение о независимости наблюдений. Задача учителя заключается в разбиении выборки на подмножества, и он эту задачу решает либо уверенно, либо с определенными сомнениями.

Сначала мы остановимся на уверенном в себе учителе. В результате его деятельности мы получаем разбиение выборки на подвыборки: $V = \bigcup_{i=1}^{|Y|} V_i$, далее для каждой подвыборки решается собственная задача обучения. Каждая из подвыборок используется независимо для вычисления оценки своего параметра. Независимость обучения – очень приятное свойство рассматриваемого метода обучения.

Метод максимального правдоподобия. Обучение с учителем.

Метод максимального правдоподобия – один из методов параметрической статистики, обладающий рядом весьма ценных свойств. Мы не будем останавливаться на обсуждении этих свойств, поскольку наша задача заключается в другом. Далее мы опустим индексы, помечающие параметры и соответствующие им выборки и рассмотрим задачу обучения для всех классов одновременно. Мы будем использовать обозначение V для всех подвыборок и a для всех параметров. Напомним, что функцией правдоподобия называют плотность совместного распределения элементов выборки, если в качестве аргументов плотности используются элементы выборки. Независимость наблюдений приводит к равенству: $p(x_1, \dots, x_{|V|}/a) = \prod_{i=1}^{|Y|} p(x_i/a)$. Поэтому

очень часто вместо функции правдоподобия, используя монотонно возрастающую функцию логарифм и применяют логарифм функции правдоподобия. Метод заключается в вычислении максимума функции правдоподобия или логарифма функции правдоподобия по оцениваемому параметру:

$$\max_a \prod_{i=1}^{|Y|} p(x_i/a) \text{ или } \max_a \sum_{i=1}^{|V|} ln\, p\,(x_i/a). \quad (2.1)$$

То значение параметра, на котором достигается максимум функции правдоподобия, называется максимально правдоподобной оценкой. Метод максимального правдоподобия позволяет учесть естественные ограничения на значение оцениваемого параметра. В этом случае возникает задача оптимизации с ограничениями:

$$\max_{a\in A} \prod_{i=1}^{|Y|} p(x_i/a) \text{ или } \max_{a\in A} \sum_{i=1}^{|V|} ln\, p\,(x_i/a). \quad (2.2)$$

Оценка параметров многомерного нормального закона.

Мы рассмотрим базовый пример, который встречается во многих задачах распознавания образов. Плотность многомерного нормального закона распределения полностью определяется вектором средних $m = Ex$ и ковариационной матрицей $C = E(x-m)(x-m)^T$. Таким образом, параметр $a = \{m, C\}$. Ковариационная матрица – симметричная и неотрицательно определенная матрица. То, что она – симметричная, следует непосредственно из определения. Неотрицательную определенность легко проверить. Рассмотрим квадратичную форму для ненулевого вектора: $G(u) = (Cu, u)$. Подставим выражение для ковариационной матрицы и в результате получим $G(u) = E(u, x-m)^2$. Отсюда для любого ненулевого вектора u квадратичная форма с ковариационной матрицей неотрицательна. Если ковариационная матрица – невырожденная матрица, то она будет положительно определена. Вернемся к плотности многомерного нормального закона распределения:

$$p(v) = \frac{1}{\sqrt{(2\pi)^d |C|}} exp\left(-\frac{1}{2}(C^{-1}(v-m), v-m)\right). \quad (2.3)$$

Для плотности (2.3) удобно использовать логарифм функции правдоподобия. Таким образом, максимально правдоподобная оценка параметра получается в результате решения следующей задачи:

$$\min_{m,C}\left[lndetC + \frac{1}{|V|}\sum_{i=1}^{|V|}(C^{-1}(x_i-m), x_i-m)\right], \quad (2.4)$$

Проще всего найти оценку для среднего. После дифференцирования целевой функции по m получаем, что максимально правдоподобной оценкой для m является выборочное среднее:

$$\bar{x} = \frac{1}{|V|}\sum_{i=1}^{|V|} x_i. \quad (2.5)$$

Задача (2.4) эквивалентна задаче: $\min_A \left[-lndetA + \frac{1}{|V|}\sum_{i=1}^{|V|}(A(x_i - \bar{x}), x_i - \bar{x})\right]$, причем матрица A – симметричная и положительно определенная, $A \in S^{++} \subset R^{d\times d}$.

Легко устанавливается, что множество симметричных и положительно определенных матриц является выпуклым конусом. Для того, чтобы показать, что функция $f(A) = lndetA$ является вогнутой функцией, воспользуемся теоремой.

Теорема. Функция $f(x)$ определенная на выпуклом множестве D – выпуклая функция тогда и только тогда, когда функция $g(t) = f(x + ty), t \in T = \{t|x + ty \in D\}$, $x \in D$, выпуклая функция.

Рассмотрим матрицу $A + tB$, $A \in S^{++}$, B – произвольная симметричная матрица. Функция $g(t) = lndet(A + tB) = lndetA + lndet\left(E + tA^{-\frac{1}{2}}BA^{-\frac{1}{2}}\right)$. Матрица $A^{-\frac{1}{2}}BA^{-\frac{1}{2}}$ – симметричная матрица, поэтому ее собственные числа – вещественные числа - λ_i- вещественные числа. Поэтому функция $g(t) = lndetA + \sum_{i=1}^{d} ln(1 + t\lambda_i)$ – вогнутая функция на множестве $1 + t\lambda_i > 0$. Применяя теорему, делаем вывод, что функция $lndetA$ - вогнутая функция. Второе слагаемое в целевой функции – линейная функция, поэтому целевая функция задачи – выпуклая функция. Вычисляем производную, приравниваем ее к нулю, в результате получаем равенство

$$C = \frac{1}{|V|}\sum_{i=1}^{|V|}(x_i - \bar{x})(x_i - \bar{x})^T. \quad (2.6)$$

Таким образом, максимально правдоподобные оценки для параметров нормального закона – это выборочное среднее и выборочная ковариационная матрица.

Получив по обучающей выборке оценки параметров нормальных законов, описывающих классы, мы можем использовать далее эти законы в байесовском решающем правиле. Для этого необходимо вычислять определители ковариационных матриц и обратные ковариационные матрицы. Если размерность признаков велика, то это вызывает вычислительные трудности, не говоря о сложности решающего правила.

Задача обучения для нормальных законов с ковариационными матрицами специальной структуры. Метод подпространств.

Рассмотрим один специальный случай многомерных нормальных законов, для которых ковариационная матрица имеет специальный вид:

$$C = AA^T + \sigma^2 E. \quad (2.11)$$

Столбцы матрицы A ортогональны, но необязательно нормированы. То есть, $A^T A = \Lambda$, Λ – диагональная матрица с положительными элементами, E – единичная матрица. Пусть число столбцов матрицы равно l, причем l существенно меньше размерности пространства наблюдений. Определим собственные векторы и собственные числа матрицы C. Легко проверить, что собственными векторами матрицы C будут векторы $u_i = a_i, i = 1, \ldots, l, u_{l+i} = b_i, i = 1,2, \ldots, d - l$. В этом выражении a_i – столбцы матрицы A, λ_i– диагональные элементы матрицы Λ, b_i– произвольный набор ортонормированных векторов, принадлежащих подпространству, ортогонально дополняющему подпространство столбцов матрицы A. Собственные числа ковариационной матрицы $\mu_i = \lambda_i + \sigma^2, i = 1, \ldots, l; \mu_{l+i} = \sigma^2, i = 1, \ldots, d - l$. Обратная ковариационная матрица $C^{-1} = \frac{1}{\sigma^2}[E_d - A(\Lambda + \sigma^2 E_l)^{-1} A^T]$, определитель ковариационной матрицы $|C| = (\sigma^2)^{d-l} \prod_{i=1}^{l} (\lambda_i + \sigma^2)$.

Чтобы не повторять проделанные выкладки, приведем только результат. Максимально правдоподобная оценка среднего значения – это выборочное среднее. Чтобы получить оценки остальных параметров, сначала находим l главных собственных векторов выборочной ковариационной матрицы u_i и соответствующие им собственные числа μ_i. Далее вычисляем $\sigma^2 = \frac{1}{d-l}\left(\sum_{i=1}^{d} r_{i,i} - \sum_{i=1}^{l} \mu_i\right)$, где $r_{i.i}$– диагональные элементы выборочной ковариационной матрицы. Теперь мы можем вычислить $\lambda_i = \mu_i - \sigma^2$. Столбцы матрицы A – это собственные векторы, нормированные таким образом, чтобы выполнялось равенство $A^T A = \Lambda$.

Метод максимального правдоподобия. Обучение с сомневающимся учителем.

Имеется обучающая выборка, относительно каждого элемента выборки x_i известен набор неотрицательных чисел $z_{i,j}, j = 1, \ldots, |Y|$, которые можно рассматривать как вероятности принадлежности элемента выборки образам. Если рассматривается такая интерпретация, то $\sum_{j=1}^{|Y|} z_{i,j} = 1$. Для уверенного учителя в этом наборе только одно число равно единице, остальные равны нулю. Сомневающийся учитель не может вынести однозначного суждения о принадлежности классам элементов выборки, поэтому плотность совместного закона распределения элементов выборки для изучаемой модели будет иметь вид: $p(x_1, \ldots, x_{|V|}) = \prod_{i=1}^{|V|} \sum_{j=1}^{|Y|} p_j p(x_i/a_j)$, и $lnp(x_1, \ldots, x_{|V|}) = \sum_{i=1}^{|V|} ln \sum_{j=1}^{|Y|} p_j p(x_i/a_j)$. В выражении присутствуют дополнительные неизвестные параметры p_j, которые необходимо оценить, используя выборку. Далее воспользуемся тем, что $\sum_{j=1}^{|Y|} z_{i,j} = 1$, и представим логарифм функции правдоподобия следующим образом $lnp(x_1, \ldots, x_{|V|}) = \sum_{i=1}^{|V|} \sum_{k=1}^{|Y|} z_{i,k} ln \sum_{j=1}^{|Y|} p_j p(x_i/a_j)$. Применим формулу Байеса $z_{i,k} =$

$\frac{p(x_i/a_k)p_k}{\sum_{j=1}^{|V|} p_j p(x_i/a_j)}$ и в результате получим $lnp(x_1,\dots,x_{|V|}) = \sum_{i=1}^{|V|}\sum_{k=1}^{|V|} z_{i,k}(lnp_k + lnp(x_i/a_k) - lnz_{i,k})$. По переменным p_k функция $lnp(x_1,\dots,x_{|V|})$ строго выпуклая функция, которая достигает максимума по этим переменным при $p_k^* = \frac{1}{|V|}\sum_i^{|V|} z_{i,k}$. Заметим, что эти оценки не зависят от параметров a_j, поэтому максимально правдоподобные оценки параметров a_j получаются в результате решения серии задач оптимизации:

$$\max_{a_j \in A_j} \sum_{i=1}^{|V|} z_{i,j}\, ln\, p\,(x_i/a_j)\,. \tag{2.12}$$

Для многомерных нормальных законов мы получим следующие оценки для средних и ковариационных матриц:

$$\bar{x}_j = \frac{1}{\sum_{i=1}^{|V|} z_{i,j}} \sum_{i=1}^{|V|} z_{i,j} x_i\,, C_j = \frac{1}{\sum_{i=1}^{|V|} z_{i,j}} \sum_{i=1}^{|V|} z_{i,j}(x_i - \bar{x}_j)(x_i - \bar{x}_j)^T\,. \tag{2.13}$$

Если необходимо упростить решающее правило, то для каждой из матриц C_j можно использовать представление (2.11).

Задача обучения без учителя.

Теперь у нас нет даже сомневающегося учителя. Единственная информация, которой мы располагаем – это выборка, число классов и плотности условных законов распределений, известных с точностью до неопределенных параметров. При предположении о независимости и одинаковой распределенности элементов выборки логарифм их совместной плотности распределения будет иметь вид:

$$ln\, p\,(v_1,\dots,v_{|V|}) = \sum_{i=1}^{|V|} ln \sum_{j=1}^{|Y|} p_j p(v_i/a_j)\,. \tag{2.14}$$

В формуле (2.14) p_j задают распределение вероятностей на множестве значений ненаблюдаемого параметра Y. Таким образом, p_j необходимо включить в множество оцениваемых параметров. Таким образом, задача самообучения заключается в вычислении

$$\max_{\{p_j, a_j\}} F(\{p_j, a_j\}) = \sum_{i=1}^{|V|} ln \sum_{j=1}^{|Y|} p_j p(x_i/a_j). \tag{2.15}$$

Задача (2.15) значительно сложнее задачи обучения. Целевая функция $F(p,a)$, как правило, имеет локальные экстремумы, поэтому локальные алгоритмы поиска не гарантируют сходимости к решению задачи. Приведенный ниже алгоритм обладает одним важным качеством: этот алгоритм является монотонным.

Алгоритм самообучения.

Алгоритм самообучения является итерационным алгоритмом.

На стадии инициализации выбираются начальные значения параметров: p_j^0 и a_j^0.

На итерации с номером t считаем известными значения параметров:p_j^tиa_j^t. Определим величины $\alpha_{i,k}^t = \frac{p(x_i/a_k^t)p_k^t}{\sum_{j=1}^{|Y|} p\left(x_i/a_j^t\right)p_j^t}$, для которых справедливо: $\sum_{k=1}^{|Y|} \alpha_{i,k}^t = 1, \alpha_{i,k}^t \geq 0$. Это обстоятельство позволяет записать целевую функцию в задаче (2.15) в виде:

$$F\left(\{p_j, a_j\}\right) = \sum_{i=1}^{|V|}\sum_{k=1}^{|Y|} \alpha_{i,k}^t \, ln \sum_{j=1}^{|Y|} p_j p\left(x_i/a_j\right) =$$

$$= \sum_{i=1}^{|V|}\sum_{k=1}^{|Y|} \alpha_{i,k}^t \left(ln\, p_k + ln\, p\,(x_i/a_k) - ln \frac{p_k p(x_i/a_k)}{\sum p_j p(x_i/a_j)}\right).$$

Рассмотрим вспомогательную функцию $\bar{F}_t\left(\{p_j, a_j\}, \{\alpha_{i,k}\}\right) = \sum_{i=1}^{|V|}\sum_{k=1}^{|Y|} \alpha_{i,k}\left(ln\, p_k + ln\, p\,(x_i/a_k) - ln\, \alpha_{i,k}\right)$. Очевидно равенство $F\left(\{p_j^t, a_j^t\}\right) = \bar{F}\left(\{p_j^t, a_j^t\}, \{\alpha_{i,k}^t\}\right)$. Вычислим максимум вспомогательной функции $\bar{F}\left(\{p_j, a_j\}, \{\alpha_{i,k}^t\}\right)$ по переменным $\{p_j, a_j\}$. Эта задача совпадает с задачей обучения с сомневающимся учителем. Пусть $\{p_j^{t+1}, a_j^{t+1}\}$ – решение задачи обучения. Поскольку по переменным $\{p_j\}$ вспомогательная функция – строго вогнутая функция, то $F\left(\{p_j^t, a_j^t\}\right) < \bar{F}\left(\{p_j^{t+1}, a_j^{t+1}\}, \{\alpha_{i,k}^t\}\right)$, если $p_j^t \neq p_j^{t+1}$ хотя бы для одного j. Вычислим $\max\limits_{\{\alpha_{i,k}\}} \bar{F}\left(\{p_j^{t+1}, a_j^{t+1}\}, \{\alpha_{i,k}\}\right)$, при ограничениях: $\sum_{k=1}^{|Y|} \alpha_{i,k} = 1, \alpha_{i,k} \geq 0$. Решение этой задачи имеет вид: $\alpha_{i,k}^{t+1} = \frac{p(x_i/a_k^{t+1})p_k^{t+1}}{\sum_{j=1}^{|Y|} p\left(x_i/a_j^{t+1}\right)p_j^{t+1}}$. Таким образом справедлива цепочка неравенств $F\left(\{p_j^t, a_j^t\}\right) \leq \bar{F}\left(\{p_j^{t+1}, a_j^{t+1}\}, \{\alpha_{i,k}^t\}\right) \leq \bar{F}\left(\{p_j^{t+1}, a_j^{t+1}\}, \{\alpha_{i,k}^{t+1}\}\right) = F\left(\{p_j^{t+1}, a_j^{t+1}\}\right)$.

При незначительном изменении целевой функции можно останавливать итерации.

Поскольку на каждой итерации целевая функция не уменьшается, а при выполнении определенного условия увеличивается, то алгоритм является монотонным. Если предположить, что параметрическое семейство плотностей условных законов распределения ограничено сверху, то целевая функция тоже ограничена сверху, и, следовательно, монотонно возрастающая последовательность значений целевой функции будет сходящейся.

Проанализируем поведение последовательности α^t. Теперь мы будем рассматривать алгоритм, как средство генерации последовательности α^t. Напомним, что $\sum_{k=1}^{|Y|} \alpha_{i,k}^t = 1, \alpha_{i,k}^t \geq 0$. При любых i и t набор $\left\{\alpha_{i,k}^t\right\}_{k=1}^{|Y|}$ задает распределение вероятностей на множестве Y.

Установлено, что последовательность значений функции $F(\{p_k^t\},\{a_k^t\}) \uparrow F^*$. Отсюда следует, что $lim\left(F(\{p_k^{t+1}\},\{a_k^{t+1}\}) - F(\{p_k^t\},\{a_k^t\})\right) = 0$. Поскольку $F(\{p_k^{t+1}\},\{a_k^{t+1}\}) - F(\{p_k^t\},\{a_k^t\}) =$

$$= \sum_{k=1}^{|Y|}\sum_{i=1}^{|V|} \alpha_{i,k}^{t+1}\left[(ln\, p_k^{t+1} - ln\, p_k^t) + \left(ln\, p\,(x_i/a_k^{t+1}) - ln\, p\,(x_i/a_k^t)\right)\right] +$$

$+\sum_{i=1}^{|V|}\sum_{k=1}^{|Y|} \alpha_{i,k}^{t+1}\, ln\frac{\alpha_{i,k}^{t+1}}{\alpha_{i,k}^t}$ и каждое из двух слагаемых положительно, то $lim\sum_{i=1}^{|V|}\sum_{k=1}^{|Y|} \alpha_{i,k}^{t+1}\, ln\frac{\alpha_{i,k}^{t+1}}{\alpha_{i,k}^t} = 0$.

Для дальнейшего нам понадобится неравенство Кульбака.

Лемма. Неравенство Кульбака. Пусть $\{x_i\}$ и $\{y_i\}$ – распределение вероятностей на множестве Y, причем распределение $\{y_i\}$ абсолютно непрерывно по отношению к распределению $\{x_i\}$, тогда $\sum_{i=1}^{|Y|} x_i\, ln\frac{x_i}{y_i} \geq \frac{1}{2}\sum_{i=1}^{|Y|}(x_i - y_i)^2$.

Из неравенства Кульбака следует, что евклидова норма $\|\alpha^{t+1} - \alpha^t\| \to 0$. Обратимся к итерациям алгоритма самообучения. Сначала вычисляются элементы $\alpha^t = \{\alpha_{i,k}^t\}$, затем $\{p_k^{t+1}\}$ и $\{a_k^{t+1}\}$, после этого вычисляются элементы $\alpha^{t+1} = \{\alpha_{i,k}^{t+1}\}$. Итерации алгоритма самообучения можно представить как результат действия оператора: $\alpha^{t+1} = Sl(\alpha^t)$. Тогда будет справедливо равенство нулю предела:

$$\|Sl(\alpha^t) - \alpha^t\| \to 0. \qquad (2.19)$$

Очевидно, что последовательность α^t принадлежат ограниченному и замкнутому подмножеству евклидового пространства $R^{|V|\times|Y|}$. Из такой последовательности можно выделить сходящуюся подпоследовательность $\alpha^{t_1},\ldots,\alpha^{t_n},\ldots$ с пределом a. Очевидно, что подпоследовательность обладает свойством (2.19) и поэтому

$$\lim_{j\to\infty} Sl(\alpha^{t_j}) = a. \qquad (2.20)$$

Допустим, что оператор Sl – непрерывный оператор, тогда a является неподвижной точкой этого оператора. То есть, $Sl(a) = a$.

Будем считать, что $A = \{\xi_1\ldots,\xi_{|A|}\}$ – множество неподвижных точек оператора Sl - конечное множество. Рассмотрим последовательность расстояний $d^t = d(\alpha^t, A) = \min_{\xi\in A}\|\alpha^t - \xi\|$. Покажем, что эта последовательность стремится к нулю. Допустим, что это не так. Это означает, что можно подобрать такое значение $\varepsilon > 0$ и такую подпоследовательность α^{t_k}, для которой $d^{t_k} \geq \varepsilon > 0$. Выделим из этой подпоследовательности сходящуюся подпоследовательность. Естественно, что предел этой подпоследовательности не будет неподвижной

точкой, а предыдущие рассуждения привели нас к выводу, что этот предел должен быть неподвижной точкой.

Теперь мы можем доказать, что существует предел последовательности α^t и этот предел – неподвижная точка оператора $Sl(\alpha)$. Выберем $\varepsilon > 0$ и выберем T таким образом, для всех $t \geq T$ одновременно выполнялись неравенства: $\|\alpha^{t+1} - \alpha^t\| < \varepsilon/3$, $d(\alpha^t, A) = \|\alpha^t - a\| < \varepsilon/3$. Пусть $d(\alpha^{t+1}, A) = \|\alpha^{t+1} - b\|$ и $a \neq b$. Из неравенства треугольника следует, что $\|b - a\| \leq \|b - \alpha^{t+1}\| + \|\alpha^{t+1} - \alpha^t\| + \|\alpha^t - a\| < \varepsilon$. Таким образом, начиная с T, ближайшая точка к элементам последовательности α будет сохраняться, то есть, $d(\alpha^t, A) = \|\alpha^t - a\|$ для $t \geq T$. Эта неподвижная точка будет пределом последовательности α.

Вернемся к задаче Робинса, которая упоминалась в разделе «Эмпирический байесовский метод». Задача Робинса заключалась в вычислении $\max\limits_{\{p_k\}} \sum_{i=1}^{|V|} ln \sum_{k=1}^{|Y|} p_k p(x_i/k)$. Алгоритм самообучения, в котором не вычисляются параметры условных распределений $\{a_k\}$, построит последовательность, которая сходится к одной из неподвижных точек функции $Sl(.)$. Отметим, что целевая функция – строго вогнутая функция, поэтому есть все основания считать, что последовательность $\{p_k^t\}$, в которой элементы выражаются через $\{\alpha_{i,k}^t\}$ следующим образом: $p_k^{t+1} = \frac{1}{|V|}\sum_{i=1}^{|V|} \alpha_{i,k}^t$, будет сходиться к $\{p_k^*\}$, причем $p_k^* = \frac{1}{|V|}\sum_{i=1}^{|V|} \xi_{i,k}$, где ξ – неподвижная точка отображения $Sl(.)$. В алгоритме самообучения мы можем исключить промежуточные вычисления элементов последовательности α. Соответствующая формула будет иметь вид:

$$p_k^{t+1} = \frac{1}{|V|}\sum_{i=1}^{|V|} \frac{p_k^t p(x_i/k)}{\sum_{j=1}^{|Y|} p_j^t p(x_i/j)}. \quad (2.21)$$

Перейдя к пределу в левой и правой частях уравнения (2.21), получим равенство:

$$p_k^* = \frac{1}{|V|}\sum_{i=1}^{|V|} \frac{p_k^* p(x_i/k)}{\sum_{j=1}^{|Y|} p_j^* p(x_i/j)}. \quad (2.22)$$

Иными словами, предельная точка алгоритма является решением системы уравнений:

$$z_k = \frac{1}{|V|}\sum_{i=1}^{|V|} \frac{z_k p(x_i/k)}{\sum_{j=1}^{|Y|} z_j p(x_i/j)}, \quad (2.23)$$

которые в свою очередь являются необходимыми и достаточными условиями оптимальности целевой функции $\sum_{i=1}^{|V|} ln \sum_{k=1}^{|Y|} z_k p(x_i/k)$ при ограничениях: $\sum_{k=1}^{|V|} z_k = 1$. Если выбрать начальные значения $p_k^0 > 0$, то из (2.21) следует, что $p_k^t > 0$, следовательно, найденное алгоритмом самообучения решение системы (2.23) будет неотрицательным. Таким образом, алгоритм самообучения решает задачу Робинса.

Задание. Вывести необходимые и достаточные условия для задачи Робинса.

2.2 Библиография

Задача обучения как задача оценки неизвестного параметра, превращает этот раздел распознавания образов в приложение математической статистики. В этой связи можно порекомендовать монографию Лемана [10]. Метод максимального правдоподобия – неединственный метод вычисления оценок. Однако, этот метод обладает многими положительными свойствами, делающими его привлекательным. В распознавальческой литературе этому методу уделяется наибольшее внимание. Достаточно посмотреть монографии по распознаванию образов [4,5,11]. Вклад распознавателей – это обучение с сомневающимся учителем. Задача самообучения как задача кластеризации рассмотрена впервые в работе М. Шлезингера[12]. М.И. Шлезингеру принадлежит и наиболее детальный анализ алгоритма самообучения, представленный в монографии [5]. Наряду со статистическим подходом к обучению и самообучению существует большое количество алгоритмов обучения и самообучения, использующих нестатистические соображения. Многие из этих алгоритмов хорошо проявили себя при решении некоторых практических задач. Нам бы хотелось в качестве удачного алгоритма самообучения привести алгоритмы k -средних и нечетких k - средних. Метод k -средних был изобретен в 1956 году Гуго Штейнгаузом [13], с описанием метода нечетких k -средних можно познакомиться в работе [14]. Отметим, что алгоритм нечетких k -средних близок к обсуждаемому здесь алгоритму.

Статистический метод подпространств был представлен в работе [15].

Часть третья. Линейные решающие правила.

Лекция 5. Методы оптимизации и линейное пороговое решающее правило.

В предыдущих разделах мы сталкивались с линейными дискриминатными функциями, например, при распознавании многомерных нормальных законов с одинаковыми ковариационными матрицами. Кроме этого, многие нелинейные функции могут быть выражены как линейные в спрямляющем пространстве:

$$g(x) = \sum_{j \in J} \alpha_j \varphi_j(x). \quad (3.1)$$

В формуле (3.1) $\{\varphi_j(x)\}$ – фиксированный набор функций. Если рассмотреть отображение, при котором точке $x \in R^n$ ставится в соответствие точка $y \in R^{|J|}$ с координатами $\varphi_j(x)$, то в пространстве $R^{|J|}$ дискриминантная функция (3.1) будет линейной, поэтому пространство $R^{|J|}$ называют спрямляющим пространством. В распознавании образов популярными являются потенциальные функции, которые строятся при помощи убывающей функции одного аргумента, например, $\varphi(z) = exp(-z/2\sigma^2)$ и набора точек $\{u_j\}$в пространстве R^n. Набор потенциальных функций будет иметь вид: $\varphi_j(x) = \varphi\left(\|x - y_j\|^2\right)$. Можно также отметить, что байесовские решающие правила выражаются как линейные дискриминантные функции в пространстве вероятностей.

Дискриминантные функции Андерсона

Остановимся на распознавании двух классов.

Линейная дискриминантная функция для распознавания двух классов будет иметь вид:

$$f(x) = \begin{cases} 1, (l, x) \geq \theta \\ 2, (l, x) < \theta \end{cases}. \quad (3.2)$$

Мы предполагаем, что вектор $l \neq 0$. Если увеличить на единицу размерность пространства, к вектору x добавить $n + 1$ координату равную -1, а вектору l - координату равную θ и не менять обозначений, то линейное решающее правило будет выглядеть следующим образом:

$$f(x) = \begin{cases} 1, (l, x) \geq 0 \\ 2, (l, x) < 0 \end{cases}. \quad (3.3)$$

Таким образом, мы имеем два эквивалентных представления линейной дискриминантной функции.

Задача Андерсона – это задача вычисления вектора l для распознавания двух классов, которые определяются как смеси многомерных нормальных законов. Таким образом, имеется множество нормальных законов J с математическими ожиданиями m_j и ковариационными матрицами C_j. Это множество разбито на два непересекающихся подмножества: $J = J_1 \cup J_2$, каждое из которых определяет свой класс. Пусть x порождается j - м нормальным

законом и $j \in J_1$. Вычислим вероятность того, что x будет распознан неверно, обозначив ее через $\alpha_j(x)$. Вероятность $\alpha_j(x)$ – это условная вероятность $P\big((l,x) < \theta/j\big)$. Условный закон распределения случайной величины (l,x)– одномерный нормальный закон распределения с математическим ожиданием $\mu_j = \big(l, m_j\big)$ и дисперсией $\sigma_j = \big(C_j l, l\big)$. условная вероятность

$$\alpha_j(l,\theta) = P\big((l,x) < \theta/j\big) = \Phi\left(\frac{\theta - (l,m_j)}{\sqrt{(C_j l,l)}}\right), j \in J_1. \tag{3.4}$$

В (3.4) Φ – функция Лапласа. Аналогичная вероятность ошибки для второго класса:

$$\beta_j(l,\theta) = P\big((l,x) \geq \theta/j\big) = 1 - \Phi\left(\frac{\theta - (l,m_j)}{\sqrt{(C_j l,l)}}\right) = \tag{3.5}$$

$$\Phi\left(\frac{(l,m_j) - \theta}{\sqrt{(C_j l,l)}}\right), j \in J_2.$$

Вычислим максимально возможные вероятности ошибки для первого класса: $\alpha(l,\theta) = \max\limits_{j \in J_1} \Phi\left(\frac{\theta - (l,m_j)}{\sqrt{(C_j l,l)}}\right)$, для второго класса: $\beta(l,\theta) = \max\limits_{k \in J_2} \Phi\left(\frac{(l,m_k) - \theta}{\sqrt{(C_k l,l)}}\right)$.

Задача Андерсона имеет следующий вид:

$$\min_{l,\theta} max\big(\alpha(l,\theta), \beta(l,\theta)\big). \tag{3.6}$$

Задача (3.6) является минимаксной оптимизационной задачей.

Рассмотрим подробнее задачу Андерсона. Из возрастания функции Лапласа сразу следует, что задача Андерсона трансформируется в задачу: $\min\limits_{l,\theta} max\left(\max\limits_{j \in J_1} \frac{\theta - (l,m_j)}{\sqrt{(C_j l,l)}}, \max\limits_{k \in J_2} \frac{(l,m_k) - \theta}{\sqrt{(C_k l,l)}}\right)$. Далее мы будем предполагать, что множества $M_1 = \{m_j : j \in J_1\}$ и $M_2 = \{m_j : j \in J_2\}$ линейно разделимы, то есть существуют такие l и θ, что $\big(l, m_j\big) > \theta, j \in J_1$ и $\big(l, m_j\big) \leq \theta, j \in J_2$. Рассмотрим множество значений параметров $П = \{(l,\theta) \in R^{n+1} : \big(l, m_j\big) > \theta, j \in J_1; (l, m_k) < \theta, k \in J_2\}$. Из предыдущего предположения следует, что $П \neq \emptyset$. Нетрудно показать, что множество $П$ является выпуклым конусом. Данная задача эквивалентна следующей задаче:

$$\max_{(l,\theta) \in П} min\left(\min_{j \in J_1} \frac{(l,m_j) - \theta}{\sqrt{(C_j l,l)}}, \min_{k \in J_2} \frac{\theta - (l,m_k)}{\sqrt{(C_k l,l)}}\right). \tag{3.7}$$

Поскольку рассмотрение параметров, не принадлежащих конусу $П$, приводит к неравенству $max\big(\alpha(l,\theta), \beta(l,\theta)\big) \geq 0.5$ и решением задачи Андерсона не

являются. Используем прием, описанный вначале раздела, и сведем задачу (3.7) к следующей задаче:

$$\max_{l\in\Pi}\min_{j\in J_1\cup J_1}\frac{(l,m_j)}{\sqrt{(C_jl,l)}},\Pi=\{l:(l,m_j)>0\}$$

В свою очередь задача (3.7) эквивалентна задаче:

$$\begin{gathered}max\ y\\ y\sqrt{(C_jl,l)}-(l,m_j)\le 0, j\in J_1\cup J_2\\ y\ge 0.\end{gathered}\tag{3.8}$$

Рассмотрим алгоритм решения этой задачи.

Алгоритм

1. Выберем начальное значение y_0 для y и начальное значение h_0 для шага h.
2. На итерации с номером t решаем задачу выпуклого программирования $\min_l \max_j y_t\sqrt{(C_jl,l)}-(l,m_j)$. Выбираем шаг h_{t+1}и вычисляем y_{t+1} в зависимости от значения целевой функции. Если оптимальное значение целевой функции отрицательно, то $h_{t+1}=h_t, y_{t+1}=y_t+h_{t+1}$; иначе $h_{t+1}=h_t/2$, $y_{t+1}=y_{t-1}+h_{t+1}$.
3. Остановка.

После решения последней оптимизационной задачи определяется вектор l.

Задача Андерсона для двух нормальных законов.

Задача Андерсона для распознавания двух нормальных законов будет иметь следующий вид:

$$\max_{l,\theta} min\left(\frac{(l,m_1)-\theta}{\sqrt{(C_1l,l)}},\frac{\theta-(l,m_2)}{\sqrt{(C_2l,l)}}\right).\tag{3.9}$$

Заморозим l. По переменной θ минимизируемая функция является кусочно-линейной и вогнутой, следовательно, при фиксированном l оптимальное значение переменной θ находится из уравнения: $\frac{\theta-(l,m_2)}{\sqrt{(C_2l,l)}}=\frac{(l,m_1)-\theta}{\sqrt{(C_1l,l)}}$решением которого будет:

$$\theta^*=(l,m_1)\frac{\sqrt{(C_2l,l)}}{\sqrt{(C_1l,l)}+\sqrt{(C_2l,l)}}+(l,m_2)\frac{\sqrt{(C_1l,l)}}{\sqrt{(C_1l,l)}+\sqrt{(C_2l,l)}}.\tag{3.10}$$

Подставим оптимальное значение θ^* в минимизируемую функцию, и в результате получим задачу для вычисления вектора l:

$$\max_l\frac{(l,m_1-m_2)}{\sqrt{(C_1l,l)}+\sqrt{(C_2l,l)}},(l,m_1-m_2)\ge 0\tag{3.11}$$

Алгоритм, который решает задачу аналогичен алгоритму, приведенному ранее.

Алгоритм.

1. Выберем начальное значение y_0 для y и начальное значение h_0 для шага h .
2. На итерации с номером t решаем задачу выпуклого программирования $\min_l \; y_t\left(\sqrt{(C_j l, l)} + \sqrt{(C_j l, l)}\right) - (l, m_1 - m_2)$. Выбираем шаг h_{t+1}и вычисляем y_{t+1} в зависимости от значения целевой функции. Если оптимальное значение целевой функции отрицательно, то $h_{t+1} = h_t, y_{t+1} = y_t + h_{t+1}$; иначе $h_{t+1} = h_t/2, y_{t+1} = y_{t-1} + h_{t+1}$.
3. Остановка.

После решения последней оптимизационной задачи определяется вектор l и вычисляется порог.

Задача Андерсона для двух нормальных законов с одинаковыми ковариационными матрицами

При равных ковариационных матрицах задача (3.11) будет выглядеть следующим образом:

$$\max_l \frac{(l, m_1 - m_2)}{\sqrt{(Cl, l)}}. \qquad (3.12)$$

Одно из возможных решений задачи (3.12) $l^* = C^{-1}(m_1 - m_2)$. Остальные решения задачи (3.12) будут иметь вид: $l_\mu = \mu l^*$, $\mu > 0$. Из (3.10) получаем что порог $\theta^* = \frac{(C^{-1}(m_1 - m_2), m_1 + m_2)}{2}$. С подобной дискриминантной функцией мы уже сталкивались при распознавании нормальных законов с одинаковыми матрицами – это дискриминантная функция Фишера.

Задача Андерсона для нескольких нормальных законов с одинаковыми ковариационными матрицами.

При одинаковых ковариационных матрицах для нескольких нормальных законов задача Андерсона будет выглядеть следующим образом:

$$\max_{l,\theta} \min\left(\min_{j \in J_1} \frac{(l, m_j) - \theta}{\sqrt{(Cl, l)}}, \min_{k \in J_2} \frac{\theta - (l, m_k)}{\sqrt{(Cl, l)}}\right). \qquad (3.13)$$

Зафиксируем l и обозначим через $\alpha(l) = \frac{1}{\sqrt{(Cl, l)}} \min_{j \in J_1}(l, m_j), \beta(l) = \frac{1}{\sqrt{(Cl, l)}} \max_{j \in J_2}(l, m_j)$. Вычислим оптимальное значение порога при данном l: $\theta(l) =$

$\frac{\sqrt{(Cl,l)}}{2}\big(\alpha(l)+\beta(l)\big)$. Подставим θ^*в (3.13): в результате получим задачу вычисления

$\max\limits_{l}\frac{1}{\sqrt{(Cl,l)}}\left(\min\limits_{j\in J_1}(l,m_i)-\max\limits_{k\in J_2}\big(l,m_j\big)\right)$. Обозначим через $z_{i,j}=m_i-m_j; i\in J_1, j\in J_2$, в результате требуется найти $\max\limits_{l}\min\limits_{i\in J_2, j\in J_1}\frac{(l,z_{i,j})}{\sqrt{(Cl,l)}}$. Пусть Z – выпуклая оболочка, натянутая на векторы $z_{i,j}$. Поскольку целевая функция линейная относительно $z_{i,j}$, то рассматриваемая задача эквивалентна задаче: $\max\limits_{l}\min\limits_{z\in Z}\frac{(l,z)}{\sqrt{(Cl,l)}}$. Мы считаем, что множество $\Pi_l=\{l\in R^n\colon (l,z)>0, z\in Z\}$ – непустое множество, функция $\frac{(l,z)}{\sqrt{(Cl,l)}}$ – положительно однородная функция по переменной l, поэтому исходная задача эквивалентна задаче: $\max\limits_{l\in\Pi_l\cap\{l\in R^n\colon\|l\|\le\alpha\}}\min\limits_{z\in Z}\frac{(l,z)}{\sqrt{(Cl,l)}}$. Поскольку по переменной z функция линейная и, следовательно, выпуклая на выпуклом и ограниченном множестве Z, по переменной l квазивогнутая на выпуклом и ограниченном множестве $\Gamma_l=\Pi_l\cap\{l\in R^n\colon\|l\|\le\alpha\}$, то мы можем применить теорему о минимаксе и получить задачу:

$$\min\limits_{z\in Z}\max\limits_{l\in\Pi_l\cap\{l\in R^n\colon\|l\|\le\alpha\}}\frac{(l,z)}{\sqrt{(Cl,l)}}. \tag{3.14}$$

Решение внутренней задачи имеет вид: $l=\beta C^{-1}z$, где β находится из уравнения: $\|l\|=\alpha$. В результате задача (3.14) трансформируется в задачу:

$$\min\limits_{z\in Z}\sqrt{(C^{-1}z,z)}. \tag{3.15}$$

Данная задача может быть решена при помощи алгоритма обучения Козинца. Для этого воспользуемся разложением Холецкого для положительно определенной и симметричной ковариационной матрицы $C=LL^T$, где L – нижнетреугольная матрица. После замены переменной $\bar{z}=L^{-1}z$получим задачу:

$$\min\limits_{z\in\bar{Z}}\sqrt{(z,z)}, \tag{3.16}$$

в которой множество $\bar{Z}=L^{-1}Z$. Введем одномерную нумерацию для векторов $\bar{z}_{i,j}=L^{-1}z_{i,j}$. Приведем алгоритм, решающий задачу

Алгоритм Козинца.

$begin$
$l=a; ep=\infty$
$DoWhile(ep\ge\varepsilon)$
$begin$
$j=|J_1||J_2|; l_1=l$
$fori=1Toj$
$begin$

$$h=\frac{(\bar{z}_i,\bar{z}_i-l)}{(\bar{z}_i-l,\bar{z}_i-l)}; l=\begin{cases}\bar{z}_i, h<0\\ \bar{z}_i+h(l-\bar{z}_i), 0\le h<1\\ l, 1\le h\end{cases}$$

end

$ep = \|l_1 - l\|$
end
end

В алгоритме a – начальное значение для вектора l, ε – выбранная точность

Задача Неймана-Пирсона

По-прежнему рассматривается задача распознавания двух классов, описываемых смесью нормальных законов. Задача Неймана-Пирсона отличается от задачи Андерсона и заключается в следующем:

$$\begin{gathered} max\, y \\ \alpha\sqrt{(C_j l, l)} + \theta - (l, m_j) \le 0, j \in J_1 \\ y\sqrt{(C_k l, l)} - \theta + (l, m_k) \le 0, k \in J_2 \\ y \ge 0, \end{gathered} \tag{3.17}$$

В задаче (3.17) множитель $\alpha = \Phi^{-1}(1-\beta)$, где β – допустимая вероятность пропуска цели. Задача (3.17) является задачей выпуклого программирования.

Задача Неймана-Пирсона для двух нормальных законов.

Задача (3.17) для двух нормальных законов будет иметь вид:

$$\begin{gathered} max\, y \\ \alpha\sqrt{(C_1 l, l)} + \theta - (l, m_1) \le 0, \\ y\sqrt{(C_2 l, l)} - \theta + (l, m_2) \le 0, \\ y \ge 0, \end{gathered} \tag{3.18}$$

Решим задачу (3.18) при фиксированном векторе l. Оптимальное значение для $y - y^* = \frac{(l, m_1 - m_2)}{\sqrt{(C_2 l, l)}} - \alpha\sqrt{\frac{(C_1 l, l)}{(C_2 l, l)}}$, при пороге $\theta^* = (l, m_1) - \alpha\sqrt{(C_1 l, l)}$. Таким образом, следует вычислить максимум функции $F(l) = \frac{(l, m_1 - m_2)}{\sqrt{(C_2 l, l)}} - \alpha\sqrt{\frac{(C_1 l, l)}{(C_2 l, l)}}$.

Линейные дискриминантные функции для конечных множеств точек

В методах предыдущего раздела используются математические ожидания и ковариационные матрицы. Если семейство математических ожиданий и ковариационных матриц – неизвестно, то требуется вычислить их оценки по обучающей выборке, что не всегда является рациональным. В этом разделе мы рассмотрим подход, при котором этого делать не требуется. Рассмотрим сначала задачу распознавания двух классов, представленных двумя множествами точек. Таким образом, имеется множество точек x_j, представляющих первый класс, если $j \in J_1$, и представляющих второй класс, если $j \in J_2$.

Линейная разделимость. Будем говорить, что множества точек линейно разделимы, если существует такой порог θ и вектор l, для которых выполняются неравенства:

$$\begin{aligned} (l, x_j) &> \theta, j \in J_1 \\ (l, x_j) &< \theta, j \in J_2 \end{aligned} \qquad (3.19)$$

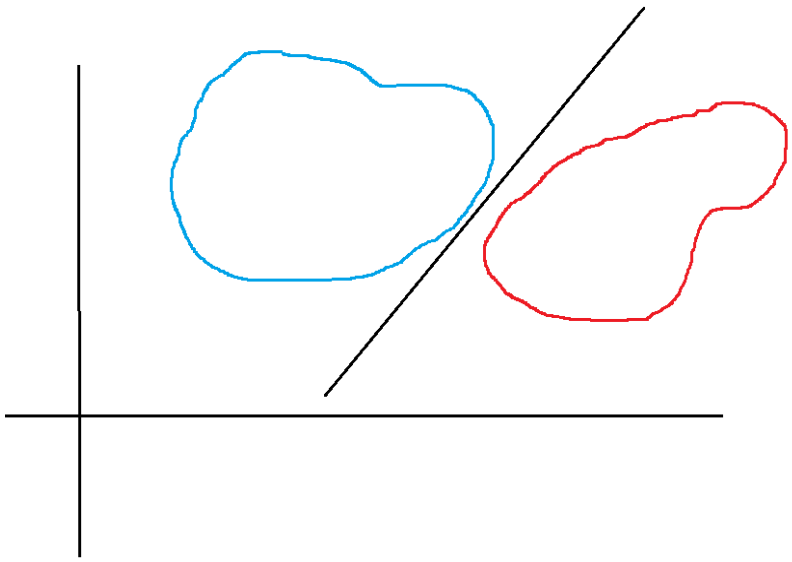


Рис 2. Линейная разделимость

Легко проверяется следующее утверждение.

Утверждение. Множества точек разделимы тогда и только тогда, когда выпуклые оболочки множеств не пересекаются.

Рассмотрим задачу линейного программирования, связанную с линейным разделением множеств:

$$\begin{aligned} &max\, z \\ &(\bar{l}, y_j) \geq z, j \in J_1 \cup J_2. \end{aligned} \qquad (3.20)$$

В задаче (3.20) векторы $\bar{l} = \begin{pmatrix} l \\ \theta \end{pmatrix}, y_j = \begin{pmatrix} x_j \\ -1 \end{pmatrix}$, если $j \in J_1$, и $y_j = \begin{pmatrix} -x_j \\ 1 \end{pmatrix}$, если $j \in J_2$. Является справедливым следующее утверждение.

Утверждение. Задача (3.20) имеет конечное решение тогда и только тогда, когда множества – линейно неразделимы.

Действительно, пусть задача (3.20) имеете решение, тогда для этого конечного решения $z^* \leq 0$. Это означает, что множества линейно неразделимы. Допустим, что множества линейно разделимы, тогда существует положительное $\tilde{z}$, для которого выполняются неравенства в (3.20). Это означает, что $max\ z = \infty$.

Решение задачи (3.20) вычисляется при помощи алгоритма подъема по обобщенному градиенту.

Алгоритм. Линейная разделимость.

$begin$

$\bar{l} = a; ep = \infty$

$Do\ While\big((ep \geq \varepsilon) \wedge (z \leq 0)\big)$

$begin$

$z = \min\limits_{j \in J_1 \cup J_2} (\bar{l}, y_j)$

$if z > 0\ then\ goto\ met1$

$j = arg \min\limits_{j \in J_1 \cup J_2} (\bar{l}, y_j)$

$\bar{l} = \bar{l} + hy_j$

$ep = abs\left(z - \min\limits_{j \in J_1 \cup J_2} (\bar{l}, y_j)\right)$

end

$met1: if z > 0$

$then\ print$ множества разделимы $else\ print$ множества неразделимы

end

В алгоритме $h > 0$– величина постоянного шага, ε – выбранная точность.

Если множества разделимы, то существует бесконечное множество гиперплоскостей, которые разделяют эти множества. Обозначим множество векторов нормалей разделяющих гиперплоскостей через L. Пусть $\alpha(l) = \min\limits_{i \in J_1} \frac{(l, x_i)}{\|l\|}, \beta(l) = \max\limits_{j \in J_2} \frac{(l, x_j)}{\|l\|}$. Естественным показателем качества разделяющей гиперплокости с вектором нормали l является разность $d(l) = \alpha(l) - \beta(l)$. Чем больше данная разность, тем лучше разделяет гиперплоскость множества. В результате возникает задача по вычислению оптимальной разделяющей гиперплоскости:

$$\max_{l \in L} \min_{z \in Z} \frac{(l,z)}{\|l\|}, \tag{3.21}$$

в которой множество Z – выпуклая оболочка, натянутая на векторы $x_i - x_j, i \in J_1, j \in J_2$. Применение теоремы о минимаксе приводит к задаче:

$$\min_{z \in Z} \max_{l \in L} \frac{(l,z)}{\|l\|}, \tag{3.22}$$

решение внутренней задачи которой имеет вид: $l = \alpha z, \alpha > 0$. Подставка решения в (3.22) приводит к задаче:

$$\min_{z \in Z} \|z\|. \tag{3.23}$$

Для решения задачи (3.23) можно применить алгоритм Козинца, который был приведен ранее. Алгоритм Козинца позволяет вычислить вектор нормали l оптимальной разделяющей гиперплоскости. Оптимальный порог $\theta = \frac{\min_{i \in J_1}(l,x_i) + \min_{j \in J_2}(l,x_j)}{2}$.

Лекция 6. Продолжение. Множества линейно неразделимы

Оптимизация с векторным критерием.

Остановимся на случае линейной неразделимости множеств, рис. .

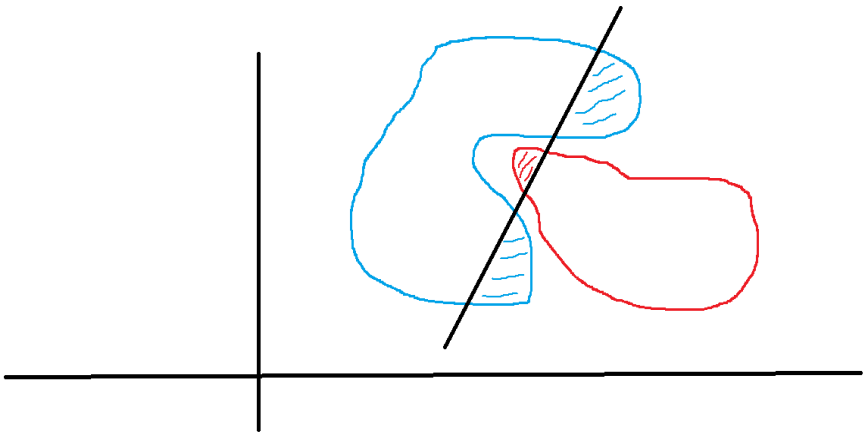


Рис.3. Линейная неразделимость

Выберем вектор нормали l и вычислим следующие величины, которые могут служить показателями качества выбора. Прежде всего, вычислим среднее значение $\alpha(l) = (l, m_1 - m_2)$, в котором $m_1 = \frac{1}{|J_1|}\sum_{i\in J_1} x_i\,, m_2 = \frac{1}{|J_2|}\sum_{j\in J_2} x_j$. Чем больше этот показатель, тем лучше решающее правило. Далее вычислим разбросы для первого и второго классов:$\beta(l) = \left(\left(\frac{1}{|J_1|}\sum_{i\in J_1} x_i x_i^T - m_1 m_1^T\right) l, l\right), \gamma(l) = \left(\left(\frac{1}{|J_2|}\sum_{j\in J_2} x_j x_j^T - m_2 m_2^T\right) l, l\right)$. Эти показатели должны быть как можно меньше. Таким образом, возникает оптимизационная задача с тремя критериями.

Для этих критериев рассмотрим выпуклую оптимизационную задачу, в которой первый и второй критерий используются в качестве ограничений. Таким образом, задача имеет следующий вид:

$$\begin{gathered} min(C_2 l, l) \\ (l, m_1 - m_2) \geq a \\ (C_1 l, l) \leq b, \end{gathered} \tag{3.24}$$

в которой $C_1 = \frac{1}{|J_1|}\sum_{i\in J_1} x_i x_i^T - m_1 m_1^T, C_2 = \frac{1}{|J_2|}\sum_{j\in J_2} x_j x_j^T - m_2 m_2^T$. Матрицы C_1 и C_2– симметричные и неотрицательно определенные матрицы. Будем считать, что они положительно определены. Параметры задачи a и b положительны и выбраны так, чтобы множество допустимых решений задачи (3.24) было непустым множеством. Легко устанавливается следующий факт.

Утверждение. Множество допустимых решений задачи (3.24) – непустое множество, если выполняется неравенство:

$$b \geq \frac{a^2}{(C_1^{-1}(m_1-m_2),m_1-m_2)}. \quad (3.25)$$

Функция Лагранжа для этой задачи будет иметь вид: $F(l,\lambda,\mu) = (C_2 l, l) + \lambda\big(a - (l, m_1 - m_2)\big) + \mu\big((C_1 l, l) - b\big)$. Условие оптимальности и условия Куна-Таккера для задачи (3.24):

$$\begin{aligned} &(C_2 + \mu C_1)l = \lambda(m_1 - m_2), \quad (3.26)\\ &\lambda\big(a - (l, m_1 - m_2)\big) = 0,\\ &\mu\big((C_1 l, l) - b\big) = 0,\\ &(l, m_1 - m_2) \geq a,\\ &(C_1 l, l) \leq b, \lambda \geq 0, \mu \geq 0. \end{aligned}$$

Проанализируем условия Куна-Таккера. Поскольку $l = 0$ не удовлетворяет ограничению: $(l, m_1 - m_2) \geq a$, то $\lambda > 0$. Тогда $(l, m_1 - m_2) = a$. Пусть $\mu = 0$, тогда вектор $l = \left[\frac{a}{(C_2^{-1}(m_2-m_1),m_2-m_1)}\right] C_2^{-1}(m_2 - m_1)$ – решение задачи, если $b \geq \left[\frac{a}{(C_2^{-1}(m_2-m_1),m_2-m_1)}\right]^2 \big(C_1 C_2^{-1}(m_2 - m_1), C_2^{-1}(m_2 - m_1)\big)$, иначе $\mu > 0$. Таким образом, если $\lambda > 0$,$\mu > 0$, то условия (3.26) превращаются в систему уравнений:

$$\begin{aligned} &(C_2 + \mu C_1)l = \lambda(m_1 - m_2), \quad (3.27)\\ &(l, m_1 - m_2) = a,\\ &(C_1 l, l) = b, \end{aligned}$$

Задача существенно упрощается, если $\mathrm{C}_1 \approx \mathrm{C}_2$ в смысле какой-либо из матричных норм. В этом случае мы можем взять в качестве общей матрицы матрицу$C = \frac{1}{2}(C_1 + C_2)$ и решить следующую задачу:

$$\begin{aligned} &min(Cl, l) \quad (3.28)\\ &(l, m_1 - m_2) \geq a \end{aligned}$$

Решение данной задачи будет иметь вид: $l = \frac{a}{(C^{-1}(m_2-m_1),m_2-m_1)} C^{-1}(m_2 - m_1)$. Поскольку параметр a входит в фактор нормализации, то он может быть выбран произвольно, например, $a = (C^{-1}(m_2 - m_1), m_2 - m_1)$. Таким образом, оптимальный вектор $l = C^{-1}(m_2 - m_1)$.

Минимизация вероятности ошибки линейного решающего правила на линейно неразделимых множествах.

Еще раз применим прием, который сводит вычисление вектора нормали к решению системы однородных линейных неравенств $(l, x) > 0$, $x \in V$. Проблема заключается в том, что система не имеет решения. Поэтому вместо решения системы рассмотрим оптимизационную задачу:
$\min_l \frac{1}{|V|} \sum_{x \in V} I_{\{(l,x) \leq 0\}}(l)$. В результате решения этой задачи будет получено решающее правило, которое неправильно распознает минимальное число элементов выборки или минимизирует эмпирическую вероятность ошибки. Сложность заключается в том, что целевая функция этой задачи не является выпуклой, что значительно усложняет задачу. Вместо этой задачи рассмотри

задачу: $\min_{l} \frac{1}{|V|} \sum_{x \in V} \big(1 - (l, x)\big)^+$. Целевая функция этой задачи – выпуклая функция. Поскольку функция $(1 - y)^+ \geq I_{\{y \leq 0\}}(y)$, см. рис. 2, то решение второй задачи является оценкой сверху решения первой задачи, то есть вероятности эмпирической ошибки решающего правила.

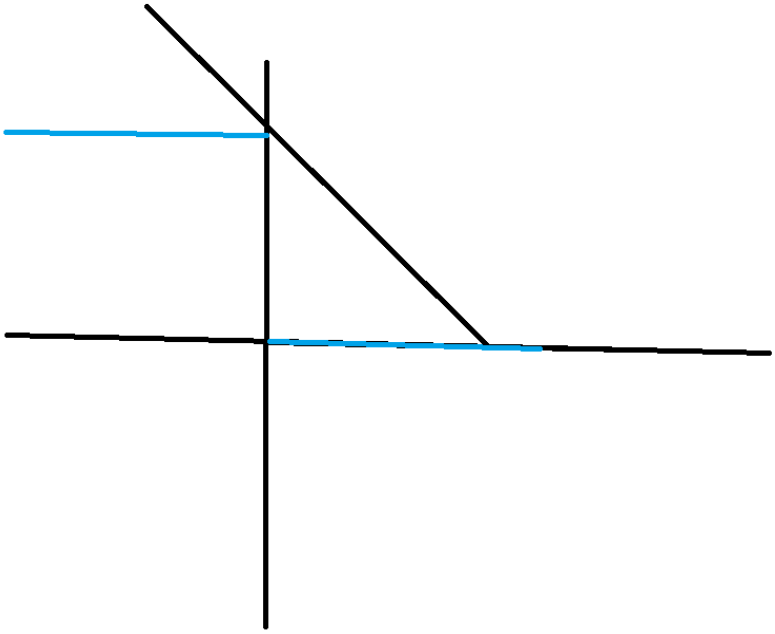


Рис. 4. Минимизация эмпирической вероятности ошибки распознавания.

Распознавание нескольких классов.

Пусть множество индексов J разбито на подмножества: $J = \bigcup_{i=1}^{k} J_i$. Линейное решающее правило для нескольких классов:

$$f(x) = arg\, \max_j \left[(l_j, x) + \theta_j\right]. \tag{3.29}$$

Решающее правило (3.29) преобразуется следующим образом:

$$f(x) = arg\, \max_j \left(\bar{l}_j, \bar{x}\right), \tag{3.30}$$

Где $\bar{l}_j = \begin{pmatrix} l_j \\ \theta_j \end{pmatrix}, \bar{x} = \begin{pmatrix} x \\ 1 \end{pmatrix}$. Для упрощения изложения далее будем рассматривать распознавание трех классов. Сформируем множество $X = X_1 \cup X_2 \cup X_3$, где $X_1 = \left\{ \begin{pmatrix} \bar{x}_i \\ -\bar{x}_j \\ 0 \end{pmatrix}, i \in J_1, j \in J_2 \right\} \cup \left\{ \begin{pmatrix} \bar{x}_i \\ 0 \\ -\bar{x}_j \end{pmatrix}, i \in J_1, j \in J_3 \right\},$

$$X_2 = \left\{ \begin{pmatrix} -\bar{x}_j \\ \bar{x}_i \\ 0 \end{pmatrix}, i \in J_2, j \in J_1 \right\} \bigcup \left\{ \begin{pmatrix} 0 \\ \bar{x}_i \\ -\bar{x}_j \end{pmatrix}, i \in J_2, j \in J_3 \right\},$$

$$X_3 = \left\{ \begin{pmatrix} -\bar{x}_j \\ 0 \\ \bar{x}_i \end{pmatrix}, i \in J_3, j \in J_1 \right\} \bigcup \left\{ \begin{pmatrix} 0 \\ -\bar{x}_j \\ \bar{x}_i \end{pmatrix}, i \in J_3, j \in J_2 \right\}.$$

Легко проверяется следующее утверждение.

Утверждение. Решающее правило (3.309) правильно распознает выборку, тогда и только тогда, когда выпуклая оболочка, натянутая на множество X строго отделена от нуля. Отделимость от нуля можно проверить, решая задачу линейного программирования, аналогичную задаче (3.20):

$$\begin{matrix} max\, z \\ (L, y) \geq z, y \in X, \end{matrix} \qquad (3.31)$$

для решения которой можно применить алгоритм «Линейная разделимость» и в зависимости от результата сделать вывод о возможности безошибочного распознавания выборки. Если $z^* = \infty$, то решающее правило (3.30) распознает выборку без ошибок, если $z^* \leq 0$, то решающее правило не может распознать выборку без ошибок. В (3.31) вектор $L = \begin{pmatrix} \bar{l}_1 \\ \bar{l}_2 \\ \bar{l}_3 \end{pmatrix}$. В случае линейной разделимости можно использовать алгоритм Козинца для вычисления оптимального вектора L. В случае линейной неразделимости может быть использована задача оптимизации:

$$\begin{matrix} min(CL, L) \\ (L, \bar{y}) \geq a \end{matrix} \qquad (3.32)$$

для вычисления параметров решающего правила (3.29). В (3.32) $\bar{y}$ – выборочное среднее, C –- выборочная ковариационная матрица для множества X. Поскольку подбная задача обсуждалась раньше, то на этом завершим рассмотрение распознавание нескольких классов.

Нелинейные решающие правила.

Рассмотрим решающее правило следующего вида:

$$l(x) = \begin{cases} 1, & f(x) \geq 0 \\ 2, & f(x) < 0 \end{cases},$$
$$f(x) = \sum_{i=1}^{N} a_i \varphi_i(x).$$

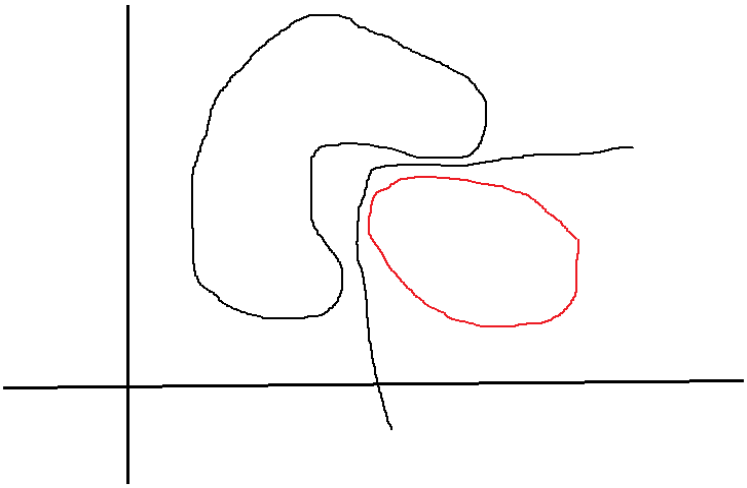


Рис 5. Нелинейное решающее правило.

Предполагается, что функции φ_i образуют базис, например, в пространстве функций от n переменных, интегрируемых с квадратом. Приведем пример такой функции: $f(x) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} \sum_{j=i}^{n} a_{i,j} x_i x_j$. Как правило, число слагаемых больше размерности исходного пространства признаков. Этот факт расширяет возможности решающего правила. В примере $N = n + \frac{n(n+1)}{2}$.

Рассмотрим отображение $\Phi: R^n \to R^N$, которое определяется равенством $\big(\Phi(x)\big)_i = \varphi_i(x)$. В пространстве R^N данное решающее правило будет иметь вид:

$$l(x) = \begin{cases} 1, & (a, y) \geq 0 \\ 2, & (a, y) < 0 \end{cases}, y = \Phi(x).$$

Данное решающее правило является линейным пороговым решающим правилом в пространстве R^N, поэтому данное пространство называется спрямляющим пространство. Мы уже сталкивались с этим понятием в разделе, посвященном байесовским решающим правилам. Для определения параметров этого правила могут быть использованы алгоритмы из предыдущих разделов в спрямляющем пространстве. На одном из них остановимся подробнее.

Обобщим задачу обучения. Рассмотрим выборку $V = \{(\Phi(x_1), z_1), \dots, (\Phi(x_i), z_i)\}$ и задачу регрессии для этой выборки:

$$\min_a \left[\sum_{i=1}^{m} \Big(\big(a, \Phi(x_i)\big) - z_i \Big)^2 + \alpha(a, a) \right] \tag{3.33}$$

с регуляризацией $\alpha(a, a)$, $\alpha > 0$. Необходимым и достаточным условием минимума является равенство:

$$\sum_{i=1}^{m} \Big(\big(a, \Phi(x_i)\big) - z_i \Big) \Phi(x_i) \; + \alpha a = 0.$$

Из этого равенства следует, что оптимальное решение принадлежит линейной оболочке $L\{\Phi(x_i)\}$. Таким образом оптимальное значение для вектора имеет следующий вид:

$$a = \sum_{i=1}^{m} \beta_i \Phi(x_i).$$

Следовательно, оптимальная функция

$$f(x) = \big(a, \Phi(x)\big) = \sum_{i=1}^{m} \beta_i \big(\Phi(x), \Phi(x_i)\big)$$

Рассмотрим функцию от двух векторных аргументов

$$K(x, y) = \big(\Phi(x), \Phi(y)\big).$$

Функцию $K(x, y)$ называют потенциальной или еще ядром. Потенциальную функцию можно использовать для выражения решающего правила:

$$f(x) = \sum_{i=1}^{m} \beta_i K(x, x_i)$$

Параметры решающего правила могут быть вычислены в результате решения оптимизационной задачи:

$$\min_{\beta} \left[\sum_{i=1}^{m} \left(z_i - \sum_{j=1}^{m} \beta_j K\big(x_i, x_j\big) \right)^2 + \alpha \sum_{j=1}^{m} \beta_j^2 \right]$$

Таким образом, вместо использования базиса $\{\varphi_i\}$ можно использовать потенциальную функцию от двух векторных аргументов $K(x, y)$. Наиболее употребительной в машинном обучении является функция Гаусса:

$$K(x, y) = e^{-1/2\sigma^2 \|x-y\|^2}.$$

Логистическая регрессия.

В этом разделе в качестве нелинейного решающего правила рассмотрим функцию следующего вида:

$$f(x) = \varphi\big((l, x)\big).$$

Функция $\varphi(y)$ – функция от одной переменной – является сигмоидной функцией, например, $\varphi(y) = \frac{1}{1+e^{-x}}$, $\varphi(y) \in (0,1)$. Значение $f(x)$ будем трактовать как вероятность того, что образец x принадлежит первому классу. Рассмотрим обучающую выборку $V = \{(z_1, x_1), \dots, (z_m, x_m)\}$, в которой $z_i \in \{0,1\}$. Равенство $z_i = 1$ означает, что образец x_i принадлежит первому классу,

равенство $z_i = 0$ означает, что образец x_i принадлежит второму классу. Применим метод максимального правдоподобия для задачи обучения. Для данной выборки логарифм максимального правдоподобия выглядит следующим образом:

$$lnL(l) = \sum_{i=1}^{m} \left(z_i ln\varphi\big((l, x_i)\big) + (1 - z_i) ln\left(1 - \varphi\big((l, x_i)\big)\right)\right).$$

Таким образом, задача обучения заключается в следующем:

$$\min_{l} \sum_{i=1}^{m} \left(z_i ln\varphi\big((l, x_i)\big) + (1 - z_i) ln\left(1 - \varphi\big((l, x_i)\big)\right)\right)$$

Задание. Доказать, что для $\varphi(y) = \frac{1}{1+e^{-x}}$ критерий данной задачи является вогнутой функцией.

3.3 Библиография

Задача Андерсона была опубликована в работе [16]. Обобщение задачи Андерсона было выполнено М.И. Шлезингером в работах М. Шлезингера [17, 18]. Алгоритм Козинца опубликован в работе [19]. Применение методов оптимимзации для решения задачи линейного разделения конечного множества точек описано в уже упомянутой работе Козинца [19], в монографии [5], в работе Вапника и Червоненкиса [23] предложен метод обучения, который получил название метод обобщенного портрета. В современной литературе этот метод называется методом опорных векторов. Метод потенциальных функций впервые был представлен в работе Айзермана, Бравермана и Розоноера [20]. Мы не уделили достаточно внимания методам, в которых использовались ряд Вольтера и Вейвлет-базис см., например, работы [21,22].

Четвертый раздел. Обучение в реальном времени.

Лекция 7. Основные понятия и определения

Представим задачу обучения как игру, протекающие в Представим задачу обучения как игру, протекающую в несколько раундов. В каждом раунде наш противник выбирает ход y_t, мы выбираем ход x_t. В результате наши потери составляют $W(x_t, y_t)$.

Смысл игры заключается в том, чтобы сделать наши потери как можно меньшими. Так за T раундов наши потери составят $\sum_{t=1}^{T} W(x_t, y_t)$. Так как потери в этой игре неизбежны, то нам необходимо определить для себя: какие потери мы будем считать допустимыми. То есть, что считать выигрышем. Несколько упростим игру. Будем считать, что выбор противника случаен, то естьy_t – независимые случайные величины с неизвестным нам законом распределения вероятностей – $Law(y)$. Противник выбирает закон распределения вероятностей вначале игры и не меняет его на протяжении игры. В этом случае естественно рассматривать средние потери $E\sum_{t=1}^{T} W(x_t, y_t)$. Пусть существует второй момент $Ey^2 < \infty$. Конкретизируем потери $W(x_t, y_t) = (x_t - y_t)^2$. Если бы закон распределения вероятностей был нам известен, то с данной функцией потерь оптимально было бы в каждом раунде игры t выбирать $x_t = Ey_1$, в результате применения оптимальной стратегии средние потери за T раундов составили бы величину $T\sigma^2$. Однако, мы не знаем закона распределения вероятностей. Из-за незнания закона распределения вероятностей мы не можем использовать оптимальную стратегию. Поэтому средние потери в результате применения иной стратегии будут отличаться от оптимальных потерь. Эту разницу, естественно, рассматривать как показатель качества выбранной нами стратегии. Таким образом, показателем качества стратегии, который далее будем называть сожалением, является величина:

$$R_T(\pi) = E\sum_{t=1}^{T}(x_t - y_t)^2 - T\sigma^2 \tag{4.1}$$

Прежде всего, отметим, что сожаление – неотрицательная величина, которая зависит от числа раундов Tи стратегии π.Стратегия генерирует последовательность ходов x_t. Отметим, что сожаление учитывает сложность игры, которая выражается через дисперсию σ^2. Стратегию следует считать успешной или выигрышной, если $\lim_{T\to 0} \frac{R_T(\pi)}{T} = 0$. Успешная стратегия в определенном смысле аппроксимирует оптимальную стратегию.

Неизвестность закона распределения вероятностей, выбранного противником, делает необходимым выполнить некоторое преобразование формулы (4.1). Сначала заменим $T\sigma^2$ на равную величину $\min_x E\sum_{t=1}^{T}(x - y_t)^2$ и перепишем сожаление следующим образом:

$$R_T(\pi) = E\sum_{t=1}^{T}(x_t - y_t)^2 - \min_x E\sum_{t=1}^{T}(x - y_t)^2.$$

Далее мы можем отказаться от использования определенного закона распределения для последовательности y и рассмотреть произвольную последовательность y. Поскольку исчезла стохастичность, то из определения сожаления следует удалить математическое ожидание:

$$R_T(\pi) = \sum_{t=1}^{T}(x_t - y_t)^2 - \min_x \sum_{t=1}^{T}(x - y_t)^2.$$

Отметим, что оптимальную стратегию мы сможем определить в конце игры, то есть в момент времени T. В момент времени t мы обладаем определенной информацией относительно целевой функции $F(x) = \sum_{t=1}^{T}(x - y_t)^2$, однако, полностью она нам неизвестна. Будем считать, что в момент времени t, когда мы принимаем решение относительно x_t, нам известна только часть последовательности y – $y_1, \dots, y_{t-1}$. Попробуем построить выигрышную стратегию, опираясь только на интуицию. Если бы мы знали всю последовательность, то мы могли бы найти $\min_x F(x)$. Оптимальное значение в этой задаче $x_T^* = \frac{1}{T}\sum_{t=1}^{T} y_t$, то есть среднее. Поскольку мы не умеем заглядывать в будущее, то в момент времени t мы наблюдаем только часть последовательности y, а именно, $y_1, y_2, \dots, y_{t-1}$. Это позволяет найти наилучшее решение для этой части последовательности $x_{t-1}^* = \frac{1}{t-1}\sum_{i=1}^{t-1} y_i$ и выбрать $x_t = x_{t-1}^*$, $t \geq 2$, и произвольно выбрать x_0. Такая стратегия является частным случаем класса стратегий, которые имеют общее название: «Следуй за лидером». Следуя этой стратегии, мы рассчитываем на то, что будущее не сильно отличается от прошлого. Позже покажем, что эта стратегия приведет к успеху.

Но прежде всего, определим предмет исследования. Для этого обобщим сожаление. Пусть имеется последовательность штрафов $l_t(x)$. Выразим сожаление $R_T(\pi, u) = \sum_{t=1}^{T} l_t(x_t) - \sum_{t=1}^{T} l_t(u)$ относительно произвольного решения u. В сущности онлайн обучение – это не что иное, как разработка и анализ алгоритмов минимизации сожаления для заданной последовательности функции потерь по отношению к произвольному конкуренту. В этих рамках онлайн-обучение можно использовать для решения широкого класса задач, которые можно сформулировать в терминах задачи минимизации сожаления. По широте использования онлайн-обучение сопоставимо с теорией игр. Отметим также, что онлайн-обучение позволяет обрабатывать данные произвольной природы зависимых или независимых, связанных или нет с конкретным, но неизвестным распределением.

Вернемся к анализу предложенной стратегии в нашем простом примере. Отметим, что в примере $l_t(x) = (x - y_t)^2$. Понадобится следующая лемма.

Лемма. Пусть $l_t(x)$ – последовательность штрафов и x_t^* – минимизатор кумулятивных потерь за t раундов, тогда

$$\sum_{t=1}^{T} l_t(x_t^*) \leq \sum_{t=1}^{T} l_t(x_T^*).$$

Доказательство проведем методом математической индукции. Для $T = 1$ неравенство очевидно. Пусть оно выполняется для $T - 1$, $T \geq 2$.

$$\sum_{t=1}^{T-1} l_t(x_t^*) \leq \sum_{t=1}^{T-1} l_t(x_{T-1}^*).$$

Последние слагаемые в суммах $\sum_{t=1}^{T} l_t(x_t^*)$ и $\sum_{t=1}^{T} l_t(x_T^*)$ одинаковые, поэтому сравним суммы $\sum_{t=1}^{T-1} l_t(x_t^*)$ и $\sum_{t=1}^{T-1} l_t(x_T^*)$. Из индуктивного предположения следует неравенство $\sum_{t=1}^{T-1} l_t(x_t^*) \leq \sum_{t=1}^{T-1} l_t(x_{T-1}^*)$. Из определения x_{T-1}^* следует неравенство $\sum_{t=1}^{T-1} l_t(x_{T-1}^*) \leq \sum_{t=1}^{T-1} l_t(x_T^*)$. Из последнего и предпоследнего неравенств следует неравенство: $\sum_{t=1}^{T} l_t(x_t^*) \leq \sum_{t=1}^{T} l_t(x_T^*)$.

То есть, при известном будущем, адаптивная стратегия выглядит предпочтительней. Эта лемма поможет доказать следующую теорему.

Теорема. Для произвольной равномерно ограниченной последовательности y, последовательности штрафов$l_t(x) = (x - y_t)^2$справедливо неравенство:

$$R_t(\pi) \leq x_0^2 + M^2 + 4M^2(1 + ln\, T).$$

В неравенстве $M = sup|y_t|$.

Применив лемму, получим первое неравенство:

$$R_T(\pi) \leq \sum_{t=1}^{T} |l(x_{t-1}^*) - l(x_t^*)|.$$

Далее используем последовательность потерь и равномерную ограниченность последовательности y:

$$R_T(\pi) \leq x_0^2 + M^2 + 2\sum_{t=1}^{T} |y_t|\,|x_t^* - x_{t-1}^*| \leq x_0^2 + M^2 + 2M\sum_{t=1}^{T} |x_t^* - x_{t-1}^*|.$$

Подставим x_t^*и x_{t-1}^*в последнее неравенство и в результате получим очередное неравенство:

$$R_T(\pi) \leq x_0^2 + M^2 + 2M\sum_{t=1}^{T}\left[\frac{1}{t(t-1)}\sum_{j=1}^{t-1}|y_j| + \frac{|y_t|}{t}\right] = x_0^2 + M^2 + 4M^2\sum_{t=1}^{T}\frac{1}{t}.$$

Далее используем неравенство $\sum_{t=1}^{T}\frac{1}{t} \leq 1 + ln\, T$ и получим доказательство теоремы.

Из теоремы следует, что для предлагаемой стратегии $\lim\limits_{T\to\infty}\frac{R_T(\pi)}{T} = 0$, то есть, согласно определению, данная стратегия является выигрышной.

Есть несколько вещей, которые необходимо подчеркнуть в этой стратегии. Во-первых, в стратегии отсутствуют параметры, подлежащие настройке. Это характерно для онлайн-обучения. Поскольку мы имеем поток данных (данные могут исчезать) и у нас нет возможности перезапускать алгоритм для настройки параметров стратегии. От прошлого нам нужно только скользящее среднее. Этот факт делает нашу стратегию вычислительно эффективной. Поэтому при разработке алгоритмов онлайн-обучения нужно стараться сохранять это свойство алгоритма. Последнее, что необходимо отметить, в алгоритме не используются градиенты. Градиенты – полезная вещь, также как и субградиенты, при дизайне алгоритмов обучения. Далее градиенты и субградиенты будут встречаться часто.

Градиентный и субградиентный спуск

В этом разделе мы рассмотрим алгоритм для выпуклых штрафов, и начнем наше описание со случая дифференцируемых штрафов.

Вернемся к исходной задаче в игровой постановке, которую формально можно представить следующим образом.

$$Fort = 1toT$$
$$Outputx_t \in V \subset R^d$$
$$Lossl_t(x_t)$$
$$endt$$

Целью игры является минимизация сожаления

$$R_T(x_1, \ldots, x_T; u) = \sum_{t=1}^{T} l_t(x_t) - \sum_{t=1}^{T} l_t(u)$$

относительно альтернативы u. Основная проблема заключается в том, что в момент выбора x_t целевая функция известна только частично, а именно, $\sum_{j=1}^{t-1} l_j(x)$. В этом сосредоточен весь смысл задачи онлайн-оптимизации. Чтобы надеяться на успех при решении этой проблемы, необходимо сделать предположения относительно последовательности функций $l_t(x)$. Например, мы можем рассматривать выпуклые штрафы $l_t(x)$. Вполне возможно, потребуются дополнительные предположения относительно последовательности функций штрафов или функций потерь. Хорошим условием является условие Липшица дополнительно к выпуклости функций потерь. Такие предположения относительно последовательности функций потерь – это попытка контроля будущего, которым мы не обладаем. С другой стороны, чем уже рассматриваемый класс функции потерь, тем проще анализ алгоритма онлайн-обучения. Хорошие алгоритмы – это те алгоритмы, которые гарантируют сублинейность сожаления при слабых предположениях относительно последовательности потерь и одновременно гарантируют минимальное сожаление для слабых противников – сильных предположениях относительно последовательности функций потерь. Поскольку, по сути дела, мы не знаем, как противник генерирует функции потерь, сожаление хорошо измеряет сложность задачи, за решение которой мы принялись. Если алгоритм гарантирует сублинейность сожаления, то это означает, что качество стратегии приближается к качеству произвольной фиксированной стратегии, в том числе стратегии, которая минимизирует потери.

После этих предварительных рассуждений остановимся на выпуклых потерях.

Элементы выпуклого анализа

Начнем с определения выпуклого множества. Выпуклым множеством называется множество, которое вместе с любыми своими точками содержит отрезок, который их соединяет. То есть для произвольных x и y, принадлежащих выпуклому множеству S, и для произвольного λ, принадлежащего интервалу $[0,1]$, выпуклая комбинация $\lambda x + (1-\lambda)y$принадлежит выпуклому множеству S. Если отказаться от условия $\lambda \in [0,1]$, то комбинация называется аффинной комбинацией. Если одновременно с двумя произвольными точками множество S

содержит их аффинную комбинацию, то множество S называют аффинным множеством.

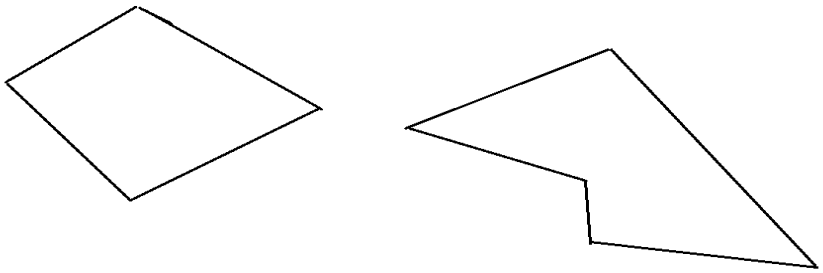


Рис.6. Выпуклое (слева) и невыпуклое (справа) множества

Пусть множество решений системы линейных уравнений $Ax = b$ содержит бесконечное множество точек, тогда это множество будет аффинным. Доказать этот факт в качестве упражнения.

Очевидно, что аффинное множество является выпуклым, а вот обратное утверждение неверно. Добавим к системе линейных уравнений дополнительное условие на ее решение. Будем рассматривать только решения с неотрицательными элементами. Если множество неотрицательных решений системы линейных уравнений содержит бесконечное число элементов, то оно будет выпуклым множеством и не будет аффинным.

Проверить, являются ли выпуклыми множествами полупространство: $(a, x) \leq b$ в R^n и множество решений системы неравенств: $y \geq \frac{1}{x}, x > 0$ в R^2.

Пересечение любой системы выпуклых множеств является выпуклым множеством.

Показать, что множество решений системы линейных неравенств $Ax \leq b$ является выпуклым множеством.

Конус. Конусом называют множество, которое вместе с любым своим элементом x содержит элемент λx для произвольного неотрицательного λ **Выпуклым конусом** называют множество, которое вместе с двумя своими элементами x и y содержит их линейную комбинацию $\lambda x + \mu y$ с неотрицательными коэффициентами.

Показать, что выпуклый конус является выпуклым множеством.

Пусть $\|.\|$– некоторая норма в R^n, необязательно эвклидова норма, тогда $\|\lambda x + (1-\lambda)y\| \leq \lambda\|x\| + (1-\lambda)\|y\|$, для произвольного $\lambda \in [0,1]$. Поэтому шар $B_r(c) = \{x: \|x\| \leq r\}$ является выпуклым множеством.

Эллипсоид. Эллипсоидом называется множество $El(c) = \{x: (C^{-1}(x - c), x - c) \leq 1\}$, где C– положительно определенная симметричная матрица. Показать, что эллипсоид является выпуклым множеством.

Множества A и B называют **отделимыми**, если существует такой вектор a, что $(a,x) \le (a,y)$, для любых $x \in A, y \in B$, и хотя бы для одной пары выполняется строгое неравенство $(a,\bar{x}) < (a,\bar{y})$.

Эквивалентное определение отделимости. Множества A и B называют отделимыми, если существует такая гиперплоскость $\Pi: (a,x) = b$, что $(a,x) \le b, (a,y) \ge b, \forall x \in A, \forall y \in B$, и хотя бы одно из множеств не лежит целиком в гиперплоскости Π. Гиперплоскость Π при этом называют **разделяющей**. Иллюстрация приведена на рис.7.

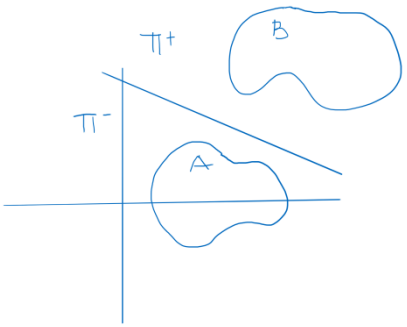


Рис.7. Разделяющая гиперплоскость

Множества A и B называют **сильно отделимыми**, если существует такой вектор a, что $\sup\limits_{x \in A}(a,x) < \inf\limits_{y \in B}(a,y)$.

Очень важную роль в выпуклом анализе играют **теоремы отделимости**.

Первая теорема звучит так. Непересекающиеся выпуклые множества отделимы.

Вторая теорема. Непересекающиеся замкнутые выпуклые множества, хотя бы одно из которых – ограниченное множество, сильно отделимы.

Тесно связана с этими теоремами теорема об опорной гиперплоскости.

Третья теорема – теорема об опорной гиперплоскости. Пусть x_0– граничная точка выпуклого множества, тогда существует такая гиперплоскость $\Pi: (a(x - x_0), x - x_0) = 0$, проходящая через эту точку, что выпуклое множество вложено в Π^+.

Выпуклой функцией называется функция, для которой область определения выпуклое множество и выполняется неравенство:

$$f(\lambda x + (1-\lambda)y) \le \lambda f(x) + (1-\lambda) f(y)$$

для любых x и y из области определения функции и любых λ из интервала $[0,1]$.

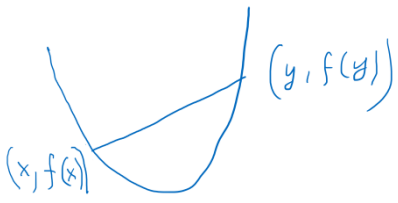


Рис.8. Хорда и график выпуклой функции.

На рис.8 представлен график выпуклой функции и произвольная хорда. Рис.8 иллюстрирует определение выпуклой функции. Приведем эквивалентное определение выпуклой функции.

Эквивалентное определение выпуклой функции. Функция называется **выпуклой функций**, если ее область определения выпуклое множество и ее надграфик $\{(x,y): x \in D, f(x) \leq y\}$– выпуклое множество. Здесь D– область определения функции.

Функция $f(x)$ называется **вогнутой**, если $-f(x)$– выпуклая функция.

Пусть $f_1(x),\dots,f_m(x)$– последовательность выпуклых функций. Показать, что функции $\sum_{i=1}^{m}\alpha_i f_i(x)\,,\alpha_i \geq 0$, и$max\{f_1(x),\dots,f_m(x)\}$ – выпуклые функции.

Функция $f(x)=x^2$является выпуклой, поэтому функция $f(x)=\sum_{i=1}^{m}\alpha_i x_i^2\,,\alpha_i \geq 0$ является выпуклой. Функции $f_1(x)=x, f_2(x)=-x$ являются выпуклыми, поэтому функция$|x|=max\{x,-x\}$ является выпуклой.

Рассмотрим произвольное семейство выпуклых функций $\{f_\alpha(x)\}_{\alpha\in A}$, функция $f(x)=\underset{\alpha\in A}{max} f_\alpha(x)$ является выпуклой.

Максимальное собственное число симметричной матрицы A является выпуклой функцией матрицы. Действительно, максимальное собственное число симметричной матрицы удовлетворяет равенству Рэлея $\lambda(A)=\underset{x}{max}\frac{(Ax,x)}{(x,x)}_{max}$. Семейство функций $f_x(A)=\frac{(Ax,x)}{(x,x)}$– семейство линейных функций, следовательно, семейство выпуклых функций. Поэтому $\lambda(A)_{max}$– выпуклая функция как максимум выпуклых функций.

Рассмотрим аффинное преобразование $Ax+b$ и выпуклую функцию $f(y)$. Показать, что суперпозиция $f(Ax+b)$ является выпуклой функцией.

Рассмотрим прямую линию $y=x+tv$, пересекающую область определения выпуклой функции $f(y)$, x принадлежит области определения, $v\in R^n$, t принадлежит интервалу $[t_i,t_s], t_i=inf\{t: x+tv\in D\}\,, t_s=sup\{t: x+tv\in D\}$. Доказать, что функция$\phi(t)=f(x+tv)$ является выпуклой функцией от одной переменной на интервале $[t_i,t_s]$.

Функция называется **строго выпуклой**, если

$$f(\lambda x+(1-\lambda)y)<\lambda f(x)+(1-\lambda)f(y),$$

для всех $x\neq y$ из области определения функции, и всех $\lambda\in(0,1)$.

Первый критерий для дифференцируемых функций звучит следующим образом. Дифференцируемая функция –выпуклая функция тогда и только тогда, когда для любых x и y из области определения функции выполняется неравенство:

$$f(y)-f(x)\geq(f'(x),x-y).$$

В неравенстве $f'(x)$– вектор частных производных или градиент. Рис.4 является иллюстрацией первого критерия.

Рис.9. График выпуклой функции и касательная.

Доказательство.

Необходимость.Начнем с неравенства, которое является определением выпуклой функции$f\big(x+\lambda(y-x)\big)\leq(1-\lambda)f(x)+\lambda f(y)$.

От правой и левой части отнимем $f(x)$, в результате получим неравенство

$f\big(x+\lambda(y-x)\big)-f(x)\leq\lambda\big(f(y)-f(x)\big)$.

Разделим левую и правую часть неравенства на положительное λ

$\frac{f\left(x+\lambda(y-x)\right)-f(x)}{\lambda}\leq f(y)-f(x)$.

Вычислим предел от левой и правой частей неравенства

$\lim_{\lambda\downarrow 0}\frac{f\left(x+\lambda(y-x)\right)-f(x)}{\lambda}\leq f(y)-f(x)$.

Из дифференцируемости функции следует окончательное неравенство

$(f'(x),y-x)\leq f(y)-f(x)$.

Достаточность.Начнем с неравенств

$f(y)-f(\lambda x+(1-\lambda)y)\geq\lambda(f'(\lambda x+(1-\lambda)y),y-x)$,$f(x)-f(\lambda x+(1-\lambda)y)\geq-(1-\lambda)(f'(\lambda x+(1-\lambda)y),y-x)$.

Левую и правую часть первого неравенства умножим на $1-\lambda$, левую и правую часть второго неравенства умножим на λ, $\lambda\in[0,1]$. Результаты сложим

$\lambda f(x)+(1-\lambda)f(y)-f(\lambda x+(1-\lambda)y)\geq 0$.

Данное неравенство доказывает, что функция $f(x)$– выпуклая функция.

Лекция 8. Продолжение

Предыдущий критерий называют **критерием первого порядка**. Для того, чтобы им воспользоваться, надо проверить выполнение неравенства для всех пар значений из области определения функции. Очень часто это сложная задача.

Критерий второго порядка звучит следующим образом. Функция $f(x)$– выпуклая функция тогда и только тогда, когда матрица вторых производных неотрицательно определена для всех значений аргумента из области определения функции.

Доказательство.

Необходимость.Воспользуемся формулой Тейлора второго порядка

$$f(y) = f(x) + (f'(x), y - x) + \frac{1}{2}(f''(x)(y - x), y - x) + o(\|y - x\|^2).$$

Здесь и далее, пока это не будет оговорено особо,$\|x\|$ – евклидова норма. В формуле Тейлора остаточный член обладает свойством $\lim\limits_{y \to x} \frac{o(\|y-x\|^2)}{\|y-x\|^2} = 0$. Положим $y = x + \theta h, \|h\| = 1, \theta > 0$ (при достаточно малом h точка y будет принадлежать области определения функции, если область определения – открытое множество). Воспользуемся первым критерием ив результате получим неравенство:

$$\frac{\theta^2}{2}(f''(x)h, h) + o(\theta^2) \geq 0.$$

Разделим обе части неравенства на положительное число θ^2 и перейдем к пределу при $\theta \downarrow 0$. В результате получим неравенство

$$(f''(x)h, h) \geq 0,$$

из которого следует неотрицательная определенность матрицы.

Достаточность.Воспользуемся формулой Тейлора первого порядка с остаточным членом в форме Лагранжа

$$f(y) = f(x) + (f'(x), y - x) + \frac{1}{2}\Big(f''(x)\big(x + \alpha(y - x)\big), x + \alpha(y - x)\Big),$$

где α– некоторое число из интервала [0,1]. Из неотрицательной определенности матрицы вторых производных и первого критерия следует выпуклость функции $f(x)$.

Заметим, если матрица вторых производных положительно определена, то функция $f(x)$– строго выпуклая функция.

Второй критерий является более конструктивным по сравнению с первым критерием.

Используя второй критерий, доказать, что функция $f(x) = (Ax, x) + (b, x) + c$ является выпуклой тогда и только тогда, когда матрица A является неотрицательно определенной.

Неравенство Йенсена. Следующий пример играет важную роль в теории вероятностей и математической статистике. Пусть множество значений случайной величины ξ принадлежит области определения выпуклой функции

функции $f(x)$. Рассмотрим случайную величину $f(\xi)$. Для ее математического ожидания справедливо неравенство $Ef(\xi) \geq f(E\xi)$.

Для доказательства используем выпуклость надграфика выпуклой функции и теорему об опорной гиперплоскости, здесь опорной прямой. Опорную прямую проведем к надграфику функции в граничной точке $\big(E\xi, f(E\xi)\big)$. Пусть уравнение опорной прямой имеет вид

$$a\big(y - f(E\xi)\big) + b(x - E\xi) = 0.$$

Очевидно, что коэффициент $a > 0$. Рассмотрим случайную точку $\big(\xi, f(\xi)\big)$ которая принадлежит надграфику функции. Поскольку любая точка надграфика функции содержится в неотрицательном полупространстве, то выполняется неравенство $a\big(f(\xi) - f(E\xi)\big) + b(\xi - E\xi) \geq 0$. Отсюда следует неравенство $a\big(Ef(\xi) - f(E\xi)\big) \geq 0$. Поскольку $a > 0$, то выполняется неравенство Йенсена.

Арифметическое и геометрическое среднее. Используя вогнутость и возрастание логарифма, доказать, что геометрическое среднее меньше либо равно арифметического среднего.

2.2 Локальные и глобальные минимумы функции

Рассмотрим произвольную функцию $f(x)$ с областью определения S. Точка x_0 называется точкой локального минимума функции $f(x)$, если существует такой шар $B_\varepsilon(x_0)$, что $f(x_0) \leq f(x)$ для всех x, принадлежащих $B_\varepsilon(x_0)$. В этом определении неявно предполагается, что точка x_0– внутренняя точка области определения.

Критерий локального минимума. Точка x_0является точкой локального минимума функции тогда и только тогда, когда вектор частных производных$f'(x_0) = 0$ и матрица частных производных второго порядка $f''(x_0)$– неотрицательно определена, то есть $(f''(x_0)h, h) \geq 0, \forall h$.

Доказательство.

Необходимость.Выберем $\alpha > 0$таким образом, чтобы точки $y_1 = x_0 + \alpha h$ и $y_2 = x_0 - \alpha h$ принадлежали окрестности точки x_0. Для этих точек применим формулу Тейлора первого порядка$f(y) - f(x) = (f'(x_0), y - x) + o(\|y - x\|)$. В результате получим одновременное выполнение двух неравенств

$$(f'(x_0), h) \geq 0 \text{ и } (f'(x_0), h) \leq 0$$

одновременно для любого h. Из одновременного выполнения двух неравенств следует справедливость равенства $(f'(x_0), h) = 0$ для всех h. Отсюда вытекает равенство нулю вектора частных производных$f'(x_0) = 0$.

Для точки y_1 применим формулу Тейлора второго порядка $f(y) - f(x) = (f'(x_0), y - x) + \frac{1}{2}(f''(x_0)(y - x), y - x) + o(\|y - x\|^2)$, в результате получим неравенство

$$\frac{\alpha^2}{2}(f''(x_0)h, h) + o(\alpha^2) \geq 0.$$

Разделив обе части неравенства на $\alpha > 0$ и, вычислив предел при $\alpha \downarrow 0$ от обеих частей, получим неравенство

$$(f''(x_0)h, h) \geq 0, \text{ для всех } h,$$

которое означает неотрицательную определенность матрицы вторых частных производных.

Достаточность. Для y из окрестности точки x_0 применим формулу Тейлора первого порядка с остаточным членом в форме Лагранжа $f(y) - f(x_0) = \frac{1}{2}\left(f''\left(x_0 + \beta(y - x_0)\right)(y - x_0), y - x_0\right)$. Из неотрицательной определенности матрицы вторых частных производных следует неравенство $f(y) - f(x_0) \geq 0$, которое устанавливает, что x_0 является точкой локального минимума.

Выпуклая функция обладает одной важной особенностью. Любой ее локальный минимум является глобальным минимумом.

Действительно, пусть x_0– точка локального минимума выпуклой функции $f(x)$ и существует такая точка x_*, для которой выполняется неравенство $f(x_*) < f(x_0)$. Рассмотрим выпуклую комбинацию $\lambda x_* + (1 - \lambda)x_0$ и выберем λ таким образом, чтобы точка $\lambda x_* + (1 - \lambda)x_0$ принадлежала шару $B_\varepsilon(x_0)$. Так как функция выпукла, то выполняется неравенство $f(\lambda x_* + (1 - \lambda)x_0) \leq \lambda f(x_*) + (1 - \lambda)f(x_0)$. Так как $f(x_*) < f(x_0)$, то $f(\lambda x_* + (1 - \lambda)x_0) < f(x_0)$. Последнее противоречит тому, что точка x_0– точка локального минимума.

Критерий минимума выпуклой функции. Точка x_0 является точкой минимума выпуклой функции $f(x)$ на выпуклом множестве S тогда и только тогда, когда для всех $y \in S$ выполняется неравенство

$$(f'(x), y - x_0) \geq 0.$$

Доказательство.

Необходимость. Используем формулу Тейлора первого порядка

$$f(x_0 + \alpha h) - f(x_0) = \alpha(f'(x_0), h) + o(\alpha), \|h\| = 1, \alpha > 0.$$

Для вывода неравенства

$$\alpha(f'(x_0), h) + o(\alpha) \geq 0$$

разделим левую и правую часть неравенства на α и вычислим от обеих частей предел при $\alpha \downarrow 0$. В результате получим неравенство

$$(f'(x), y - x_0) \geq 0.$$

Достаточность. Неравенство $f(y) - f(x_0) \geq 0$ следует из первого критерия выпуклости функции $f(y) - f(x_0) \geq (f'(x_0), y - x_0)$.

Неравенство выполняется для всех $y \in S$, поэтому точка x_0 является точкой минимума функции $f(x)$.

Если точка x_0– внутренняя точка области определения функции $f(x)$, то мы можем для любого h выбрать α таким образом, что точки $y_1 = x_0 + \alpha h, y_2 = x_0 - \alpha h$ принадлежат области определения. Использование критерия для этих точек устанавливает выполнение двух неравенств

$$(f'(x_0), h) \geq 0, -(f'(x_0), h) \geq 0$$

одновременно. Из одновременного выполнения этих неравенств, следует равенство

$$(f'(x_0), h) = 0,$$

которое выполняется для всех h, поэтому $f'(x_0) = 0$.

Критерий минимума внутренней точки выпуклой функции. Внутренняя точка $x_0 \in S$ будет точкой минимума функции $f(x)$ тогда и только тогда, когда $f'(x_0) = 0$.

Рассмотрим несколько примеров. Пусть $f(x)$– выпуклая функция.

$S = R_+^n$. Точка x_0– точка минимума $\Leftrightarrow f'(x_0) \geq 0, \frac{\partial f(x_0)}{\partial x_i} x_{0,i} = 0$.

$S = \{x: Ax = b\}$. Точка x_0– точка минимума $\Leftrightarrow (f'(x_0), z) = 0, \forall z \in ker\, A$.

Доказать в качестве упражнения.

Рассмотрим задачу вычисления минимума функции на выпуклом множестве S.

Метод проекции субградиента

Задача, которой посвящен этот раздел, заключается в следующем:

$$\min_{x \in S} f(x).$$

Множество S – выпуклое и замкнутое множество, $f(x)$– выпуклая функция. Рассматриваемый в этом разделе метод называется методом проекции градиента, если функция $f(x)$– дифференцируемая функция, или проекции субградиента, если функция $f(x)$– недифференцируемая функция.

Под проекцией точки y на множество S – $P_S(y)$ – понимается решение задачи

$$\min_{x \in S} \|y - x\|.$$

Решение этой задачи может не существовать; например, для множества $S = (0,1)$ и $y = 2$ решения не существует, для множества$S = \{(x_1, x_2): x_1^2 + x_2^2 = 1\}$ и $y = (0,0)$ решений бесконечное множество.

Является справедливой следующая теорема.

Теорема о проекции точки. Если множество S– выпуклое и замкнутое множество, то проекция точки y – $P_S(y)$существует, причем проекция – единственная.

Проекция характеризуется неравенством тупого угла (см. рис.5):

$$\big(y - P_s(y), x - P_s(y)\big) \leq 0, \forall x \in S,$$

которое непосредственно выводится из критерия оптимальности.

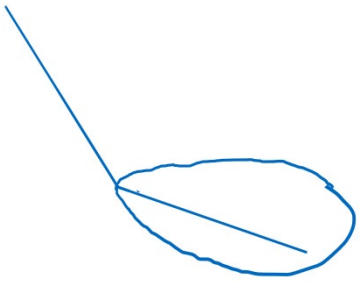


Рис. 5. Неравенство тупого угла.

Проекция обладает следующим важным свойством.

Теорема. Проекция является нерасширяющим отображением, то есть

$$\|P_S(y) - P_S(x)\| \leq \|y - x\|.$$

Для доказательства дважды воспользуемся неравенством тупого угла:

$\big(x - P_S(x), P_S(y) - P_S(x)\big) \leq 0, \big(y - P_S(y), P_S(x) - P_S(y)\big) \leq 0.$

Складывая эти неравенства, получим неравенство:

$\left(y - x - \left(P_S(y) - P_S(x)\right), P_S(y) - P_S(x)\right) \geq 0.$

Рассмотрим два вектора, для которых выполняется неравенство: $(a - b, b) \geq 0$. С помощью несложных преобразований $\|a\|^2 = \|a - b + b\|^2 = \|a - b\|^2 + 2(a - b, b) + \|b\|^2 \geq \|b\|^2$ устанавливается, что $\|a\|^2 \geq \|b\|^2$. Теперь, чтобы доказать теорему, достаточно положить $a = y - x, b = P_S(y) - P_S(x)$.

Метод проекции градиента заключается в генерации последовательности точек:

$$x_{t+1} = P_S\left(x_t - h_t f'(x_t)\right),$$

h_t– выбранная подходящим способом последовательность шагов.

При $S = R^n$

$$x_{t+1} = x_t - h_t f'(x_t).$$

Этот метод называется **методом спуска по градиенту**. Если функция – недифференцируемая функция, то вместо градиента используется субградиент.

Остановимся на определении субградиента выпуклой функции. Для этого вернемся к первому критерию выпуклости:$f(y) - f(x) \geq (f'(x), y - x)$.

Поскольку вектор частных производных (градиент)может не существовать, то заменим его на подходящий вектор g и приведем, исходя из этой замены, определение субградиента.

Субградиентом выпуклой функции $f(x)$ называется всякий вектор $g(x)$, для которого выполняется неравенство:

$$f(y) - f(x) \geq (g(x), y - x).$$

Неравенство должно выполняться для всех y из области определения функции. Множество всех субградиентов называется **субдифференциалом** функции –$\partial f(x)$.

Это множество может быть пустым, содержать один элемент или бесконечное множество элементов. Если функция дифференцируема, то $\partial f(x) = \{f'(x)\}$. Если дифференциал состоит из бесконечного множества элементов, то он является выпуклым множеством. Дифференциал не имеет смысла для точек, в которых выпуклые функции равны минус бесконечности, и дифференциал – пустое множество для точек, в которых функции равны плюс бесконечности. Если точка является внутренней точкой множества точек, в которых $f(x) < \infty$, то дифференциал в этой точке –непустое множество.

Определение.Если $\partial f(x) \neq \emptyset$, то функция субдифференцируемая в точке (x).

Выпуклая функция $y = |x|$– недифференцируемая функция при $x = 0$.

В качестве упражнения доказать, что $\partial|x| = \begin{cases} -1, x < 0 \\ [-1,1], x = 0. \\ 1, x > 0 \end{cases}$

Рассмотрим функцию, которая является максимумом семейства выпуклых функций

$F(x) = max\{f_1(x), f_2(x), \dots, f_m(x)\}$. Если каждая функция семейства – субдифференцируемая функция в точке x, функции семейства – непрерывные функции в точке x, то функция $F(x)$– субдифференцируемая функция в точке x, причем субдифференциал этой функции – выпуклая оболочкаобъединения активной части субдифференциалов функций семейства:$\partial F(x) = conv \bigcup_{i \in A(x)} \partial f_i(x)$, $A(x) = \{j: f_j(x) = F(x)\}$.

Установим связь между константой Липшица и субградиентом.

Теорема 1. Если функция $f(x)$ удовлетворяет условию Липшица: $|f(y) - f(x)| \le L\|y - x\|$, то субградиент равномерно ограничен этой константой: $\|g(x)\| \le L$.

Доказательство. Выберем $y = x + \varepsilon \frac{g}{\|g\|}, \varepsilon > 0$ и подставим в определяющее субградиент неравенство:

$$L\varepsilon \ge |f(y) - f(x)| \ge f(y) - f(x) \ge (g, y - x) = \varepsilon\|g\|.$$

Разделим на $\varepsilon > 0$ и получим необходимое неравенство.

Теорема 2. Если субградиент функции $f(x)$ равномерно ограничен константой L: $\|g\| \le L$, то функция удовлетворяет условию Липшица:

$$|f(y) - f(x)| \le L\|y - x\|.$$

Доказательство. Используем определение субградиента в точке x и неравенство Коши:

$$f(x) - f(y) \le (-g, y - x) \le |(-g, y - x)| \le \|g\|\|y - x\| \le L\|y - x\|.$$

Аналогичным образом используем определение субградиента в точке y:

$$f(y) - f(x) \le L\|y - x\|.$$

Отсюда и из предыдущего неравенства следует доказательство теоремы.

Замена градиентана субградиент в тех точках, в которых градиент не существует, приводит к новой формуле генерации последовательности точек:

$$x_{t+1} = P_S(x_t - h_t g_t), g_t \in \partial f(x_t).$$

В этой формуле присутствует неоднозначность, связанная с выбором элемента из множества $\partial f(x_t)$.

При анализе метода проекции субградиента изучают поведение последовательности $f(x_t)$. Для выпуклых функций, удовлетворяющих условию Липшица, базовый результат содержится в теореме.

Теорема. Пусть выпуклая функция $f(x)$ удовлетворяет условию Липшица: $|f(x) - f(y)| \le L\|x - y\|$, тогда для последовательности x_t, полученной проекцией субградиента, справедливо неравенство:

$$f_T^* - f^* \le \frac{\|x_1 - x_*\|^2 + L^2 \sum_{t=1}^{T} h_t^2}{2 \sum_{t=1}^{T} h_t}.$$

В этом неравенстве $f_T^* = \min\limits_{1 \le t \le T} f(x_t), f^* = \min\limits_{x \in S} f(x), x^* = arg \min\limits_{x \in S} f(x)$.

Доказательство.Используем обозначение $r_t^2 = \|x_t - x^*\|^2$. Рассмотрим разность $r_{t+1}^2 - r_t^2 = \|P_s(x_t - h_t g_t) - x^*\|^2 - \|x_t - x^*\|^2 \le \|x_t - x^* - h_t g_t\|^2 - \|x_t - x^*\|^2 = -2h_t(x_t - x^*, g_t) + h_t^2\|g_t\|^2$.

Из определения субградиента следует неравенство $2h_t\big(f(x_t) - f(x^*)\big) \le r_t^2 - r_{t+1}^2 + h_t^2\|g_t\|^2$. Просуммируем левую и правую часть этого неравенства и в

результате получим неравенство: $2\sum_{t=1}^{T} h_t \left(f(x_t) - f(x^*)\right) \le r_1^2 + \sum_{t=1}^{T} h_t^2 \|g_t\|^2$. Отсюда несложно получить доказательство теоремы.

Поскольку оптимальное значение x^* нам неизвестно, то мы предположим, что $\|x_1 - x_*\| \le R$. При постоянном шаге $h_t = h$, неравенство приобретает вид:

$$f_T^* - f^* \le \frac{R^2 + L^2 T h^2}{2Th}.$$

Минимальное значение правой части неравенства достигается при шаге $h = \sqrt{\frac{R^2}{TL^2}}$. Для такого шага выполняется неравенство: $f_T^* - f^* \le \frac{RL}{\sqrt{T}}$.

Шаг h зависит от числа итераций T и от трудно определяемых констант R и L. Хорошим свойством оценки является ее стремление к нулю при T, стремящемся к бесконечности. Задав погрешность ε, можно определить число итераций $T \ge \frac{R^2 L^2}{\varepsilon^2}$.

Аналитической сложностью алгоритма называется число шагов алгоритма, необходимых для достижения требуемой точности ε. Аналитическая сложность данного алгоритма имеет порядок $\frac{1}{\varepsilon^2}$.

Рассмотрим последовательность шагов $h_t = \frac{1}{t}$, которая не зависит от числа итераций T. Найдем оценку снизу для суммы $\sum_{t=1}^{T} \frac{1}{t}$. Для этого рассмотрим интеграл $\int_1^T \frac{1}{t} dt = \ln T$. Для интеграла справедливо неравенство: $\int_1^T \frac{dt}{t} = \sum_{i=1}^{T-1} \int_i^{i+1} \frac{dt}{t} \le \sum_{i=1}^{T-1} \frac{1}{i} = \sum_{i=1}^{T} \frac{1}{i} - \frac{1}{T}$. Отсюда оценка снизу для суммы $\sum_{t=1}^{T} \frac{1}{t}$: $\sum_{t=1}^{T} \frac{1}{t} \ge \ln T - \frac{1}{T}$. Найдем оценку сверху для суммы $\sum_{t=1}^{T} \frac{1}{t^2}$, для этого используем интеграл $\int_1^T \frac{dt}{t^2} = 1 - \frac{1}{T}$. Для интеграла справедливо неравенство $\int_1^T \frac{dt}{t^2} \ge \sum_{t=1}^{T} \frac{1}{t^2} - 1$. Отсюда следует оценка сверху для суммы $\sum_{t=1}^{T} \frac{1}{t^2} \le 2 - \frac{1}{T}$. Применив оценку снизу и оценку сверху, получим оценку для разности:

$$f_T^* - f^* \le \frac{R^2 + L^2\left(2 - \frac{1}{T}\right)}{2\left(\ln T - \frac{1}{T}\right)},$$

которая стремится к нулю при T, стремящемся к бесконечности. Аналитическая сложность алгоритма при таком выборе шага имеет порядок $exp\left(\frac{1}{\varepsilon}\right)$ что значительно хуже первоначального способа выбора шага.

Более тонкую оценку можно получить для шага $h_t = \frac{R}{L\sqrt{t}}$. Для этого шага является справедливым неравенство:

$$f_T^* - f^* \le RL \frac{3(1 + \ln 2)}{\sqrt{T+2}}.$$

Аналитическая сложность алгоритма эквивалентна $\frac{1}{\varepsilon^2}$; такая же, как и в первоначальном варианте. Отметим, что исчезла зависимость шага от числа итераций, но сохранилась зависимость от констант R и L.

Важно отметить, что если $\sum_{t=1}^{\infty} h_t = \infty, h_t \to 0$, то $f_T^* - f_t \to 0$. Действительно, для любого положительного ε можно выбрать такое N, что для всех $t \geq N$, $h_t \leq \varepsilon$. Отсюда, для всех $T \geq N$ выполняется неравенство: $\sum_{t=1}^{T} h_t^2 \leq \sum_{t=1}^{N-1} h_t^2 + \varepsilon \sum_{t=N}^{T} h_t$. Отсюда следует, что для отношения $\frac{\sum_{t=1}^{T} h_t^2}{\sum_{t=1}^{T} h_t}$ справедливо неравенство $\frac{\sum_{t=1}^{T} h_t^2}{\sum_{t=1}^{T} h_t} \leq \frac{\sum_{t=1}^{N-1} h_t^2}{\sum_{t=1}^{T} h_t} + \varepsilon$. Переходя к пределу при $T \to \infty$, получим равенство нулю предела отношения $\frac{\sum_{t=1}^{T} h_t^2}{\sum_{t=1}^{T} h_t}$, так как ε – произвольное положительное число. Отсюда следует, что$f_T^* - f_t \to 0$.

Лекция 9. Онлайн субградиентный спуск

Вернемся к рассмотрению основной задачи – разработке стратегии вычисления последовательности решений $x_t \in V$, для которой отношение сожаления к периоду $\frac{R_T(x_1,\dots,x_T;u)}{T} = \frac{1}{T}(\sum_{t=1}^{T} l_t(x_t) - \sum_{t=1}^{T} l_t(u))$стремится к нулю при T, стремящемся к бесконечности для любого альтернативного постоянного решения $u \in V$.

Ранее рассматривалась стратегия «Следуй за лидером», которая заключалась в следующем:

$$x_{t+1} = arg\min_{x} \sum_{i=1}^{t} l_i(x),$$

которая в одной из задач привела к успеху. Возникает вопрос: всегда ли такая стратегия приводит к успеху? Для того, чтобы получить отрицательный ответ, рассмотрим пример.

Отрицательный пример. Пусть множество, из которого можно брать решения $V = [-1,1]$, последовательность потерь $l_t(x) = zx + I_V(x)$, $z_1 = -0.5, z_t = \begin{cases} 1, t = 2,4, \\ -1, t = 3,5,\dots \end{cases}$. Для данной функции потерь кумулятивные потери $L_t = \sum_{i=1}^{t} l_i(x) = -0.5x + tI_V(x) + x\sum_{i=2}^{t}(-1)^i = \begin{cases} -0.5x + tI_V(x), t = 1,3,\dots \\ 0.5x + tI_V(x), t = 2,4,\dots \end{cases}$Первое решение x_1- произвольная точка интервала$[-1,1]$,например, ноль. Для четных t $x_{t+1} = -1$, для нечетных t $x_{t+1} = -1$. Для этой стратегии последовательность кумулятивных потерь получается следующей:$L_1 = 1$, $L_T = 1.5T - 0.5, T \geq 2$. Выберем $u = 0$, для этого u сожаление $R(x_1, x_2, \dots, x_T, 0) = 0.5T - 0.5$, поэтому стратегия «Следуй за лидером» не является выигрышной стратегией для данной последовательности потерь.

Далее мы рассмотрим алгоритм типа субградиентного спуска, с помощью которого построим стратегию, которая будет обеспечивать сублинейный рост сожаления для выпуклой, удовлетворяющей условию Липшица, последовательности потерь.

В анализируемом методе в момент времени t определено значение x_t и получен штраф $l_t(x)$, следующее значение вычисляется следующим образом:

$$x_{t+1} = P_v(x_t - h_t g_t), g_t \in \partial l_t(x_t).$$

Назовем эту стратегию online субградиентным спуском и проанализируем сожаление для этой стратегии. Будем следовать описанной ранее схеме. Введем обозначение

$$r_t = \|x_t - u\|^2$$

и рассмотрим разность

$$r_{t+1}^2 - r_t^2 \leq \|x_t - u - h_t g_t\|^2 - \|x_t - u\|^2 = -2h_t(g_t, x_t - u) + h_t^2\|g_t\|^2.$$

Используем определение субградиента для получения следующего неравенства:

$$r_{t+1}^2 - r_t^2 \leq 2h_t\big(l_t(u) - l_t(x_t)\big) + h_t^2\|g_t\|^2.$$

Отсюда оценка сожаления имеет вид:

$$R(x_1, x_2, \dots, u) \le \frac{1}{2}\sum_{t=1}^{T}\frac{r_t^2 - r_{t+1}^2}{h_t} + \frac{L^2}{2}\sum_{t=1}^{T} h_t.$$

Допустим шаг является постоянным h_t и V – ограниченное множество с диаметром D, тогда оценка сожаление приобретает вид:

$$R(x_1, x_2, \dots, u) \le \frac{D}{2h} + \frac{L^2 T h}{2}.$$

Оптимальное значение для шага $h = \sqrt{\frac{D^2}{L^2 T}}$. Для такого шага оценка сверху

$$R(x_1, x_2, \dots, u) \le LD\sqrt{T}.$$

Отсюда следует, что стратегия сублинейная и поэтому выигрышная. Для переменного шага также несложно получить оценку сожаления:

$$R(x_1, x_2, \dots, u) \le \frac{D^2}{2h_T} + \frac{L^2}{2}\sum_{t=1}^{T} h_t.$$

Упражнение. Показать, что переменный шаг $h_t = \frac{1}{\sqrt{t}}$ обеспечивает сублинейность сожаления.

Задача стохастического программирования

К задачам стохастического программирования относится широкий пласт задач, в которых целевые функции и ограничения определены с точностью до случайных параметров.

В прямых методах решения задач стохастического программирования используются стохастические аналоги градиента или субградиента. Приведем определение стохастического градиента или стохастического субградиента. Пусть $F(x)$ – выпуклая функция и $x_0, \dots, x_s$ – случайная последовательность. Случайный вектор ς_s называется **стохастическим градиентом** функции или **стохастическим субградиентом** функции $F(x)$, зависящим от последовательности $x_0, \dots, x_s$, если условное математическое ожидание $E(\varsigma_s / x_0, \dots, x_s) = g(x_s)$, а вектор $g(x_s)$– градиент или субградиент функции $F(x)$ в точке x_s.

Мы ограничимся рассмотрением задач следующего вида:

$$\min_{x \in V} F(x) = \min_{x \in V} E f(x, \xi).$$

Закон распределения случайного элемента ξ либо неизвестен, либо вычисление математического ожидания связано с вычислительными сложностями. Мы предполагаем, что $f(x, \xi)$– выпуклая функция по переменной x при любом значении ξ. Из выпуклости параметрического семейства функций $f(x, \xi)$ следует

выпуклость функции $F(x)$. Мы предполагаем существование градиента или субградиента для каждой функции семейства $f(x,\xi)$. Вместо неизвестного закона распределения предлагается случайная последовательность независимых случайных величин с общим законом распределения. Будем рассматривать $f(x,\xi)$ как параметрическое семейство функций. Пусть $g(x,\xi)$– субградиент функции$f(x,\xi)$:$f(y,\xi)-f(x,\xi)\geq\left(g_\xi(x),y-x\right)$. Рассмотрим математическое ожидание разности. При этом мы предполагаем существование математического ожидания $Eg_\xi(x)$ и используем свойство линейности скалярного произведения. Вернемся к целевой функции задачи стохастического программирования. Из предыдущего неравенства следует неравенство $F(y)-F(x)\geq\left(Eg_\xi(x),y-x\right)$. Таким образом, согласно определению субградиента, вектор $Eg_\xi(x)$ является субградиентом функции целевой функции$F(x)$, поэтому, согласно определению, случайный вектор $g_\xi(x)$– стохастический субградиент функции $F(x)$. По аналогии с детермированным случаем метод генерации последовательности x_t с применением формулы:

$$x_{t+1}=P_V\left(x_t-h_{t+1}g_{\xi_{t+1}}(x_t)\right)$$

называется методом проекции стохастического градиента. Пусть множество V является выпуклым замкнутым и ограниченным множеством с диаметром $\mathbb{R}$, и выполняется почти, наверное, неравенство $\left\|g_\xi(x)\right\|\leq L$. Внесем некоторые изменения, обусловленные случайностью, в проведенный ранее анализ метода проекции субградиента. Введем обозначение $r_t=E\|x_t-x_*\|^2$ здесь x_*– решение задачи, то есть $F(x_*)=\min\limits_{x\in V}F(x)$. Получим оценку сверху разности $r_{t+1}-r_t$. Сначала используем свойство оператора проектирования $r_{t+1}-r_t\leq E\left[\left\|x_t-h_tg_{\xi_{t+1}}(x_t)\right\|^2-\|x_t-x_*\|^2\right]$. Отсюда $r_{t+1}-r_t\leq E\left[-2h_t\left(g_{\xi_{t+1}}(x_t),x_t-x_*\right)+h_t^2L^2\right]$. Воспользуемся тем, что $g_{\xi_t}(x_t)$–субградиент выпуклой функции $f(x,\xi_t)$: $r_{t+1}-r_t\leq E\left[2h_t\left(f(x_*,\xi_t)-f(x_t,\xi_t)\right)+h_t^2L^2\right]$. Осталось вычислить математическое ожидание$E\left(f(x_*,\xi_t)-f(x_t,\xi_t)\right)=EE(f(x_*,\xi_t)-f(x_t,\xi_t)/\xi_{t-1})$. Условное математическое ожидание $E(f(x_*,\xi_t)-f(x_t,\xi_t)/\xi_{t-1})=F(x_*)-F(x_t)$. Подставим вычисленное математическое ожидание в оценку сверху разности и в результате получим $r_{t+1}-r_t\leq 2h_tE\left(F(x_*)-F(x_t)\right)+h_t^2L^2$. Выберем некоторое T. Следующее неравенство втекает из предыдущего неравенства, чтобы в этом убедиться, достаточно просуммировать предыдущее неравенство по t, итак, $2\sum_{t=1}^T h_tE\left(F(x_t)-F(x_*)\right)\leq R+L^2\sum_{t=1}^T h_t^2$. Разделим обе части неравенства на $2\sum_{t=1}^T h_t$, введем обозначение $\nu_t=\frac{h_t}{\sum_{j=1}^T h_j}$, и воспользуемся тем, что функция $F(x)$ – выпуклая функция, в результате получим неравенство:

$$EF(\sum_{t=1}^T \nu_t x_t)-F(x_*)\leq\frac{R^2+L^2\sum_{t=1}^T h_t^2}{2\sum_{t=1}^T h_t}.$$

Для постоянного шага неравенство имеет следующий вид:

$EF\left(\frac{1}{T}\sum_{t=1}^{T}x_t\right)-F(x_*)\leq\frac{R^2+L^2Th^2}{2Th}$. Выберем оптимальное значение для шага, вычисляя минимум правой части неравенства. Оптимальное значение $h_*=\frac{R}{L\sqrt{T}}$. Подставим данный постоянный шаг в правую часть неравенства, чтобы получить следующую оценку:

$$EF\left(\frac{1}{T}\sum_{t=1}^{T}x_t\right)-F(x_*)\leq\frac{RL}{\sqrt{T}}.$$

Из этого неравенства следует: среднее значение $EF\left(\frac{1}{T}\sum_{t=1}^{T}x_t\right)$ стремится к $F(x_*)$ со скоростью $\frac{1}{\sqrt{T}}$.

Альтернативный способ вычисления условного минимума регрессии заключается в следующем. Выбирается достаточно большое значение для T и рассматривается приближенная задача:

$$\min_{x\in V}F_e(x)=\min_{x\in V}\frac{1}{T}\sum_{i=1}^{T}f(x,\xi_i).$$

Отметим, что задача решается для фиксированного набора значений $\xi_1,\ldots,\xi_T$. Связь между T и точностью приближения ε устанавливает следующая теорема.

Теорема. Если дисперсия случайной величины $f(x,\xi)$ равномерно ограничена на множестве V: $\sup_{x\in V}Df(x,\xi)\leq C$, то

$$P\left(\left|F(x)-\frac{1}{T}\sum_{t=1}^{T}f(x,\xi)\right|\leq\varepsilon\right)\geq\sigma, \text{ для } T\geq\frac{C}{(1-\sigma)\varepsilon^2}.$$

Доказательство опирается непосредственно на неравенство Чебышева: $P(|\eta-E\eta|\leq\varepsilon)\geq 1-\frac{D\eta}{\varepsilon^2}$. Среднее $E\frac{1}{T}\sum_{i=1}^{T}f(x,\xi_i)=F(x)$, дисперсия$D\frac{1}{T}\sum_{i=1}^{T}f(x,\xi_i)=\frac{Df(x,\xi)}{T}\leq\frac{C}{T}$.

Для решения приближенной задачи можно использовать хорошо развитые методы решения задач математического программирования. В частности, метод проекции субградиента. Однако, вычисление субградиента может потребовать серьезных вычислительных затрат из-за большого числа слагаемых. Уменьшить вычислительную сложность позволяет прием, к описанию которого мы приступаем. Для дальнейшего нам понадобится следующая теорема.

Теорема. Если $g_i(x)$– субградиенты функций $f(x,\xi_i)$, то $\sum_{i\geq1}\alpha_i f(x,\xi_i)$– субградиент функции $F(x)=\sum_{i\geq1}\alpha_i f(x,\xi_i)$, при условии, что коэффициенты $\alpha_i\geq0$.

Доказательство. Определение субградиента означает выполнение неравенств:

$$f(y,\xi_i)-f(x,\xi_i)\geq(g_i(x),y-x).$$

Дальше умножим каждое неравенство на соответствующий положительный множитель α_iи просуммируем:

$$\sum_{i\geq1}\alpha_i\big(f(y,\xi_i)-f(x,\xi_i)\big)\geq\sum_{i\geq1}\alpha_i(g_i(x),y-x).$$

Свойство линейности скалярного произведения доказывает теорему.

Рассмотрим случайную величину θ, равновероятно распределенную на множестве значений $\{1,\ldots,T\}$, и случайный вектор $g_\theta(x)$. Математическое ожидание случайного вектора $Eg_\theta(x)=\frac{1}{T}\sum_{i=1}^{T}g_i(x)$. Применив теорему,

получим, что субградиент функции $F_e(x)$–$g_e(x)$ равен $Eg_\theta(x)$. Следовательно, $g_\theta(x)$–стохастический субградиент функции $F_e(x)$.

В данном случае метод проекции стохастического субградиента заключается в генерации последовательности:

$$x_{t+1} = P_v\left(x_t - h_{t+1}g_{\theta_{t+1}}(x_t)\right).$$

Случайные величины θ_i– независимые копии случайной величины θ.

Для T итераций при постоянном шаге $h^* = \frac{R}{L\sqrt{T}}$ получим оценку

$$EF_e\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - F_e(x_*) \le \frac{RL}{\sqrt{T}}.$$

Рассмотрим эту же задачу с точки зрения задачи вычисления выигрышной стратегии, например, полученной при помощи онлайн субградиентного спуска. Для этого рассмотрим последовательность потерь $l_t(x) = \alpha_t f(x, \xi_t)$. Сожаление для этой стратегии $R_T(u) = \sum_{t=1}^{T} \alpha_t f(x_t, \xi_t) - \sum_{t=1}^{T} \alpha_t f(u, \xi_t)$– условно сублинейно, то есть сублинейно для фиксированной последовательности случайных величин $\xi_1, \ldots, \xi_t, \ldots$ Вычислим математическое ожидание сожаления $ER_T = E\sum_{t=1}^{T} \alpha_t E(f(x_t, \xi_t)/\xi_1, \xi_2, \ldots, \xi_{t-1}) - \sum_{t=1}^{T} \alpha_t F(u)$. Условное математическое ожидание $E(f(x_t, \xi_t)/\xi_1, \ldots, \xi_{t-1}) = F(x_t)$. Отсюда среднее сожаление$ER_T = \sum_{t=1}^{T} \alpha_t EF(x_t) - \sum_{t=1}^{T} \alpha_t F(u)$. Из выпуклости функции $F(x)$ следует, что среднее сожаление

$$\frac{1}{\sum_{t=1}^{T} \alpha_t} ER_T \ge EF\left(\frac{1}{\sum_{t=1}^{T} \alpha_t}\sum_{t=1}^{T} \alpha_t x_t\right) - F(u).$$

Пример 1. Положим в предыдущем неравенстве$\alpha_t = 1$ и $u^* = argmin\, F\,(u)$, в результате получим теорему.

Теорема. Для любой выигрышной стратегии x_t, используемой для случайной последовательности штрафов $l_t(x) = f(x, \xi_t)$, в которой ξ_t– независимые копии случайной величины ξ, среднее значение $EF\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) \to F(u^*)$,при $T \to \infty$.

Пример 2. Рассмотрим задачу бинарной классификации, точнее, задачу обучения для решающего правила вида: $d(x) = \begin{cases} 1, (l,x) \ge 0 \\ 2, (l,x) < 0 \end{cases}$. Дана обучающая выборка $V = \{(x_i, y_i)\}$ объема N. Элементы $y_i \in \{-1,1\}$. Предполагается, что множества $V_1 = (x_i : y_i = 1)$ и $V_2 = (x_i : y_i = -1)$линейно не разделяются,то есть не существует гиперплоскости $(l,x) = 0$, что $(l,x) \ge 0$ для всех $x \in V_1$ и $(l,x) < 0$для всех $x \in V_2$. Рассмотрим штраф $max\{1 - y(l,x), 0\}$,с помощью которого определяется качество решающего правила для обучающей выборки:

$$F(l) = \frac{1}{N}\sum_{i=1}^{N} max\{1 - y_i(l, x_i), 0\}$$

Таким образом, задача обучения заключается в вычислении минимума функции$F(l)$.

Для использования метода спуска по стохастическому субградиенту требуется вычислить субградиент функции $f_i(l) = max\{1 - y_i(l, x_i), 0\}$. Для этого найдем субградиент функции от одной переменной $\varphi(x) = \max\{x, 0\}$. По

определению субградиента $g_\phi(x) = \begin{cases} 1, x > 0 \\ \alpha, x = 0 \\ 0, x < 0 \end{cases}, \alpha \in [0,1]$. Отсюда субградиент функции $g_i(l) = \begin{cases} -y_i x_i, 1 - y_i(l, x_i) > 0 \\ \alpha(-y_i x_i), 1 - y_i(l, x_i) = 0. \\ 0, 1 - y_i(l, x_i) < 0 \end{cases}$

Знание субградиента позволяет для решения задачи обучения применить описанную выше процедуру спуска по стохастическому субградиенту.

Упражнение к разделу 3. В предыдущем примере заменить функцию штрафа $f_i(l) = \left| y_i - \frac{exp((l,x_i)) - exp(-(l,x_i))}{exp((l,x_i)) + exp(-(l,x_i))} \right|$ и предложить алгоритм обучения.

Задача обучения в общей постановке

Предыдущий раздел завершился рассмотрением задачи обучения в одной из наиболее используемых постановок. Рассматривая задачу обучения в более общей постановке, мы предположим, что у нас имеется решающее правило, точнее класс решающих правил $\varphi_l(x)$, параметризованных вектором l. Мы можем рассмотреть случайный вектор $\xi = \binom{x}{y}$ и переопределить вектор параметров: $l := \binom{l}{-1}$. Функция $f(l, \xi) = [(l, \xi)]^2$. Стохастический градиент $g_\xi(l) = (l, \xi)\xi$. Проекция вектора z на множество V вычисляется достаточно просто: $\left(P_V(z)\right)_i = \begin{cases} z_i, i \neq d+1 \\ -1, i = d+1 \end{cases}$. Эта задача непосредственно связана с задачей линейного прогноза ненаблюдаемой случайной величины y по наблюдаемым случайным величинам x.

Рассмотренный в этом разделе подход к обучению называется методом минимизации эмпирического риска. Основной смысл этого метода заключается в замене целевой функции задачи обучения $R(x) = Ef(x, \xi)$ на функцию $R_e(x) = \frac{1}{T}\sum_{t=1}^{T} f(x, \xi_t)$.Эта замена связана с тем, что закон распределения случайной величины ξ при решении задачи обучения неизвестен. Вместо него используется обучающая выборка, с помощью которой можно вычислить эмпирический риск. Понимая, что точное решение невозможно, задача обучения рассматривается как задача вычисления минимума функции $F_e(x)$. Важно найти оценку для абсолютной величины разности $|F(x^*) - F(x_T)|$, где $x^* = arg\min_{x \in V} F(x), x_T = arg\min_{x \in V} F_e(x)$, например, в виде $P(|F(x^*) - F_e(x_T)| \geq \varepsilon) \leq \sigma$. Ранее подобная оценка была найдена с использованием равномерной интегрируемости и неравенства Чебышева. В следующем разделе мы рассмотрим использование мартингалов для получения искомой оценки.

6. Выпуклый анализ для сильной выпуклости

Дадим определение сильной выпуклости.

Определение. Функция $f(x)$ называется μ**-сильно выпуклой**, если для нее существует оценка снизу разности $f(y)-f(x)$:

$$f(y)-f(x)\geq(g,y-x)+\frac{\mu}{2}\|y-x\|^2.$$

Является очевидным утверждение.

Теорема. Пусть $f(x)-$ функция выпуклая и дважды дифференцируемая. Если выполняется неравенство $(f''(x)y,y)\geq\mu\|y\|^2$, для всех x, то она μ-сильно выпуклая функция.

Добавим, что в этом случае $g=f'(x)$.

Для доказательства достаточно воспользоваться формулой Тейлора второго порядка $f(y)-f(x)=(f'(x),y-x)+\frac{1}{2}(f''(\lambda x+(1-\lambda)y)y-x,y-x)$, $\lambda\in(0,1)$. Из формулы Тейлора вытекает необходимое неравенство.

В качестве примера рассмотрим квадратичную функцию $f(x)=(A(x-a),x-a)$. Для этой функции формула Тейлора будет иметь вид: $f(y)-f(x)=2(A(x-a),y-x)+(A(x-a),x-a)$. Пусть $\mu>0-$ наименьшее собственное число положительной и симметричнойматрицы A, тогда квадратичная функция $f(x)$ будет 2μ- сильно выпуклой.

Рассмотрим в качестве примера недифференцируемую в нуле функцию $f(x)=|x|+x^2$. Субградиентом в нуле этой функции является любое α из интервала $(-1,1)$. Оценка снизу определяется неоднозначно $|x|+x^2\geq\alpha x+x^2$, поскольку неоднозначно определяется субградиент.

Следующее свойство также является простым свойством.

Теорема. Если функция $f_1(x)-\mu_1$-сильно выпуклая, функция $f_2(x)-\mu_2$-сильно выпуклая, то сумма $f_1(x)+f_2(x)-(\mu_1+\mu_2)$-сильно выпуклая функция.

Доказательство основывается на том, что любой субградиент суммы равен сумме субградиентов слагаемых.

Online субградиентный спуск для сильно выпуклых функций

Рассмотрим основную задачу, которая изучается в этом пособии. Имеется последовательность штрафов $l_t(x)$. Функции $l_t(x)$определены на выпуклом и замкнутом множестве V, на котором они являются сильно выпуклыми функциями: $l_t(y)-l_t(x)\geq(g_t,y-x)+\frac{\mu_t}{2}\|y-x\|$

Рассмотримonlineсубградиентный спуск, порождающий последовательность стратегий:

$$x_{t+1}=P_V(x_t-h_tg_t).$$

Применив методику, изложенную ранее можем доказать следующую оценку для разности $l_t(x_t)-l_t(u)$: $l_t(x_t)-l_t(u)\leq\left(\frac{1}{2h_t}-\frac{\mu_t}{2}\right)\|x_t-u\|^2-\frac{1}{2h_t}\|x_{t+1}-u\|^2+\frac{h_t}{2}\|g_t\|^2$. Выберем шаг таким образом, чтобы $\frac{1}{2h_t}-\frac{\mu_t}{2}=\frac{1}{2h_{t-1}}$, положив $h_1=\frac{1}{\mu_1}$. Применив метод математической индукции нетрудно доказать формулу для шага:

$$h_t=\frac{1}{\mu_1+\mu_2+....+\mu_t}.$$

С этим шагом оценка разности будет иметь вид: $l_t(x_t)-l_t(u)\leq\frac{1}{h_{t-1}}\|x_t-u\|^2-\frac{1}{2h_t}\|x_{t+1}-u\|^2+\frac{h_t}{2}\|g_t\|^2,t=2,...,T;l_1(x_1)-l_1(u)\leq$

Отсюда $\leq-\frac{\mu_1}{2}\|x_2-u\|^2+\frac{1}{2\mu_1}\|g_1\|^2.$

несложно получить оценку для сожаления:

$$R(u)=\sum_{t=1}^{T}(l_t(x_t)-l(u))\leq-\frac{\mu_1}{2}\|x_2-u\|^2+\frac{1}{2}\sum_{i=2}^{T}\left(\frac{1}{h_{i-1}}\|x_i-u\|^2-\frac{1}{h_i}\|x_{i+1}-u\|^2\right)+$$

$$+\frac{1}{2}\sum_{i=1}^{T}h_i\|g_i\|^2$$

После несложных преобразований получаем оценку для сожаления:

$$R(u)=\sum_{t=1}^{T}(l_t(x_t)-l(u))\leq-\frac{1}{2h_T}\|x_{T+1}-u\|^2+\frac{1}{2}\sum_{i=1}^{T}h_i\|g_i\|^2\leq\frac{1}{2}\sum_{i=1}^{T}h_i\|g_i\|^2.$$

Предположим, что $\mu_t\geq\mu>0$ и потери удовлетворяют условию Липшица с общей константой Липшица $-L$.При этих предположениях оценка сожаления будет выглядеть следующим образом:

$$R(u)=\sum_{t=1}^{T}(l_t(x_t)-l_t(u))\leq\frac{L^2}{2\mu}(1+\ln T).$$

Данная оценка сожаления для сильно выпуклых потерь лучше, чем оценка для выпуклых потерь, полученная ранее. Поскольку предыдущая оценкаимела порядок роста $0(\sqrt{T})$, а данная оценка имеет порядок роста $O(\ln T)$. Но в обоих случаях алгоритм online субградиентного спуска является сублиненйным алгоритмом. Такой результат для сильно выпуклых функций был предсказуем, поскольку оценка снизу для сильно выпуклой функции более тонкая. Для выпуклой функции оценка снизу является линейной, для сильно выпуклой функции – квадратичной.

Упражнения к разделу

1. Пусть потери $l_t(x) = \|x - y_t\|^2$. Показать, что $l_t(x)$– сильно выпуклые функции, определить μ и показать, что алгоритм on-line градиентного спуска при $x_0 = 0$ совпадает с алгоритмом «следуй за лидером».

2. Для $l_t(x)$ из предыдущего упражнения вычислить оценку сверху для сожаления.

3. Для двойственной нормы доказать равенство $\|l\|_p^* = \|y\|_q$, в котором $\frac{1}{p} + \frac{1}{q} = 1$.

Лекция 10. Машинное обучение. Оптимизация. Стохастический анализ

Мартингалы

Напомним определение мартингала. Рассмотрим стохастический базис $\langle \Omega, (F_t)_{t=0}^T, F, P \rangle$. Последовательность σ - алгебр F_t удовлетворяет условиям:

$$F_0 = \sigma(\Omega, \bar{\Omega})\text{– тривиальная алгебра,}$$
$$F_t \subseteq F_{t+1}, F_t \subseteq F.$$

Определение. Случайная последовательность X_t называется **адаптированной последовательностью** относительно стохастического базиса $\langle \Omega, (F_t)_{t=0}^T, F, P \rangle$, если для всех t случайная величина X_t измерима относительно σ - алгебры F_t, случайная последовательность является **предсказуемой последовательностью** , если X_t - измеримая случайная величина относительно σ - алгебры F_{t-1}, для всех t.

Определение. Адаптированная последовательность X_t называется **мартингалом**, если выполняются условия:

$$E|X_t| < \infty,$$
$$E(X_t/F_{t-1}) = X_{t-1}.$$

Условие $E|X_t| < \infty$ гарантирует существование условного математического ожидания $E(X_t/F_{t-1})$.

Рассмотрим мартингал с конечным числом членов $-T < \infty$, для мартингала справедливо равенство: $X_t = E(X_T/F_t)$ и мартингал X_t является равномерно интегрируемой последовательностью. Можно поступить иначе. Рассмотрим абсолютно интегрируемую случайную величину ξ. Построим случайную последовательность X_t следующим образом:

$$X_t = E(\xi/F_t).$$

Данная последовательность является мартингалом.

Риск, эмпирический риск и сожаление

Вернемся к исходной задаче минимизации риска

$$\min F(l) = \min Ef(l, \xi).$$

Среднее вычисляется по неизвестному закону распределения случайного элемента ξ. Как уже отмечалось, для решения этой задачи можно использовать метод online градиентного или online субградиентного спуска, который неотличим от метода спуска по стохастическому градиенту или стохастическому субградиенту.

Допустим, мы применяем некоторый алгоритм, который генерирует решения $l_1, \ldots, l_t, \ldots$ Относительно решений, можно утверждать, что они предсказуемые относительно потока σ-алгебр $F_t = \sigma(\xi_1, \xi_2, \ldots, \xi_t)$ Сожаление, связанное с рассматриваемой задачей

$R(l)=\sum_{i=1}^{T} f(l_i,\xi_i)-\sum_{i=1}^{T} f(l,\xi_i)$. Напомним, что эмпирический риск вычисляется по формуле:

$$F_e(l)=\frac{1}{T}\sum_{i=1}^{T} f(l,\xi_i).$$

Рассмотрим последовательность $Z_t : Z_0 = 0, \Delta Z_t = F(l_t) - f(l_t, \xi_t)$. Потребуем, чтобы случайные величины ΔZ_t были абсолютно интегрируемые. Вычислим условное математическое ожидание $E(\Delta Z_t / F_{t-1}) = E(F(l_t) - f(l_t, \xi_t) / F_{t-1})$. При вычислении условного математического ожидания учтем, что случайная величина ξ_t не зависит от F_{t-1}, случайная величина $l_t - F_{t-1}$-измеримая. Далее

$$E(\Delta Z_t / F_{t-1}) = F(l_t) - Ef(l_t, \xi_t) = 0.$$

Следовательно, последовательность Z_t является мартингалом. Потребуем дополнительно равномерную ограниченность для приращений: $|\Delta Z_t| \le C$, хотя бы для $t = 1,2,...,T$, тогда $P(Z_T < \varepsilon) \ge 1 - \exp\left(-\frac{\varepsilon^2}{2C^2T}\right)$. Приравняв $\exp\left(-\frac{\varepsilon^2}{2C^2T}\right) = \delta$, получим равенство: $\varepsilon = C\sqrt{2T\ln\frac{1}{\delta}}$. Отсюда $P\left(Z_T < C\sqrt{2T\ln\frac{1}{\delta}}\right) \ge 1 - \delta$.

Подставим в данное неравенство Z_T, в результате получим $P\left(\sum_{i=1}^{T}(F(l_i) - f(l_i,\xi_i)) < C\sqrt{2T\ln\frac{1}{\delta}}\right) \ge 1 - \delta$. Воспользуемся определением сожаления $\sum_{i=1}^{T} f(l_i,\xi_i) = R(u) + \sum_{i=1}^{T} f(u,\xi_i)$ и из предыдущего неравенства получим $P\left(\sum_{i=1}^{T} F(l_i) \le R(u) + \sum_{i=1}^{T} f(u,\xi_i) + C\sqrt{2T\ln\frac{1}{\delta}}\right) \ge 1 - \delta$. Рассмотрим последовательность $\overline{Z}_t = \sum_{i=1}^{t}(F(u) - f(u,\xi_i))$, эта последовательность имеет независимые приращения $\Delta Z_t = F(u) - f(u,\xi_i)$. Математическое ожидание $E\Delta Z_t = 0$, следовательно, последовательность $\overline{Z}_t$ –мартингал. При том же предположении о равномерной ограниченности приращений

получаем неравенство $P\left(\sum_{i=1}^{T} f(u,\xi_i) \le TF(u) + C\sqrt{2T\ln\frac{1}{\delta}}\right) \ge 1-\delta$. Из очевидного неравенства для случайных событий $P(AB) \ge P(A) + P(B) - 1$ следует неравенство $P\left(\sum_{i=1}^{T} F(l_i) \le R(u) + TF(u) + 2C\sqrt{2T\ln\frac{1}{\delta}}\right) \ge 1-2\delta$. Разделим обе части неравенства на T, в результате получим:

$$P\left(\frac{1}{T}\sum_{i=1}^{T} F(l_i) \le \frac{R(u)}{T} + F(u) + 2C\sqrt{\frac{2\ln\frac{1}{\delta}}{T}}\right) \ge 1-2\delta.$$

Неравенство справедливо для любыхu. Пусть u^* – решение задачи, то есть $F(u^*) = \min F(u)$, тогда неравенство приобретает вид:

$$P\left(\frac{1}{T}\sum_{i=1}^{T} F(l_i) \le \frac{R(u^*)}{T} + F(u^*) + 2C\sqrt{\frac{2\ln\frac{1}{\delta}}{T}}\right) \ge 1-2\delta.$$

Если $F(l)$– выпуклая функция, то выполняется неравенство:

$$P\left(F\left(\frac{1}{T}\sum_{i=1}^{T} l_i\right) \le \frac{R(u^*)}{T} + F(u^*) + 2C\sqrt{\frac{2\ln\frac{1}{\delta}}{T}}\right) \ge 1-2\delta.$$

Если алгоритм является сублинейным, то есть $\lim_{T\to\infty}\frac{R(u)}{T} = 0$, то, выбрав достаточно большое значение для T и взяв в качестве решения среднее $\overline{u} = \frac{1}{T}\sum_{i=1}^{T} l_i$, мы можем с вероятностью $1-2\delta$ оказаться сколь угодно близко от оптимального значения $F(u^*)$.

Определенный интерес представляет случай, когда множество L, из которого выбирается решение, конечно, то есть $L = \{l_1, l_2, \ldots, l_m\}$. При выборе оптимального решения используется выборка, состоящая из T независимых и одинаково распределенных случайных величин.

Эта задача возникает, например, в следующей схеме принятия решения. Часть выборки используется для вычисления последовательности решений с помощью сублинейного алгоритма, другая часть выборки используется для выбора из найденных решений решения, минимизирующего эмпирический риск. Оценим качество такого двухэтапного метода вычисления оптимального

решения. Обозначим через $\bar{l} = \arg\min_{l \in L} R_e(l)$ оптимальное решение. Напомним, что $F_e(l) = \frac{1}{T}\sum_{i=1}^{T} f(l, \xi_i)$. Математическое ожидание $EF_e(l) = F(l)$. По-прежнему предположим, что $|F(l_i) - f(l_i, \xi_i)| \le C$. Рассмотрим вероятность $P\left(\exists i : |F(l_i) - F_e(l_i)| \ge \frac{\varepsilon}{2}\right)$. Для этой вероятности справедлива оценка сверху: $P\left(\exists i : |F(l_i) - F_e(l_i)| \ge \frac{\varepsilon}{2}\right) \le \sum_{i=1}^{k} P\left(|F(l_i) - F_e(l_i)| \ge \frac{\varepsilon}{2}\right)$. Применим те же рассуждения, что и раньше, и в результате получим оценку сверху:

$$P\left(\exists i : |F(l_i) - F_e(l_i)| \ge \frac{\varepsilon}{2}\right) \le k \exp\left(-\frac{\varepsilon^2 T}{2C^2}\right).$$

В результате для всех $l \in L$, в том числе и для $\bar{l}$, с вероятностью, не меньшей чем $1 - \delta$ будет выполняться неравенство:

$$|F(l) - F_e(l)| \le C\sqrt{\frac{2\ln\frac{k}{\delta}}{T}}\ .$$

Сопоставив данное неравенство с предыдущим неравенством и разделив поровну выборку объема T, на обучающую и контрольную, мы в результате получим неравенство:

$$F(\bar{l}) \le F(l^*) + \frac{2R(l^*)}{T} + 6C\sqrt{\frac{\ln\left(\frac{T}{2\delta}\right)}{T}},$$

которое будет выполняться с вероятностью, не меньшей, чем $1 - \delta$. Данный подход следует использовать для случая, когда риск не является выпуклой функцией.

Упражнения

1. Доказать, что F_τ - σ-алгебра. Здесь и далее τ – момент остановки.
2. Константа k является моментом остановки и $F_\tau = F_k$, если $\tau = k$.
3. Случайное событие $\{\tau = n\} \in F_\tau$, $\forall n$.
4. Пусть $B \in F_\tau$, тогда $B \cap \{\tau < n\} \in F_\tau$.

Литература

1. М. Де Грот. Оптимальные статистические решеня. Мир, М ., 1974, 491 с.
2. В. Ковалевский. Задача распознавания образов с точки зрения математической статистики. Наукова Думка, Киев 1965, с. 3–41.
3. Chow, C. Statistical independence and threshold functions. IEEE Transactions on Computers, 14, 1965, p. 247–252.
4. R. Duda, P. Hart, D. Stork. Pattern classification. P. Willey and Sons, New York, 2007, 677 p.
5. М. Шлезингер, В. Главач. Десять лекций по распознаванию образов. Kluwer Academic Publishers, Бостон, 2005, 546 с.
6. J. Neyman and E. Pearson, E. On the problem of the most efficient tests of statistical hypotheses. Phil. Trans. Royal Soc. London, 231, 1933, p. 289–337.
7. B. Rudloff. A generalized Neyman-Pearson lemma for hedge problems in incomplete markets/ B. Rudloff //Workshop „Stochastische Analysis“,27.09.2004 – 29.09.2004, p. 241-249.
8. А. Ширяев. Статистический последовательный анализ. Наука, 1969, 231 с.
9. Г. Белявский, В. Сибирцев. Способы решения некоторых задач оптимальной обработки и распознавания изображений// в сб. Обработка и распознавание сигналов, ИК АНУССР,Киев, 1975, с 3-17.
10. Э. Леман. Теория точеного оценивания. Наука, 1991, 444 с.
11. В. Ковалевский. Методы оптимальных решений в распознавании изображений. Наука, 1976, 328 с.
12. М. Шлезингер. Взаимосвязь обучения и самообучения в распознавании образов. Кибернетика, 2, 1968, с 81-88.
13. H. Steinhaus . Sur la division des corps materiels en parties. Bull. Acad. Polon. Sci., C1. III vol IV, 1956, p. 801—804.
14. J. Bezdek, J. Keller, R. Krishnapuram and N. Pal (). Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. Springer, 1999, 397 p.
15. Г. Белявский. Метод линейных подпространств в распознавании образов//в сб.Распознавание образов, ИК АНУССР, Киев, 1975, с. 10-21.
16. T. Anderson, T. and R. Bahadur. Classification into two multivariate normal distributions with different covariance matrices// Annals of Mathematical Statistics, 33, 1962, p. 420–431.
17. М. Шлезингер. Исследование одного класса распознающих устройств// Автоматика, 2, 1972, с. 38–42.
18. М. Шлезингер Синтез линейного решающего правила для одного класса задач. Издательство академии наук, 5, 1972, с. 157–160.
19. Б. Козинец. Рекуррентный алгоритм разделения выпуклых оболочек двух множеств// в сб. Алгоритмы обучения распознаванию образов, Советское радио, Москва, 1973, с 43-50.
20. М. Айзерман, Е. Браверманн, Л. Розоноер. Метод потенциальных функций в теории обучения машин. Наука, 1970, 423 с.

21. G. Belyavskiy, V. Misyura, E. Puchkov. Prediction intervals for time series using neural networks based on wavelet-core// Far East Journal of Mathematical Sciences, 100,V 3, 2016, c. 413-425.
22. Е. Пучков, Г. Белявский. Разработка методов динамического построения искусственной нейронной сети на основе ряда Вольтера и вейвлет-ядра//Вестник государственного университета путей сообщения, 3, 2015, с. 127-131.
23. В. Вапник, Червоненкис. Теория распознавания образов. Статистические проблемы обучения, Наука, 1974, 297 с.
24. G. Lugossi. Lecture on prediction of individual sequences. Presented at 2001, a statistics Odissey Center Emile Borel institute Henri Poincare, 95
25. F. Orabona. A modern introduction to online lerning. Arxiv:1912.132 13 V 2, 21 may 2020,117.
26. S. Boyd, L. Vanderberghe. Convex optimization. Cambridge, 2009, 716.
27. Shapiro A., Dentcheva D., Ruszczynski A. Lecture on stochastic programming. Modelingand theory. MPS-SIAM series on Optimization, 2014.
28. Shalev-Shvartz, Ben-David. Understanding machine learning. Cambridge, 2014, 397.
29. P. Baldi, L. Mazliak. Martigales and Markov Chains. Chapman, 2002, 192.

10 Лекций. Презентация

Методы оптимизации для машинного обучения

Проф. Белявский Г.И.

Ростов-на-Дону, 2022

Лекция 1

Объект анализа характеризуется парой $(x, y) \in X \times Y$

x – наблюдаемый параметр, y – скрытый параметр

D – множество решений

Отображение $l: X \to D$ называется детерминированной решающей функцией или байесовским решающим правилом

Лекция 1

$W: Y \times D \to R$ – штрафная функция

$W(y, d)$ – величина штрафа за принятое решение d при значении скрытого параметра равном y.

Например, $W(y, d) = \|y - d\|$

$\langle \Omega, F, P \rangle$ - вероятностное пространство, $x: \Omega \to X, y: \Omega \to Y$ - случайными величины

Лекция 1

$R(l) = EW\big(y, l(x)\big)$- средний риск, задача: $\min\limits_{l \in L} R(l)$

$R(l) = EE\big(W\big(y, l(x)\big)/F_x\big) = E\big(\, r(l,x)\big)$, $r(l,x) = E\big(W\big(y, l(x)\big)/F_x\big)$- условный средник риск

Решение:

$$l(x) = arg \min_{d \in D} r(d,x)$$

Пусть X и Y содержат не более чем счетное число элементов, тогда

$r(l,x) = \sum_{u \in Y} W\big(u, l(x)\big)P(u/x), \quad P(u/v) = \frac{P(u,v)}{P(v)}$

Лекция 1

$$l(x) = arg \min_{d \in D} \sum_{u \in Y} W(u,d) P(u/x)$$

Формула Байеса: $P(u/v) = \frac{P(v/u)P(u)}{P(v)}$

Байесовское решающее правило:

$$l(x) = arg \min_{d \in D} \frac{1}{P(x)} \sum_{u \in Y} W(u,d) P(x/u) P(u).$$

Лекция 1

Разбиение $R^n = \bigcup_{i=1}^{|D|} Q_i, P(Q_i \cap Q_j \neq \emptyset) = 0, i \neq j$

$l(x) = i, \forall x \in Q_i$

Множества

$$Q_i = \left\{ x: \sum_{r=1}^{|Y|} (a_{i,r} - a_{j,r}) p(x/r) \leq 0 \right\}$$

$a_{i,j} = W(j,i)P(j)$

Лекция 1

$$Q_1 = \left\{x: ln\frac{p(x/1)}{p(x/2)} \leq \bar{\theta}\right\},\ Q_2 = \left\{x: ln\frac{p(x/1)}{p(x/2)} > \bar{\theta}\right\}$$

$$p(x/1) = \frac{1}{\sqrt{(2\pi)^n |C|}} exp\left(-\frac{1}{2}(C^{-1}(x - m_1), x - m_1)\right)$$

$$p(x/2) = \frac{1}{\sqrt{(2\pi)^n |C|}} exp\left(-\frac{1}{2}(C^{-1}(x - m_2), x - m_2)\right)$$

$$Q_1 = \{x: (C^{-1}(m_1 - m_2), x) \leq \bar{\theta}\}\ ,\ Q_2 = \{x: (C^{-1}(m_1 - m_2), x) > \bar{\theta}\}$$

Минимаксное решающее правило:

$$l(x) = arg \min_{q \in S_m} \max_{p \in S_n} \sum_{j=1}^{m} \sum_{i=1}^{n} q_j W(i,j) P(x/i) p_i$$

Матричная игра с матрицей проигрышей:

$$R(x) = \left(W(i,j) P(x/i) \right)$$

Лекция 2.

Задача линейного программирования:

$$max \sum_{i=1}^{m} z_i$$

$$(R(x)z)_i \leq 1, \quad i = 1,2, \dots, n$$

$$z_i \geq 0$$

Оптимальное решающее правило:

$$l(x) = \frac{1}{\sum_1^m z_i^*(x)} z^*(x)$$

Двойственная задача заключается в вычислении

$$min \sum_{i=1}^{n} u_i$$

при ограничениях:

$$(R^T(x)u)_i \geq 1, \;\; i = 1,2, \dots, n$$

$$u_i \geq 0.$$

Наихудшее маргинальное распределение: $p_i = \frac{1}{\sum_{j=1}^{n} u_j^*} u_i^*$

Пример

Штрафная функция $W(u,d) = \begin{cases} 0, u = d \\ 1, u \neq d \end{cases}$

Задача:

$$max(z_1 + z_2), p(x/1)z_2 \leq 1$$
$$p(x/2)z_1 \leq 1, z_1 \geq 0, z_2 \geq 0$$

Решение:

$$l(x) = \begin{pmatrix} \dfrac{p(x/1)}{p(x/1) + p(x/2)} \\ \dfrac{p(x/2)}{p(x/1) + p(x/2)} \end{pmatrix}$$

Лекция 2

Выборка: $V=\{x_1,\ldots,x_k\}$

$X=\{\nu_1,\nu_2,\ldots,\nu_m\}$ – конечное множество и множество

$Y=\{y_1,y_2,\ldots,y_n\}$- конечное множество

Эмпирическая вероятностная мера на множестве X:

$P_X(\nu_i)=\frac{1}{k}\sum_{j=1}^{k} I_{\nu_i}(x_j)$

Приближенное равенство: $P_X(\nu_i)\approx\sum_{j=1}^{n} P(\nu_i/y_j)p_j$

Лекция 2

Оптимизационная задача:

$$\min_{p} \left\| P_X - P_{X/Y} p \right\|,$$

$$\sum_{i=1}^{n} p_i = 1, p_i \geq 0$$

Вектор $P_X = \left(P_X(\nu_i)\right)$ и матрица $P_{X/Y} = \left(P\left(\nu_i / y_j\right)\right)$

Эквивалентная задача:

«Вычисление вектора с минимальной нормой в выпуклой оболочке, натянутой на множество векторов $g_i = P_X - P^i_{X/Y}$»

$P^i_{X/Y}$- й столбец матрицы $P_{X/Y}$

Решение задачи: $g^* = \sum_{i=1}^{n} g_i p_i^*$

Маргинальное распределение:

$$P(y_i) = p_i^*$$

Задача вычисления максимума логарифма функции правдоподобия:

$$\max_{\{p\}} \sum_{i=1}^{k} ln \sum_{j=1}^{n} p_j p(x_i/y_j)$$

Необходимые условия максимума:

$$p_k = \frac{1}{N} \sum_{i=1}^{N} p(y_k/x_i)\,,$$

$$p(y_k/x_i) = \frac{p(x_i/y_k)p_k}{\sum_{j=1}^{n} p(x_i/u_j)p_j}, k = 1,2,\ldots,n.$$

Метод Гаусса-Зейделя:

$$p^t\left(\frac{y_k}{x_i}\right) = \frac{p\left(\frac{x_i}{y_k}\right)p_k^{t-1}}{\sum_{j=1}^{n} p\left(\frac{x_i}{y_j}\right)p_j^{t-1}}, \quad p_k^t = \frac{1}{N}\sum_{i=1}^{N} p^t\left(\frac{u_k}{x_i}\right),$$

$$k = 1,2,\ldots,n.$$

Задача Неймана-Пирсона

Решающее правило:

$$l(x) = \begin{cases} 1, x \in D_1 \\ 2, x \in D_2 \end{cases}, X = D_1 \cup D_2, D_1 \cap D_2 = \emptyset$$

Задача Немана-Пирсона:

$$\max_{D_2} P(x \in D_2/2)$$

$$P(x \in D_2/1) \leq \alpha$$

Лемма Неймана-Пирсона. Пусть P и Q – две меры на измеримом пространстве (Ω, F). Пусть мера P – абсолютно непрерывна по отношению к мере Q. Тогда для фиксированного λ справедливо утверждение. Если $A \in F$ таково, что $Q(A) \leq Q\left(A^{\lambda}\right)$, то $P(A) \leq P\left(A^{\lambda}\right)$, где $A^{\lambda} = \left\{\omega \in \square : \frac{dP}{dQ} > \lambda\right\}$.

Положим

$$P(A) = P(x \in A/2) \text{ и } Q(A) = P(x \in A/1)$$

Решаем уравнение:

$$Q(A^{\lambda}) = \alpha$$

Определяем

$$D_2 = A^{\lambda^*}$$

λ^* - решение уравнения

Задача Неймана-Пирсона для двух нормальных законов

с общей ковариационной матрицей C и математическими ожиданиями: m_1, m_2

Решение задачи:

$$D_2 = \{x: (g,x) < \lambda^*\}, g = C^{-1}(m_1 - m_2)$$

$$\lambda^* = \mu_1 + \sigma_1 \Phi^{-1}(\alpha), \sigma_1^2 = (Cg, g), \mu_1 = (g, m_1),$$

$\Phi(x)$ – функция Лапласа

Задача Неймана- Пирсона и задача о рюкзаке.

Пусть

$X = \{v_1, \ldots, v_i, \ldots, v_k\}, Y = \{1,2\},$

$P(x = v_i/1) = p_i^1$ и $P(x = v_i/2) = p_i^2$

Плотность

$$\frac{dP}{dQ}(v_i) = \frac{p_i^2}{p_i^1}$$

Будем считать, что плотность – неубывающая последовательность

Характеристическое свойство множества A^{λ}

Если $v_i \in A^{\lambda}$ и $i \geq 2$, то элементы $v_1, \ldots, v_{i-1}$ также принадлежат множеству A^{λ}.

Алгоритм, использующий утверждение.

begin

if $p_1^1 > \alpha$ then begin $D_2 = \emptyset$; stop end; add v_1 to D_2 ;

$s = p_1^1 + p_2^1$; $i = 2$;

do while ($s \leq \alpha$)

add v_i to D_2;

$i = i + 1$; $s = s + p_i^1$;

end while; $s = s - p_i^1$;

if $s < \alpha$ then print (“Algorithm cannot solve problem”) else print(D_2);

end.

Если алгоритм, использующий лемму Неймана-Пирсона, не решает задачу, то возможны два варианта поведения.

Первый вариант – скорректировать α, а именно, добавить оператор $if(\alpha - s) > \left(s + p_i^1 - \alpha\right) then\ begin\ \alpha := s + p_i^1; Add\ v_i to D_2 end\ else\ \alpha := s.$

Второй вариант

Определим переменные

$$z_i = \begin{cases} 1, v_i \in D_2 \\ 0, v_i \in D_1 \end{cases}$$

Задача о рюкзаке:

$$max \sum_{i=1}^{k} z_i p_i^2 , \sum_{i=1}^{k} z_i p_i^1 \leq \alpha, z_i \in \{0,1\}$$

$$D_2 = \{v_i : z_i = 1\}$$

Нечеткая задача о рюкзаке

$$max \sum_{i=1}^{k} z_i p_i^2 , \sum_{i=1}^{k} z_i p_i^1 \leq \alpha, 0 \leq z_i \leq 1$$

Нечеткое множество $D_2 = \{(v_i, z_i)\}$.

Лекция 2.

$$P(x \in D_2/1) = \sum_{i=1}^{k} z_i p_i^1 \, , P(x \in D_2/2) = \sum_{i=1}^{k} z_i p_i^2$$

Решение нечеткой задачи:

При помощи функции Лагранжа запишем двойственную задачу:

$$\min_{\lambda \geq 0} \max_{0 \leq z_i \leq 1} \left[\sum_{i=1}^{k} z_i \left(p_i^2 - \lambda p_i^1 \right) + \lambda \alpha \right]$$

Лекция 2

$f(\lambda) = \max\limits_{0 \le z_i \le 1} \left[\sum_{i=1}^{k} z_i \left(p_i^2 - \lambda p_i^1\right) + \lambda\alpha\right]$ - кусочно-линейная выпуклая функция

Вычислим индекс

$$j = max\left\{l: \sum_{i=1}^{l} p_i^1 \le \alpha\right\}$$

Пусть $1 \le j < k$,

На интервале $\left(\frac{p_{j+1}^2}{p_{j+1}^1}, \infty\right)$ $f(\lambda)$– неубывающая функция

На интервале $\left(0, \frac{p_{j+1}^2}{p_{j+1}^1}\right)$ функция $f(\lambda)$– невозрастающая функция

Минимум функции $f(\lambda)$ достигается при $\lambda = \frac{p_{j+1}^2}{p_{j+1}^1}$

Оптимальные значения z:

$$z_i = \begin{cases} 1, i \leq j \\ \dfrac{\alpha - \sum_{i=1}^{j} p_i^1}{p_{j+1}^1}, i = j+1 \\ 0, i > j+1 \end{cases}$$

Пусть $j = k$, тогда $1 \leq \alpha$, это противоречит тому, что $\alpha < 1$. Пусть $j = 0$, тогда все $z_i = 0$ и $D_2 = \emptyset$

Алгоритм

begin

for $i = 1$to k

$z_i = 0;$

if $p_1^1 > \alpha$ then begin print z; stop end;

$z_1 = 1;$

$s = p_1^1 + p_2^1 \; ; i = 2;$

do while ($s \leq \alpha$).

$z_i = 1;$

$i = i + 1\ ; s = s + p_i^1;$

end while;

$z_i = \frac{\alpha - s + p_i^1}{p_i^1}\ ;$

print z;

end.

Задача Неймана Пирсона. Несколько источников опасных состояний

$$max \sum_{i=1}^{k} z_i p_i^1$$

$$\sum_{i=1}^{k} z_i p_i^j \leq \alpha, j = 2, \ldots, r$$

$$0 \leq z_i \leq 1$$

Лекция 3

Задача Неймана Пирсона. Несколько источников ложных тревог

$$max\, w$$

$$\sum_{i=1}^{k} z_i p_i^1 \leq \alpha$$

$$\sum_{i=1}^{k} z_i p_i^j \geq w, j = 2, \ldots, r$$

$$0 \leq z_i \leq 1$$

Лекция 3

$X^n = X \times X \ldots \times X$,

$X^n = D_0^n \cup D_1^n \cup D_2^n$

$P(x^n \in D_2^n/1)$- вероятность пропуска цели

$P(x^n \in D_1^n/2)$- вероятность ложной тревоги

$P(x \in D_0^n/1)$ и $P(x \in D_0^n/2)$ – вероятности отказа от распознавания

Задача:

$$minmax\{P(x^n \in D_0^n/1), P(x^n \in D_0^n/2)\}$$
$$P(x^n \in D_2^n/1) \leq \alpha_1$$
$$P(x^n \in D_1^n/2) \leq \alpha_2$$

Две вспомогательные задачи:

$$max\ P\ (x^n \in D_2^n/2),\ \ P(x^n \in D_2^n/1) \leq \alpha_1 \text{ и}$$

$max\ P\ (x^n \in D_1^n/1), P(x^n \in D_1^n/2) \leq \alpha_2$

Решение первой задачи: $A^n(\lambda^*), A^n(\lambda) = \left\{v^n: \frac{dP_2^n}{dP_1^n}(v^n) > \lambda\right\}$

Лекция 3

λ_n^* – решение уравнения $P(x^n \in A^n(\lambda)/1) = \alpha_1$

Решение второй задачи $B^n(\mu_n^*)$, где $B^n(\mu) = \left\{v^n : \frac{dP_2^n}{dP_1^n}(v^n) < \mu\right\}$, μ_n^* – решение уравнения $P(x^n \in B(\mu)/2) = \alpha_2$

$\mu_n^* \leq \lambda_n^* \Longrightarrow D_0^n = C^n(\lambda_n^*, \mu_n^*), D_1^n = B^n(\mu_n^*), D_2^n = A^n(\lambda_n^*)$

где $C^n(\lambda_n^*, \mu_n^*) = X \backslash (A^n(\lambda_n^*) \cup B^n(\mu_n^*))$

$\mu_n^* > \lambda_n^*$

$D_0^n = \emptyset, D_1^n = \left\{v^n : \frac{dP_2^n}{dP_1^n}(v^n) < \theta\right\}, D_2^n = \left\{v^n : \frac{dP_2^n}{dP_1^n}(v^n) \geq \theta\right\}, \theta \in (\lambda_n^*, \mu_n^*)$

момент остановки $\tau = min\{n: \neg(x^n \in D_0)\}$, далее

$l(x^n) = \begin{cases} 1, x^n \in D_1^n \\ 2, x^n \in D_2^n \end{cases}.$

Задача о разладке

Рассматривается случайная последовательность $x_1, x_2, \dots x_n, \dots$ и случайная величина $y \in \{1,2, \dots\}$ - дискретное случайное время

Закон распределения:

$$p(x_1, x_2, \dots x_n) = \sum_{i=1}^{n} p(x_1, x_2, \dots x_n / y = i) p(y = i)$$

Условная вероятность:

$$p(x_1, x_2, \dots x_n / y = i) = p_\infty(x_1, \dots, x_{i-1}) p_0(x_i, x_{i+1}, \dots, x_n),$$

где $p_\infty(x_1, \dots, x_{i-1}) = \prod_{k=1}^{i-1} p_\infty(x_k)$, $p_0(x_i, x_{i+1}, \dots, x_n) = \prod_{j=i}^{n} p_0(x_j)$

Требуется выбрать одно из двух решений $y \le n$ и $y > n$

Матрица штрафа

$$W = \begin{pmatrix} 0 & \alpha \\ \beta & 0 \end{pmatrix}$$

Оптимальное решение:

$$d = \begin{cases} y \le n, \dfrac{P(y \le n/x_1, \dots x_n)}{P(y > n/x_1, \dots x_n)} > \dfrac{\alpha}{\beta} \\ y > n, \dfrac{P(y \le n/x_1, \dots x_n)}{P(y > n/x_1, \dots x_n)} \le \dfrac{\alpha}{\beta} \end{cases}$$

Оптимальная остановка:

$$\tau = min\left\{n: \psi_n(x_1, \dots, x_n) > \frac{\alpha}{\beta}\right\}$$

Отношение правдоподобия:

$$\psi_n(x_1, \dots, x_n) = \frac{P(y \le n/x_1, \dots x_n)}{P(y > n/x_1, \dots x_n)}$$

Лекция 3

Рекуррентное уравнение:

$$\psi_n(x_1, \dots, x_n) =$$

$$\frac{1}{P(y>n)} \frac{p_0(x_n)}{p_\infty(x_n)} \big(P(y > n-1) \psi_{n-1}(x_1, \dots, x_{n-1}) +$$

$$P(y = n) \big), \psi_1(x_1) = \frac{1}{P(y>1)} P(y = 1) \frac{p_0(x_1)}{p_\infty(x_1)}$$

Оптимальная остановка:

$$\tau = min \left\{ n : \psi_n(x_1, \dots, x_n) > \frac{\alpha}{\beta} \right\}$$

Распознавании скрытых марковских цепочек

Задача вычисления ненаблюдаемой последовательности $y = \{y_0, \dots, y_n\}$ по наблюдаемой последовательности $x = \{x_1, \dots, x_n\}$:

$$\max_y p(x, y) = \max_y p(y_0) \prod_{i=1}^{n} p(x_i/y_i) p(y_i/y_{i-1})$$

Эквивалентная задача:

$$\max_y \left\{ lnp(y_0) + \sum_{i=1}^{n} \left(lnp(x_i/y_i) + lnp(y_i/y_{i-1}) \right) \right\}$$

Действие первое. Определение семейства функций Беллмана:

$$V_k(u) = \max_{y_k^n} \sum_{i=k}^{n} \left(lnp(x_i/y_i) + lnp(y_i/y_{i-1}) \right), y_{k-1} = u; k = 1,2, \dots, n;$$

$$V_0 = \max_{y} \left\{ lnp(y_0) + \sum_{i=1}^{n} \left(lnp(x_i/y_i) + lnp(y_i/y_{i-1}) \right) \right\}$$

Действие второе. Уравнения Беллмана:

$$V_k(u) = \max_{y_k} [lnp(x_k/y_k) + lnp(y_k/u) + V_{k+1}(y_k)],$$

$$W_k(u) = \operatorname{argmax}_{y_k} [lnp(x_k/y_k) + lnp(y_k/u) + V_{k+1}(y_k)],$$

$$V_n(u) = \max_{y_n}[lnp(x_n/y_n) + lnp(y_n/u)]$$

Действие 3. Решение основной задачи

$V_0 = \max_{y_0}[lnp(y_0) + V_1(y_0)]$, $y_0^* = \operatorname{argmax}_{y_0}[lnp(y_0) + V_1(y_0)]$,

$y_k^* = W_k(y_{k-1}^*)$

Пример. Задача сглаживания

$$p(x_k/\ y_k) = \frac{1}{\sqrt{2\pi\sigma_1^2}} e^{-\frac{(x_k - y_k)^2}{2\sigma_1^2}},\ p(x_k/\ y_k) =$$

$$\frac{1}{\sqrt{2\pi\sigma_2^2}} e^{-\frac{(y_k - y_{k-1})^2}{2\sigma_2^2}},\ p(y_0) = \frac{1}{\sqrt{2\pi\sigma_3^2}} e^{-\frac{(y_0 - m)^2}{2\sigma_3^2}}$$

Исходная задача: $\max_y \left\{ -\frac{1}{2} \left(ln2\pi\sigma_3^2 + \frac{(y_0 - m)^2}{\sigma_3^2} + \right.\right.$

$$\left.\left. \sum_{i=1}^{n} \left(ln2\pi\sigma_1^2 + \frac{(x_i - y_i)^2}{\sigma_1^2} + ln2\pi\sigma_2^2 + \frac{(y_i - y_{i-1})^2}{\sigma_2^2} \right) \right) \right\}$$

Эквивалентная задача:

$$\min_y \left\{ \frac{(y_0 - m)^2}{\sigma_3^2} + \sum_{i=1}^{n} \left(\frac{(x_i - y_i)^2}{\sigma_1^2} + \frac{(y_i - y_{i-1})^2}{\sigma_2^2} \right) \right\}$$

Краевые функции:

$$V_n(u) = \min_{y_n} \left[\frac{(x_n - y_n)^2}{\sigma_1^2} + \frac{(y_n - u)^2}{\sigma_2^2} \right] = \frac{(x_n - u)^2}{\sigma_1^2 + \sigma_2^2} \text{ и } W_n(u) = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} x_n + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} u$$

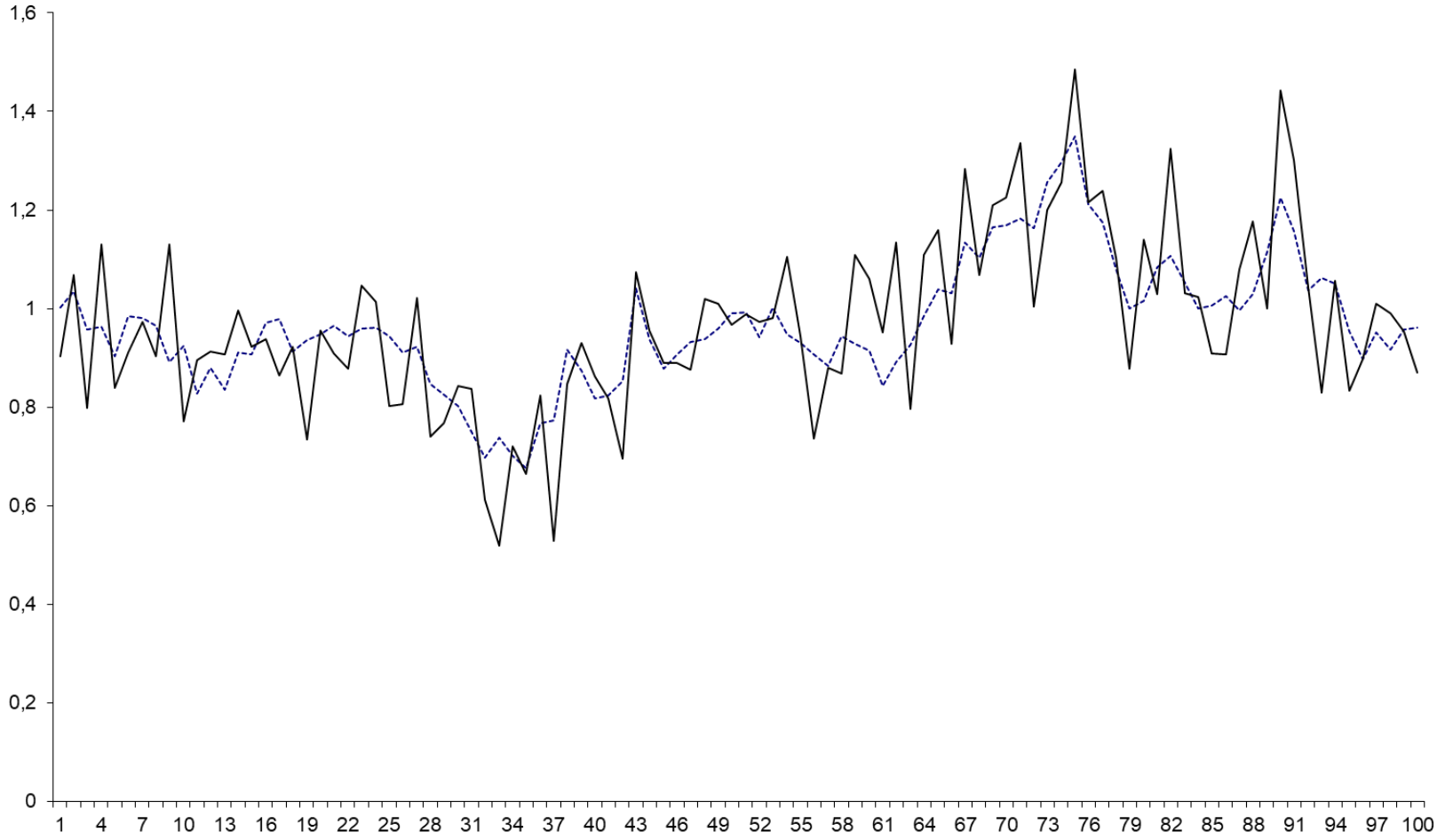
Предположим, что $V_k(u) = A_k u^2 + 2B_k u + C_k$, причем $A_k > 0$

Из уравнения: $V_k(u) = \min_{y_k} \left[\frac{(x_k - y_k)^2}{\sigma_1^2} + \frac{(y_k - u)^2}{\sigma_2^2} + A_{k+1}y_k^2 - 2B_{k+1}y_k + C_{k+1}\right]$

получим:

$A_k = \frac{A_{k+1}\sigma_1^2 + 1}{A_{k+1}\sigma_1^2\sigma_2^2 + \sigma_1^2 + \sigma_2^2}$, $B_k = \frac{B_{k+1}\sigma_1^2 + x_k}{A_{k+1}\sigma_1^2\sigma_2^2 + \sigma_1^2 + \sigma_2^2}$, $C_k = C_{k+1} + x_k^2\sigma_2^2$, $W_k(u) = \frac{B_{k+1}\sigma_1^2\sigma_2^2 + x_k\sigma_2^2 + u\sigma_1^2}{A_{k+1}\sigma_1^2\sigma_2^2 + \sigma_1^2 + \sigma_2^2}$

Лекция 3



Сглаживание

Лекция 4

Метод максимального правдоподобия. Обучение с учителем.

Задача:

$$\max_{a \in A} \sum_{i=1}^{|V|} ln\, p\,(x_i/a)$$

Многомерный нормальный закон:

$$p(v) = \frac{1}{\sqrt{(2\pi)^d |C|}} exp\left(-\frac{1}{2}(C^{-1}(v-m), v-m)\right)$$

Лекция 4

Задача:

$$\min_{m,C}\left[lndetC + \frac{1}{|V|}\sum_{i=1}^{|V|}(C^{-1}(x_i - m), x_i - m)\right]$$

Оптимальное значение для m:

$$m^* = \bar{x} = \frac{1}{|V|}\sum_{i=1}^{|V|} x_i$$

Эквивалентная задача:

$\min\limits_{A\in S^{++}} \left[-lndetA + \frac{1}{|V|}\sum_{i=1}^{|V|}(A(x_i - \bar{x}), x_i - \bar{x})\right]$, S^{++} - множество симметричных положительно определенных матриц, $C^{-1} = A$

Теорема. Функция $f(x)$ определенная на выпуклом множестве D – выпуклая функция тогда и только тогда, когда функция $g(t) = f(x + ty), t \in T = \{t|x + ty \in D\}$, $x \in D$, выпуклая функция.

Рассмотрим $f(A) = lndetA$

Рассмотрим матрицу $A + tB$, $A \in S^{++}$, B – произвольная симметричная матрица, и функцию

Лекция 4

$$g(t) = lndet(A + tB) = lndetA + lndet\left(E + tA^{-\frac{1}{2}}BA^{-\frac{1}{2}}\right) = lndetA + \sum_{i=1}^{d} ln(1 + t\lambda_i), \lambda_i -$$

собственные числа симметричной матрицы $A^{-\frac{1}{2}}BA^{-\frac{1}{2}}$

$g(t)$- вогнутая функция на множестве $1 + t\lambda_i > 0$

Следовательно, функция $f(A)$ – вогнутая функция

Вычисляем производную, приравниваем ее к нулю, в результате получаем равенство

$$C^* = \frac{1}{|V|}\sum_{i=1}^{|V|}(x_i - \bar{x})(x_i - \bar{x})^T$$

Специальный вид:

$$C = AA^T + \sigma^2 E$$

$A^T A = \Lambda$, Λ – диагональная матрица с положительными элементами, E – единичная матрица

Собственными векторами матрицы C будут векторы $u_i = a_i, i = 1,\ldots,l, u_{l+i} = b_i, i = 1,2,\ldots,d-l.$

a_i – столбцы матрицы A, λ_i– диагональные элементы матрицы Λ, b_i– произвольный набор ортонормированных векторов, принадлежащих подпространству, ортогонально дополняющему подпространство столбцов матрицы A

Собственные числа ковариационной матрицы $\mu_i = \lambda_i + \sigma^2, i = 1,\ldots,l; \mu_{l+i} = \sigma^2, i = 1,\ldots,d-l$. Обратная ковариационная матрица $C^{-1} = \frac{1}{\sigma^2}[E_d - A(\Lambda + \sigma^2 E_l)^{-1}A^T]$, определитель ковариационной матрицы $|C| = (\sigma^2)^{d-l}\prod_{i=1}^{l}(\lambda_i + \sigma^2)$

Максимально правдоподобная оценка среднего значения – это выборочное среднее. Чтобы получить оценки остальных параметров, сначала находим l главных собственных векторов выборочной ковариационной матрицы u_i и соответствующие им собственные числа μ_i. Далее вычисляем $\sigma^2 = \frac{1}{d-l}\left(\sum_{i=1}^{d} r_{i,i} - \sum_{i=1}^{l} \mu_i\right)$, где $r_{i.i}$– диагональные элементы выборочной ковариационной матрицы. Теперь мы можем вычислить $\lambda_i = \mu_i - \sigma^2$. Столбцы матрицы A – это собственные векторы, нормированные таким образом, чтобы выполнялось равенство $A^T A = \Lambda$.

Лекция 4

Метод максимального правдоподобия. Обучение с сомневающимся учителем

Имеется обучающая выборка, относительно каждого элемента выборки x_i известен набор неотрицательных чисел $z_{i,j}, j = 1, \ldots, |Y|, (\sum_{j=1}^{|Y|} z_{i,j} = 1)$ которые можно рассматривать как вероятности принадлежности элемента выборки образам

Логарифм функции правдоподобия

$$lnp(x_1, \ldots, x_{|V|}) = \sum_{i=1}^{|V|} ln \sum_{j=1}^{|Y|} p_j p(x_i/a_j) =$$
$$\sum_{i=1}^{|V|} \sum_{k=1}^{|Y|} z_{i,k} ln \sum_{j=1}^{|Y|} p_j p(x_i/a_j)$$

После применения формулы Байеса $z_{i,k} = \frac{p(x_i/a_k)p_k}{\sum_{j=1}^{|Y|} p_j p(x_i/a_j)}$

$$lnp(x_1, \dots, x_{|V|}) = \sum_{i=1}^{|V|}\sum_{k=1}^{|Y|} z_{i,k}\left(lnp_k + lnp(x_i/a_k) - lnz_{i,k}\right)$$

Оптимальные значения

$$p_k^* = \frac{1}{|V|}\sum_{i}^{|V|} z_{i,k}$$

Оптимальные значения параметров получаются в результате решения серии задач

$$\max_{a_j \in A_j} \sum_{i=1}^{|V|} z_{i,j} \ln p\left(x_i / a_j\right)$$

Для многомерных нормальных законов

$$\bar{x}_j = \frac{1}{\sum_{i=1}^{|V|} z_{i,j}} \sum_{i=1}^{|V|} z_{i,j} x_i \,, C_j$$

$$= \frac{1}{\sum_{i=1}^{|V|} z_{i,j}} \sum_{i=1}^{|V|} z_{i,j} \left(x_i - \bar{x}_j\right)\left(x_i - \bar{x}_j\right)^T$$

Задача обучения без учителя

$$\max_{\{p_j, a_j\}} \left[F\left(\{p_j, a_j\}\right) = \sum_{i=1}^{|V|} ln \sum_{j=1}^{|Y|} p_j p\left(x_i / a_j\right) \right]$$

Алгоритм

На стадии инициализации выбираются начальные значения параметров: p_j^0 и a_j^0.

На итерации с номером t считаем известными значения параметров:p_j^tиa_j^t. Определим величины

$\alpha_{i,k}^t = \frac{p(x_i/a_k^t)p_k^t}{\sum_{j=1}^{|Y|} p\left(x_i/a_j^t\right)p_j^t}$, для которых справедливо:

$\sum_{k=1}^{|Y|} \alpha_{i,k}^t = 1, \alpha_{i,k}^t \geq 0.$

Это обстоятельство позволяет записать целевую функцию в задаче (2.15) в виде:

$$F\left(\{p_j, a_j\}\right) = \sum_{i=1}^{|V|}\sum_{k=1}^{|Y|} \alpha_{i,k}^t \, ln \sum_{j=1}^{|Y|} p_j p\left(x_i/a_j\right) =$$

$$= \sum_{i=1}^{|V|}\sum_{k=1}^{|Y|} \alpha_{i,k}^t \left(ln\, p_k + ln\, p\,(x_i/a_k) - ln \frac{p_k p(x_i/a_k)}{\sum p_j p\left(x_i/a_j\right)} \right)$$

Вспомогательная функция $\bar{F}_t\left(\{p_j, a_j\}, \{\alpha_{i,k}\}\right) = \sum_{i=1}^{|V|} \sum_{k=1}^{|Y|} \alpha_{i,k}\left(ln\, p_k + ln\, p\,(x_i/a_k) - ln\, \alpha_{i,k}\right)$

Очевидно равенство

$F\left(\{p_j^t, a_j^t\}\right) = \bar{F}\left(\{p_j^t, a_j^t\}, \{\alpha_{i,k}^t\}\right).$

Пусть $\{p_j^{t+1}, a_j^{t+1}\}$ доставляют

максимум вспомогательной функции $\bar{F}\left(\{p_j, a_j\}, \{\alpha_{i,k}^t\}\right)$

по переменным $\{p_j, a_j\}$

Вспомогательная функция – строго вогнутая функция по p_k :

$p_j^t \neq p_j^{t+1}$ хотя бы для одного $j \Rightarrow F\left(\left\{p_j^t, a_j^t\right\}\right) < \bar{F}\left(\left\{p_j^{t+1}, a_j^{t+1}\right\}, \left\{\alpha_{i,k}^t\right\}\right)$

Вычислим $\max\limits_{\{\alpha_{i,k}\}} \bar{F}\left(\left\{p_j^{t+1}, a_j^{t+1}\right\}, \left\{\alpha_{i,k}\right\}\right)$,

$\sum_{k=1}^{|Y|} \alpha_{i,k} = 1, \alpha_{i,k} \geq 0$

Решение:

$$\alpha_{i,k}^{t+1} = \frac{p(x_i/a_k^{t+1})p_k^{t+1}}{\sum_{j=1}^{|Y|} p\left(x_i/a_j^{t+1}\right)p_j^{t+1}}$$

Справедлива цепочка неравенств:

$$F\left(\left\{p_j^t, a_j^t\right\}\right) \le \bar{F}\left(\left\{p_j^{t+1}, a_j^{t+1}\right\}, \left\{\alpha_{i,k}^t\right\}\right)$$
$$\le \bar{F}\left(\left\{p_j^{t+1}, a_j^{t+1}\right\}, \left\{\alpha_{i,k}^{t+1}\right\}\right) = F\left(\left\{p_j^{t+1}, a_j^{t+1}\right\}\right)$$

При незначительном изменении целевой функции можно останавливать итерации

Утверждение

Последовательность значений целевой функции будет сходящейся последовательностью

Поведение последовательности α^t

Последовательность значений функции

$$F\left(\left\{p_k^t\right\}, \left\{a_k^t\right\}\right) \uparrow F^*$$

Отсюда следует, что

$$lim\left(F(\{p_k^{t+1}\},\{a_k^{t+1}\}) - F(\{p_k^t\},\{a_k^t\})\right) = 0$$

Поскольку $F(\{p_k^{t+1}\},\{a_k^{t+1}\}) - F(\{p_k^t\},\{a_k^t\}) =$

$$= \sum_{k=1}^{|Y|}\sum_{i=1}^{|V|} \alpha_{i,k}^{t+1}\left[(ln\, p_k^{t+1} - ln\, p_k^t)\right.$$

$$\left. + \left(ln\, p\,(x_i/a_k^{t+1}) - ln\, p\,(x_i/a_k^t)\right)\right] +$$

$+\sum_{i=1}^{|V|}\sum_{k=1}^{|Y|}\alpha_{i,k}^{t+1}\,ln\frac{\alpha_{i,k}^{t+1}}{\alpha_{i,k}^{t}}$ и каждое из двух слагаемых

положительно, то $lim\sum_{i=1}^{|V|}\sum_{k=1}^{|Y|}\alpha_{i,k}^{t+1}\,ln\frac{\alpha_{i,k}^{t+1}}{\alpha_{i,k}^{t}}=0$.

Лемма. Неравенство Кульбака. Пусть $\{x_i\}$ и $\{y_i\}$ – распределение вероятностей на множестве Y, причем распределение $\{y_i\}$ абсолютно непрерывно по отношению к распределению $\{x_i\}$, тогда $\sum_{i=1}^{|Y|}x_i\,ln\frac{x_i}{y_i}\geq\frac{1}{2}\sum_{i=1}^{|Y|}(x_i-y_i)^2$

Из неравенства Кульбака следует

$$\|\alpha^{t+1} - \alpha^t\| \to 0$$

Итерации алгоритма самообучения можно представить как результат действия оператора:

$$\alpha^{t+1} = Sl(\alpha^t)$$

Будет справедливо равенство нулю предела:

$$\|Sl(\alpha^t) - \alpha^t\| \to 0$$

Выделим сходящуюся подпоследовательность $\alpha^{t_1}, \ldots, \alpha^{t_n}, \ldots$ с пределом a

Очевидно:

$$\lim_{j\to\infty} Sl(\alpha^{t_j}) = a$$

Допустим, что оператор Sl – непрерывный оператор, тогда a является неподвижной точкой этого оператора:

$$Sl(a) = a$$

Будем считать, что $A = \{\xi_1 \ldots, \xi_{|A|}\}$ – множество неподвижных точек оператора Sl

Последовательность расстояний

$$d^t = d(\alpha^t, A) = \min_{\xi\in A} \|\alpha^t - \xi\|$$

стремится к нулю. Допустим, что это не так. Это означает, что можно подобрать такое значение $\varepsilon > 0$ и такую подпоследовательность α^{t_k}, для которой $d^{t_k} \geq \varepsilon > 0$. Выделим из этой подпоследовательности сходящуюся подпоследовательность. Естественно, что предел этой подпоследовательности не будет неподвижной точкой

Теперь мы можем доказать, что существует предел последовательности α^t и этот предел – неподвижная точка оператора $Sl(\alpha)$. Выберем $\varepsilon > 0$ и выберем T таким образом, для всех $t \geq T$ одновременно

выполнялись неравенства: $\|\alpha^{t+1} - \alpha^t\| < \varepsilon/3$, $d(\alpha^t, A) = \|\alpha^t - a\| < \varepsilon/3$. Пусть $d(\alpha^{t+1}, A) = \|\alpha^{t+1} - b\|$ и $a \neq b$. Из неравенства треугольника следует, что $\|b - a\| \leq \|b - \alpha^{t+1}\| + \|\alpha^{t+1} - \alpha^t\| + \|\alpha^t - a\| < \varepsilon$. Таким образом, начиная с T, ближайшая точка к элементам последовательности α будет сохраняться, то есть, $d(\alpha^t, A) = \|\alpha^t - a\|$ для $t \geq T$. Эта неподвижная точка будет пределом последовательности α

Лекция 4

Задача Робинса

$$\max_{\{p_k\}} \sum_{i=1}^{|V|} ln \sum_{k=1}^{|Y|} p_k p(x_i/k)$$

Алгоритм самообучения, в котором не вычисляются параметры условных распределений $\{a_k\}$, построит последовательность, которая сходится к одной из неподвижных точек функции $Sl(.)$. Отметим, что целевая функция – строго вогнутая функция, поэтому есть все основания считать, что последовательность $\{p_k^t\}$, в которой элементы выражаются через $\{\alpha_{i,k}^t\}$

следующим образом: $p_k^{t+1} = \frac{1}{|V|}\sum_{i=1}^{|V|}\alpha_{i,k}^t$, будет сходиться к $\{p_k^*\}$, причем $p_k^* = \frac{1}{|V|}\sum_{i=1}^{|V|}\xi_{i,k}$, где ξ – неподвижная точка отображения $Sl(.)$. В алгоритме самообучения мы можем исключить промежуточные вычисления элементов последовательности α. Соответствующая формула будет иметь вид:

$$p_k^{t+1} = \frac{1}{|V|}\sum_{i=1}^{|V|}\frac{p_k^t p(x_i/k)}{\sum_{j=1}^{|Y|}p_j^t p(x_i/j)}$$

Перейдем к пределу в левой и правой частях

$$p_k^* = \frac{1}{|V|}\sum_{i=1}^{|V|}\frac{p_k^* p(x_i/k)}{\sum_{j=1}^{|Y|} p_j^* p(x_i/j)}$$

Это равенство является необходимым и достаточным условием максимума логарифма функции правдоподобия $\sum_{i=1}^{|V|} ln \sum_{k=1}^{|Y|} z_k p(x_i/k)$ при ограничениях: $\sum_{k=1}^{|V|} z_k = 1$

Из $p_k^0 > 0$ следует $p_k^t > 0$

Линейная дискриминантная функция для распознавания двух классов:

$$f(x) = \begin{cases} 1, (l, x) \geq \theta \\ 2, (l, x) < \theta \end{cases}$$

$$J = J_1 \cup J_2$$

Условная вероятность ошибки при распознавании первого класса:

$$\alpha_j(l, \theta) = P\big((l, x) < \theta / j\big) = \Phi\left(\frac{\theta - (l, m_j)}{\sqrt{(C_j l, l)}}\right)$$

Условная вероятность ошибки при распознавании первого класса:

$$\beta_j(l,\theta) = P\big((l,x) \geq \theta / j\big) = 1 - \Phi\left(\frac{\theta - (l,m_j)}{\sqrt{(C_j l,l)}}\right) =$$

$$\Phi\left(\frac{(l,m_j) - \theta}{\sqrt{(C_j l,l)}}\right), j \in J_2$$

Максимально возможные вероятности ошибки для первого класса:

$$\alpha(l,\theta) = \max_{j\in J_1} \Phi\left(\frac{\theta - (l, m_j)}{\sqrt{(C_j l, l)}}\right)$$

для второго класса:

$$\beta(l,\theta) = \max_{k\in J_2} \Phi\left(\frac{(l, m_k) - \theta}{\sqrt{(C_k l, l)}}\right)$$

Задача Андерсона

$$\min_{l,\theta} \max\big(\alpha(l,\theta), \beta(l,\theta)\big)$$

Лекция 5

Эквивалентная задача

$$\max_{l,\theta\in\Pi} min\left(\min_{j\in J_1}\frac{(l,m_j)-\theta}{\sqrt{(C_j l,l)}},\min_{k\in J_2}\frac{\theta-(l,m_k)}{\sqrt{(C_k l,l)}}\right),\ \Pi=\{(l,\theta)\in$$

$$R^{n+1}:(l,m_j)>\theta, j\in J_1;(l,m_k)<\theta, k\in J_2\}$$

Следующая задача:

$$\max_{l\in\square}\min_{j\in J_1\cup J_1}\frac{(l,m_j)}{\sqrt{(C_j l,l)}},\square=\{l:(l,m_j)>0\}$$

$$m_j := \begin{pmatrix} m_j \\ -1 \end{pmatrix}, j \in J_1, m_j := \begin{pmatrix} -m_j \\ 1 \end{pmatrix}, j \in J_2, l := \begin{pmatrix} l \\ \theta \end{pmatrix}$$

Следующая задача:

$$max\, y$$

$$y\sqrt{(C_j l, l)} - (l, m_j) \leq 0, j \in J_1 \cup J_2$$

$$y \geq 0$$

Алгоритм

1. Выберем начальное значение y_0 для y и начальное значение h_0 для шага h .

2. На итерации с номером t решаем задачу выпуклого программирования

$\min_{l} \max_{j} y_t \sqrt{(C_j l, l)} - (l, m_j)$. Выбираем шаг h_{t+1}и вычисляем y_{t+1} в зависимости от значения целевой функции. Если оптимальное значение целевой функции отрицательно, то $h_{t+1} = h_t, y_{t+1} = y_t + h_{t+1}$; иначе $h_{t+1} = h_t/2, y_{t+1} = y_{t-1} + h_{t+1}$.

3. Остановка.

После решения последней оптимизационной задачи определяется вектор l

Задача Андерсона для двух нормальных законов:

$$\max_{l,\theta} \min \left(\frac{(l, m_1) - \theta}{\sqrt{(C_1 l, l)}}, \frac{\theta - (l, m_2)}{\sqrt{(C_2 l, l)}} \right)$$

θ находится из уравнения: $\frac{\theta - (l, m_2)}{\sqrt{(C_2 l, l)}} = \frac{(l, m_1) - \theta}{\sqrt{(C_1 l, l)}}$

Оптимальное значение:

$$\theta^* = (l, m_1)\frac{\sqrt{(C_2 l, l)}}{\sqrt{(C_1 l, l)} + \sqrt{(C_2 l, l)}}$$

$$+ (l, m_2)\frac{\sqrt{(C_1 l, l)}}{\sqrt{(C_1 l, l)} + \sqrt{(C_2 l, l)}}$$

Задача для вычисления вектора l:

$$\max_l \frac{(l, m_1 - m_2)}{\sqrt{(C_1 l, l)} + \sqrt{(C_2 l, l)}}, (l, m_1 - m_2) \geq 0$$

Алгоритм.

1. Выберем начальное значение y_0 для y и начальное значение h_0 для шага h .

2. На итерации с номером t решаем задачу выпуклого программирования

$$\min_l \; y_t\left(\sqrt{(C_j l, l)} + \sqrt{(C_j l, l)}\right) - (l, m_1 - m_2) \; .$$

Выбираем шаг h_{t+1}и вычисляем y_{t+1} в зависимости от значения целевой функции. Если оптимальное значение целевой функции отрицательно, то $h_{t+1} = h_t, y_{t+1} = y_t + h_{t+1}$; иначе $h_{t+1} = h_t/2, y_{t+1} = y_{t-1} + h_{t+1}$.

3. Остановка.

После решения последней оптимизационной задачи определяется вектор l и вычисляется порог.

Задача Андерсона для двух нормальных законов с одинаковыми ковариационными матрицами:

$$\max_{l} \frac{(l, m_1 - m_2)}{\sqrt{(Cl, l)}}$$

Оптимальные значения:

$$l^* = C^{-1}(m_1 - m_2),\ \theta^* = \frac{(C^{-1}(m_1 - m_2), m_1 + m_2)}{2}$$

Задача Андерсона для нескольких нормальных законов с одинаковыми ковариационными матрицами:

$$\max_{l,\theta} \min\left(\min_{j\in J_1} \frac{(l,m_j)-\theta}{\sqrt{(Cl,l)}}, \min_{k\in J_2} \frac{\theta-(l,m_k)}{\sqrt{(Cl,l)}}\right)$$

Эквивалентная задача:

$$\max_{l} \frac{1}{\sqrt{(Cl,l)}}\left(\min_{j\in J_1}(l,m_i)-\max_{k\in J_2}(l,m_j)\right)$$

Обозначим через $z_{i,j}=m_i-m_j; i\in J_1, j\in J_2$

Требуется найти:

$$\max_{l} \min_{i \in J_2, j \in J_1} \frac{\left(l, z_{i,j}\right)}{\sqrt{(Cl, l)}}$$

Применение теоремы о минимаксе:

$$\min_{z \in Z} \max_{l \in \Pi_l \cap \{l \in R^n : \|l\| \leq \alpha\}} \frac{(l,z)}{\sqrt{(Cl,l)}},$$

$$\Pi_l = \left\{l : \left(l, z_{i,j}\right) \geq 0\right\},$$

Z – выпуклая оболочка, натянутая на векторы $z_{i,j}$

Решение внутренней задачи: $l = \beta C^{-1} z$, β – решение уравнения $\|l\| = \alpha$

Внешняя задача:

$$\min_{z \in Z}(C^{-1}z, z)$$

Разложение Холецкого:

$C = LL^T$, где L – нижнетреугольная матрица

Введем обозначение: $\bar{z} = L^{-1}z$, получим задачу:

$$\min_{z \in \bar{Z}}(z, z), \bar{Z} = L^{-1}Z$$

Введем одномерную нумерацию для векторов $\bar{z}_{i,j} = L^{-1}z_{i,j}$ и приведем алгоритм, решающий задачу:

Алгоритм Козинца.

$$begin$$
$$l = a; ep = \infty$$
$$DoWhile(ep \geq \varepsilon)$$
$$begin$$
$$j = |J_1||J_2|; l_1 = l$$
$$fori = 1Toj$$
$$begin$$
$$h = \frac{(\bar{z}_i, \bar{z}_i - l)}{(\bar{z}_i - l, \bar{z}_i - l)}; l = \begin{cases} \bar{z}_i, h < 0 \\ \bar{z}_i + h(l - \bar{z}_i), 0 \leq h < 1 \\ l, 1 \leq h \end{cases}$$
$$end$$

$$ep = \|l_1 - l\|$$
$$end$$
$$end$$

В алгоритме a – начальное значение для вектора l, ε – выбранная точность

Задача Неймана-Пирсона

Рассматривается задача распознавания двух классов, описываемых смесью нормальных законов

Задача Неймана-Пирсона заключается в следующем:

$$max\ y$$
$$\alpha\sqrt{\left(C_j l, l\right)} + \theta - \left(l, m_j\right) \leq 0, j \in J_1$$
$$y\sqrt{\left(C_k l, l\right)} - \theta + \left(l, m_k\right) \leq 0, k \in J_2$$
$$y \geq 0$$

В задаче множитель $\alpha = \Phi^{-1}(1-\beta)$, где β – вероятность пропуска цели

Задача Неймана-Пирсона для двух нормальных законов:

$$max\, y$$
$$\alpha\sqrt{(C_1 l, l)} + \theta - (l, m_1) \leq 0,$$
$$y\sqrt{(C_2 l, l)} - \theta + (l, m_2) \leq 0,$$
$$y \geq 0$$

Решим задачу (3.18) при фиксированном векторе l.

Оптимальное значение для $y - y^* = \frac{(l, m_1 - m_2)}{\sqrt{(C_2 l, l)}} - \alpha\sqrt{\frac{(C_1 l, l)}{(C_2 l, l)}}$

, при пороге $\theta^* = (l, m_1) - \alpha\sqrt{(C_1 l, l)}$. Таким образом,

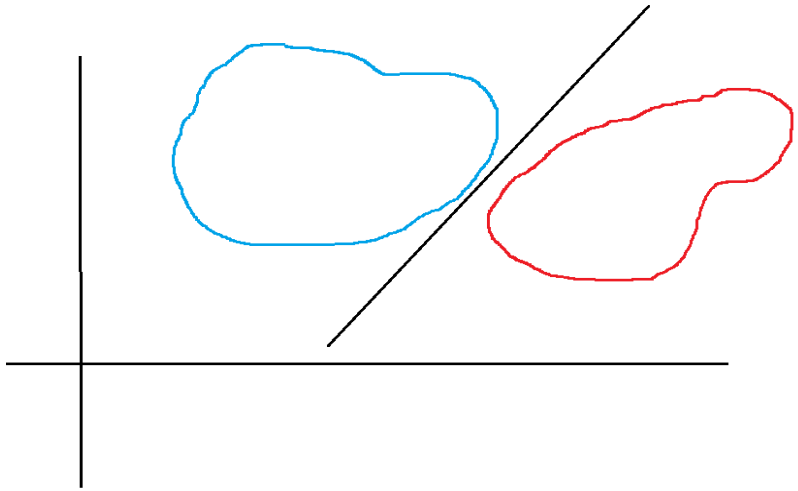
следует вычислить максимум функции $F(l) =$

$$\frac{(l,m_1-m_2)}{\sqrt{(C_2 l,l)}} - \alpha\sqrt{\frac{(C_1 l,l)}{(C_2 l,l)}}$$

Линейные дискриминантные функции для конечных множеств точек

Линейная разделимость. Будем говорить, что множества точек линейно разделимы, если существует такой порог θ и вектор l, для которых выполняются неравенства:

$$(l, x_j) > \theta, j \in J_1$$
$$(l, x_j) < \theta, j \in J_2$$



Линейная разделимость

Утверждение. Множества точек разделимы тогда и только тогда, когда выпуклые оболочки множеств не пересекаются

Рассмотрим задачу линейного программирования, связанную с линейным разделением множеств:

$$max\ z$$
$$(\bar{l}, y_j) \geq z, j \in J_1 \cup J_2.$$

В задаче векторы $\bar{l} = \begin{pmatrix} l \\ \theta \end{pmatrix}, y_j = \begin{pmatrix} x_j \\ -1 \end{pmatrix}$, если $j \in J_1$, и $y_j = \begin{pmatrix} -x_j \\ 1 \end{pmatrix}$, если $j \in J_2$. Является справедливым следующее утверждение

Утверждение. Задача имеет конечное решение тогда и только тогда, когда множества – линейно неразделимы

Лекция 5

Алгоритм. Линейная разделимость.

$$begin$$

$$\bar{l} = a; ep = \infty$$

$$Do\ While\big((ep \geq \varepsilon) \wedge (z \leq 0)\big)$$

$$begin$$

$$z = \min_{j \in J_1 \cup J_2} (\bar{l}, y_j)$$

$$if\ z > 0\ then\ goto\ met1$$

$$j = arg \min_{j \in J_1 \cup J_2} (\bar{l}, y_j)$$

$$\bar{l} = \bar{l} + hy_j$$

$$ep = abs\left(z - \min_{j \in J_1 \cup J_2} (\bar{l}, y_j)\right)$$

end

$met1: if z > 0$

$then\ print$ множества разделимы $else\ print$ множества нераз

end

В алгоритме $h > 0$– величина постоянного шага, ε – выбранная точность

Обозначим множество векторов нормалей разделяющих гиперплоскостей через L

Пусть $\alpha(l) = \min\limits_{i \in J_1} \frac{(l,x_i)}{\|l\|}, \beta(l) = \max\limits_{j \in J_2} \frac{(l,x_j)}{\|l\|}$

Естественным показателем качества разделяющей гиперплокости с вектором нормали l является разность $d(l) = \alpha(l) - \beta(l)$

Задача по вычислению оптимальной разделяющей гиперплоскости:

$$\max\limits_{l \in L} \min\limits_{z \in Z} \frac{(l,z)}{\|l\|},$$

в которой множество Z – выпуклая оболочка, натянутая на векторы $x_i - x_j, i \in J_1, j \in J_2$

Решение внутренней задачи которой имеет вид: $l = \alpha z$

Внешняя задача:

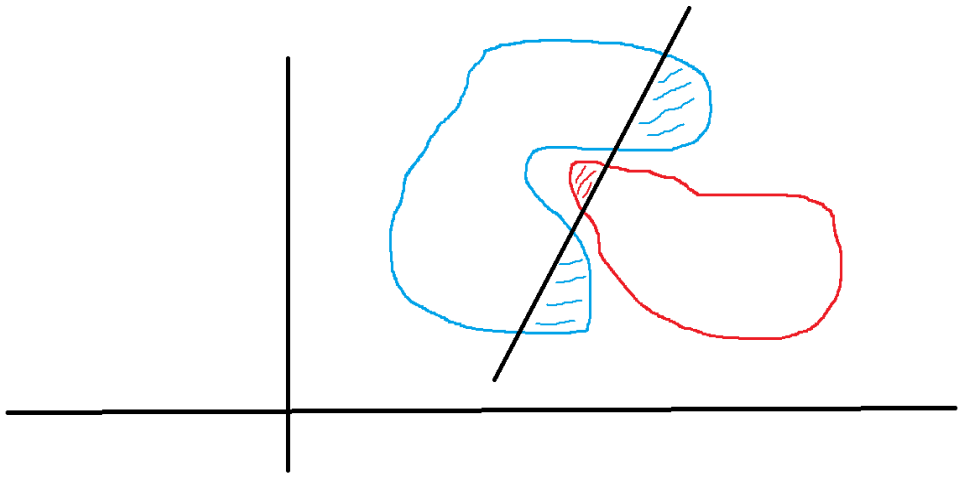
$$\min_{z \in Z} \|z\|,$$

для решения которой можно применить алгоритм Козинца. Алгоритм Козинца позволяет вычислить вектор нормали l оптимальной разделяющей гиперплоскости

Оптимальный порог $\theta = \frac{\min_{i \in J_1}(l, x_i) + \min_{j \in J_2}(l, x_j)}{2}$

Лекция 6

Множества линейно неразделимы



Множества линейно неразделимы

Выберем вектор нормали l и вычислим следующие величины, которые могут служить показателями качества выбора

Прежде всего, вычислим среднее значение $\alpha(l) = (l, m_1 - m_2)$, в котором $m_1 = \frac{1}{|J_1|}\sum_{i \in J_1} x_i$, $m_2 = \frac{1}{|J_2|}\sum_{j \in J_2} x_j$

Далее вычислим разбросы

для первого класса

$$\beta(l) = \left(\left(\frac{1}{|J_1|} \sum_{i \in J_1} x_i x_i^T - m_1 m_1^T \right) l, l \right)$$

для второго класса

$$\gamma(l) = \left(\left(\frac{1}{|J_2|} \sum_{j \in J_2} x_j x_j^T - m_2 m_2^T \right) l, l \right)$$

Таким образом, возникает оптимизационная задача с тремя критериями

Одна из возможных постановок задачи

$$min(C_2 l, l)$$
$$(l, m_1 - m_2) \geq a$$
$$(C_1 l, l) \leq b$$

$$C_1 = \frac{1}{|J_1|} \sum_{i \in J_1} x_i x_i^T - m_1 m_1^T, C_2 = \frac{1}{|J_2|} \sum_{j \in J_2} x_j x_j^T - m_2 m_2^T$$

Утверждение. Множество допустимых решений задачи – непустое множество, если выполняется неравенство:

$$b \geq \frac{a^2}{(C_1^{-1}(m_1 - m_2), m_1 - m_2)}$$

Функция Лагранжа:

$$F(l, \lambda, \mu) = (C_2 l, l) + \lambda\big(a - (l, m_1 - m_2)\big) + \mu\big((C_2 l, l) - b\big)$$

Условие оптимальности и условия Куна-Таккера

$$(C_2 + \mu C_1)l = \lambda(m_1 - m_2),$$
$$\lambda\big(a - (l, m_1 - m_2)\big) = 0,$$
$$\mu\big((C_1 l, l) - b\big) = 0,$$

$$(l, m_1 - m_2) \geq a,$$
$$(C_1 l, l) \leq b, \lambda \geq 0, \mu \geq 0$$

Проанализируем условия Куна-Таккера. Поскольку $l = 0$ не удовлетворяет ограничению: $(l, m_1 - m_2) \geq a$, то $\lambda > 0$. Тогда $(l, m_1 - m_2) = a$. Пусть $\mu = 0$, тогда вектор $l = \left[\frac{a}{(C_2^{-1}(m_2 - m_1), m_2 - m_1)}\right] C_2^{-1}(m_2 - m_1)$ – решение задачи, если $b \geq \left[\frac{a}{(C_2^{-1}(m_2 - m_1), m_2 - m_1)}\right]^2 \left(C_1 C_2^{-1}(m_2 - m_1), C_2^{-1}(m_2 - m_1)\right)$, иначе $\mu > 0$. Таким образом, если $\lambda > 0, \mu > 0$, то условия (3.26) превращаются в систему уравнений:

$$(C_2 + \mu C_1)l = \lambda(m_1 - m_2),$$
$$(l, m_1 - m_2) = a,$$
$$(C_1 l, l) = b$$

Задача существенно упрощается, если $\mathrm{C}_1 \approx \mathrm{C}_2$ в смысле какой-либо из матричных норм

В этом случае мы можем взять в качестве общей матрицы матрицу $\mathrm{C} = \frac{1}{2}(\mathrm{C}_1 + \mathrm{C}_2)$ и рассмотреть задачу

$$min(Cl, l)$$
$$(l, m_1 - m_2) \geq a$$

Решение

$$l = \frac{a}{(C^{-1}(m_2 - m_1), m_2 - m_1)} C^{-1}(m_2 - m_1)$$

Пусть $a = (C^{-1}(m_2 - m_1), m_2 - m_1)$, тогда

$$l = C^{-1}(m_2 - m_1)$$

Минимизация вероятности ошибки

Эмпирическая вероятность ошибки

$$\sum_{x \in V} I_{\{(l,x) \leq 0\}}(l)$$

Задача

$$\min_l \frac{1}{|V|} \sum_{x \in V} I_{\{(l,x) \leq 0\}}(l)$$

Сложность заключается в том, что целевая функция этой задачи не является выпуклой

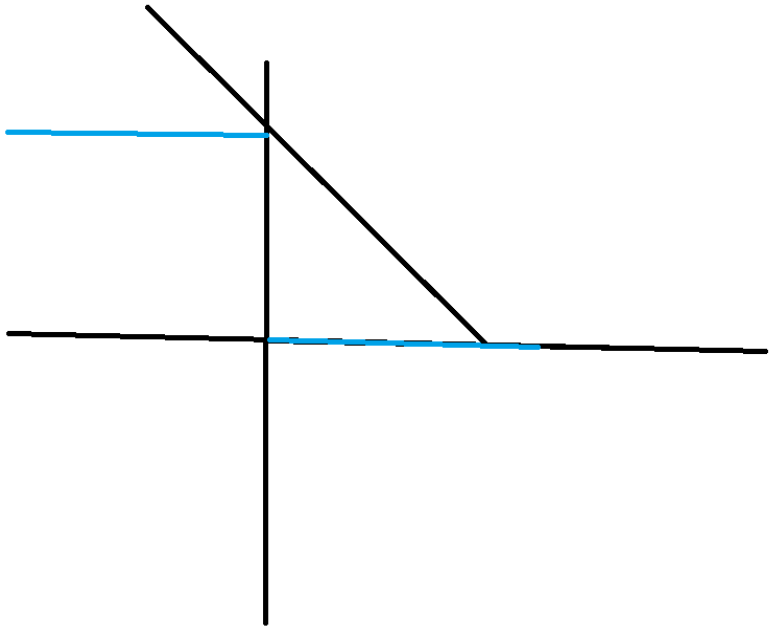
Вместо этой задачи рассмотри задачу:

$$\min_{l} \frac{1}{|V|} \sum_{x \in V} \big(1 - (l, x)\big)^{+}$$

Целевая функция этой задачи – выпуклая функция

Справедливо неравенство

$$(1 - y)^{+} \geq I_{\{y \leq 0\}}(y)$$



Минимизация эмпирической вероятности ошибки распознавания

Распознавание нескольких классов

Множество индексов J разбито на подмножества: $J = \bigcup_{i=1}^{k} J_i$

Решающее правило

$$f(x) = arg\,max_j \left[\left(l_j, x\right) + \theta_j\right]$$

Решающее правило преобразуется следующим образом:

$$f(x) = arg\,max_j \left(\bar{l}_j, \bar{x}\right)$$

Для упрощения изложения далее будем рассматривать распознавание трех классов. Сформируем множество

Лекция 6

$$X = X_1 \cup X_2 \cup X_3, \text{ где } X_1 = \left\{ \begin{pmatrix} \bar{x}_i \\ -\bar{x}_j \\ 0 \end{pmatrix}, i \in J_1, j \in J_2 \right\} \cup \left\{ \begin{pmatrix} \bar{x}_i \\ 0 \\ -\bar{x}_j \end{pmatrix}, i \in J_1, j \in J_3 \right\},$$

$$X_2 = \left\{ \begin{pmatrix} -\bar{x}_j \\ \bar{x}_i \\ 0 \end{pmatrix}, i \in J_2, j \in J_1 \right\} \bigcup \left\{ \begin{pmatrix} 0 \\ \bar{x}_i \\ -\bar{x}_j \end{pmatrix}, i \in J_2, j \in J_3 \right\},$$

$$X_3 = \left\{ \begin{pmatrix} -\bar{x}_j \\ 0 \\ \bar{x}_i \end{pmatrix}, i \in J_3, j \in J_1 \right\} \bigcup \left\{ \begin{pmatrix} 0 \\ -\bar{x}_j \\ \bar{x}_i \end{pmatrix}, i \in J_3, j \in J_2 \right\}$$

Утверждение. Решающее правило правильно распознает выборку, тогда и только тогда, когда выпуклая оболочка, натянутая на множество X строго отделена от нуля

Отделимость от нуля можно проверить, решая задачу линейного программирования:

$$max\ z$$
$$(L, y) \geq z, y \in X$$

Если задача имеет решение, то выпуклая оболочка неотделима от нуля, если задача не имеет решения, то выпуклая оболочка отделима от нуля

Лекция 6

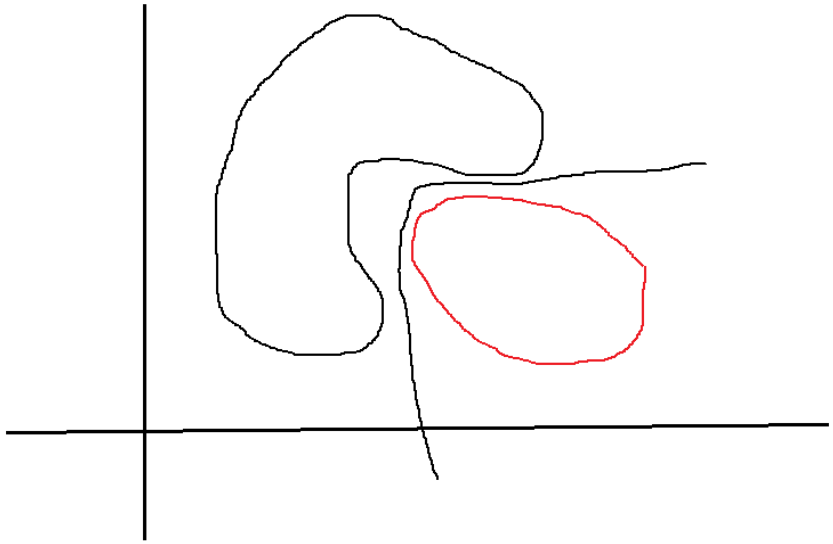
Нелинейные решающие правила

$$l(x) = \begin{cases} 1, & f(x) \geq 0 \\ 2, & f(x) < 0 \end{cases},$$

$$f(x) = \sum_{i=1}^{N} a_i \varphi_i(x)$$

Приведем пример такой функции:

$$f(x) = a_0 + \sum_{i=1}^{n} a_i x_i + \sum_{i=1}^{n} \sum_{j=i}^{n} a_{i,j} x_i x_j$$



Нелинейное решающее правило

Рассмотрим отображение $\Phi: R^n \to R^N$, которое определяется равенством $\big(\Phi(x)\big)_i = \varphi_i(x)$. В пространстве R^N данное решающее правило будет иметь вид:

$$l(x) = \begin{cases} 1, & (a, y) \geq 0 \\ 2, & (a, y) < 0 \end{cases}, y = \Phi(x)$$

Рассмотрим выборку $V = \{(\Phi(x_1), z_1), \dots, (\Phi(x_i), z_i)\}$ и задачу регрессии для этой выборки:

$$\min_a \left[\sum_{i=1}^{m} \Big(\big(a, \Phi(x_i)\big) - z_i \Big)^2 + \alpha(a, a) \right]$$

с регуляризацией $\alpha(a, a)$

Необходимое и достаточное условие:

$$\sum_{i=1}^{m} \Big(\big(a, \Phi(x_i)\big) - z_i \Big) \Phi(x_i) \; + \alpha a = 0$$

Оптимальное значение для вектора имеет следующий вид:

$$a = \sum_{i=1}^{m} \beta_i \Phi(x_i)$$

Оптимальная функция

$$f(x) = \big(a, \Phi(x)\big) = \sum_{i=1}^{m} \beta_i \big(\Phi(x), \Phi(x_i)\big)$$

Потенциальная функция

$$K(x, y) = \big(\Phi(x), \Phi(y)\big)$$

Решающее правило:

$$f(x) = \sum_{i=1}^{m} \beta_i K(x, x_i)$$

Оптимизационная задача:

$$\min_{\beta} \left[\sum_{i=1}^{m} \left(z_i - \sum_{j=1}^{m} \beta_j K(x_i, x_j) \right)^2 + \alpha \sum_{j=1}^{m} \beta_j^2 \right]$$

Функция Гаусса: $K(x, y) = e^{-1/2\sigma^2 \|x-y\|^2}$

Лекция 6

Логистическая регрессия

$$f(x) = \varphi\big((l, x)\big)$$

Логистическая функция

$$\varphi(y) = \frac{1}{1 + e^{-x}}$$

$f(x)$ будем трактовать как вероятность того, что образец x принадлежит первому классу

Рассмотрим обучающую выборку $V = \{(z_1, x_1), \dots, (z_m, x_m)\}$, в которой $z_i \in \{0,1\}$

Равенство $z_i = 1$ означает, что образец x_i принадлежит первому классу, равенство $z_i = 0$ означает, что образец x_i принадлежит второму классу

Логарифм максимального правдоподобия выглядит следующим образом:

$$lnL(l) = \sum_{i=1}^{m} \left(z_i ln\varphi\big((l, x_i)\big) + (1 - z_i) ln\left(1 - \varphi\big((l, x_i)\big)\right)\right)$$

Задача обучения:

$$\min_{l} \sum_{i=1}^{m} \left(z_i ln\varphi\big((l, x_i)\big) + (1 - z_i) ln\left(1 - \varphi\big((l, x_i)\big)\right)\right)$$

Представим задачу обучения как игру, протекающую в несколько раундов. В каждом раунде наш противник выбирает ход y_t, мы выбираем ход x_t. В результате наши потери составляют $W(x_t, y_t)$

За T раундов потери составят $\sum_{t=1}^{T} W(x_t, y_t)$

Будем считать, что выбор противника случаен, то естьy_t – независимые случайные величины с неизвестным нам законом распределения вероятностей – $Law(y)$. Противник выбирает закон распределения вероятностей вначале игры и не меняет его на протяжении игры. В

этом случае, естественно, рассматривать средние потери: $E\sum_{t=1}^{T} W(x_t, y_t)$

Потери: $W(x_t, y_t) = (x_t - y_t)^2$

С данной функцией потерь оптимально было бы в каждом раунде игры t выбирать $x_t = Ey_1$, в результате применения оптимальной стратегии средние потери за T раундов составили бы величину $T\sigma^2$

Сожаление:

$$R_T(\pi) = E\sum_{t=1}^{T}(x_t - y_t)^2 - T\sigma^2$$

Стратегию следует считать успешной или выигрышной, если $\lim_{T\to 0} \frac{R_T(\pi)}{T} = 0$

Перепишем сожаление следующим образом:

$$R_T(\pi) = E\sum_{t=1}^{T}(x_t - y_t)^2 - \min_x E\sum_{t=1}^{T}(x - y_t)^2$$

Можно удалить математическое ожидание:

$$R_T(\pi, y_1, \dots, y_T) = \sum_{t=1}^{T}(x_t - y_t)^2 - \min_x \sum_{t=1}^{T}(x - y_t)^2$$

Лекция 7

В момент времени t мы обладаем определенной информацией относительно целевой функции $F(x) = \sum_{t=1}^{T}(x - y_t)^2$

Будем считать, что в момент времени t когда мы выбираем x_t, нам известна только часть последовательности $y - y_1, \ldots, y_{t-1}$

Наилучшее решение для этой части последовательности

$$x_t^* = \frac{1}{t-1}\sum_{i=1}^{t-1} y_i$$

Пусть имеется последовательность штрафов $l_t(x)$.
Выразим сожаление $R_T(\pi, u) = \sum_{t=1}^{T} l_t(x_t) - \sum_{t=1}^{T} l_t(u)$ относительно произвольного решения u

Вернемся к анализу предложенной стратегии в нашем простом примере. Отметим, что в примере $l_t(x) = (x - y_t)^2$

Лемма. Пусть $l_t(x)$ – последовательность штрафов и x_t^* – минимизатор кумулятивных потерь за t раундов, тогда

$$\sum_{t=1}^{T} l_t(x_t^*) \leq \sum_{t=1}^{T} l_t(x_T^*)$$

Доказательство проведем методом математической индукции. Для $T = 1$ неравенство очевидно. Пусть оно выполняется для $T - 1, T \geq 2$.

$$\sum_{t=1}^{T-1} l_t(x_t^*) \leq \sum_{t=1}^{T-1} l_t(x_{T-1}^*).$$

Последние слагаемые в суммах $\sum_{t=1}^{T} l_t(x_t^*)$ и $\sum_{t=1}^{T} l_t(x_T^*)$ одинаковые, поэтому сравним суммы $\sum_{t=1}^{T-1} l_t(x_t^*)$ и $\sum_{t=1}^{T-1} l_t(x_T^*)$. Из индуктивного предположения следует неравенство $\sum_{t=1}^{T-1} l_t(x_t^*) \leq \sum_{t=1}^{T-1} l_t(x_{T-1}^*)$. Из определения x_{T-1}^* следует неравенство $\sum_{t=1}^{T-1} l_t(x_{T-1}^*) \leq \sum_{t=1}^{T-1} l_t(x_T^*)$. Из последнего и предпоследнего неравенств следует неравенство: $\sum_{t=1}^{T} l_t(x_t^*) \leq \sum_{t=1}^{T} l_t(x_T^*)$.

То есть, при известном будущем, адаптивная стратегия выглядит предпочтительней. Эта лемма поможет доказать следующую теорему

Теорема. Для произвольной равномерно ограниченной последовательности y, последовательности штрафов$l_t(x) = (x - y_t)^2$справедливо неравенство:

$$R_t(\pi) \leq x_0^2 + M^2 + 4M^2(1 + ln\, T).$$

В неравенстве $M = sup|y_t|$.

Применив лемму, получим первое неравенство:

$$R_T(\pi) \leq \sum_{t=1}^{T} |l(x_{t-1}^*) - l(x_t^*)|.$$

Далее используем последовательность потерь и равномерную ограниченность последовательности y:

$R_T(\pi) \le x_0^2 + M^2 + 2\sum_{t=1}^{T}|y_t|\,|x_t^* - x_{t-1}^*| \le x_0^2 + M^2 + 2M\sum_{t=1}^{T}|x_t^* - x_{t-1}^*|$.

Подставим x_t^* и x_{t-1}^* в последнее неравенство и в результате получим очередное неравенство:

$R_T(\pi) \le x_0^2 + M^2 + 2M\sum_{t=1}^{T}\left[\frac{1}{t(t-1)}\sum_{j=1}^{t-1}|y_j| + \frac{|y_t|}{t}\right] =$

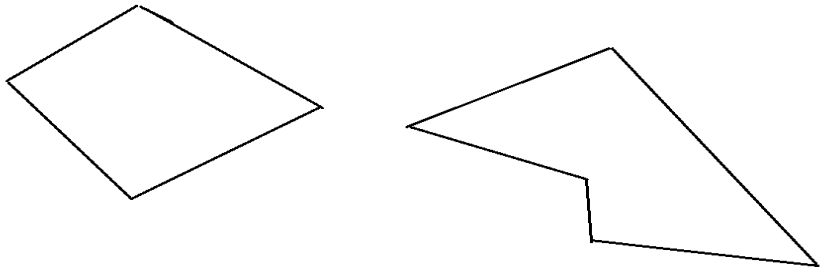
$x_0^2 + M^2 + 4M^2\sum_{t=1}^{T}\frac{1}{t}$.

Далее используем неравенство $\sum_{t=1}^{T}\frac{1}{t} \le 1 + ln\,T$ и получим доказательство теоремы.

Из теоремы следует, что для предлагаемой стратегии $\lim_{T\to\infty} \frac{R_T(\pi)}{T} = 0$, то есть, согласно определению, данная стратегия является выигрышной

Элементы выпуклого анализа

Выпуклым множеством называется множество, которое вместе с любыми своими точками содержит отрезок, который их соединяет. То есть для произвольных x и y, принадлежащих выпуклому множеству S, и для произвольного λ, принадлежащего интервалу $[0,1]$, выпуклая комбинация $\lambda x + (1-\lambda)y$принадлежит выпуклому множеству S



Выпуклое (слева) и невыпуклое (справа) множества

Множества A и B называют **отделимыми**, если существует такой вектор$_a$, что $(a, x) \leq (a, y)$, для любых $x \in A, y \in B$, и хотя бы для одной пары выполняется строгое неравенство $(a, \bar{x}) < (a, \bar{y})$

Эквивалентное определение отделимости. Множества A и B называют отделимыми, если существует такая гиперплоскость $\Pi: (a, x) = b$, что $(a, x) \leq b, (a, y) \geq b, \forall x \in A, \forall y \in B,$ и хотя бы одно из множеств не лежит целиком в гиперплоскости Π. Гиперплоскость Π при этом называют **разделяющей**

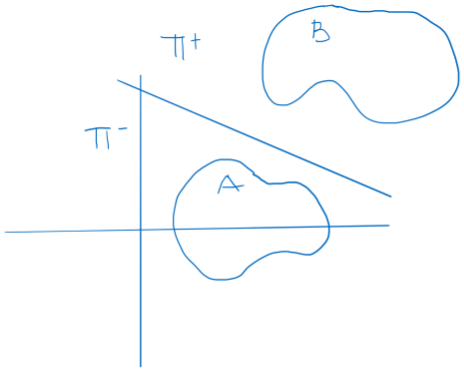


Рис.7. Разделяющая гиперплоскость

Множества A и B называют **сильно отделимыми**, если существует такой вектор a, что $\sup_{x \in A}(a, x) < \inf_{y \in B}(a, y)$.

Очень важную роль в выпуклом анализе играют **теоремы отделимости**.

Первая теорема звучит так. Непересекающиеся выпуклые множества отделимы.

Вторая теорема. Непересекающиеся замкнутые выпуклые множества, хотя бы одно из которых – ограниченное множество, сильно отделимы.

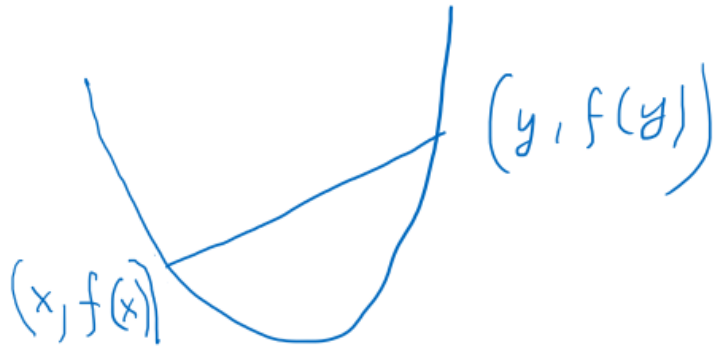
Тесно связана с этими теоремами теорема об опорной гиперплоскости.

Третья теорема – теорема об опорной гиперплоскости. Пусть x_0– граничная точка выпуклого множества, тогда существует такая гиперплоскость $П: (a(x - x_0), x - x_0) = 0$, проходящая через эту точку, что выпуклое множество вложено в $П^+$.

Выпуклой функцией называется функция, для которой область определения выпуклое множество и выполняется неравенство:

$$f(\lambda x + (1 - \lambda)y) \leq \lambda f(x) + (1 - \lambda)f(y)$$

для любых x и y из области определения функции и любых λ из интервала $[0,1]$



Хорда и график выпуклой функции

Эквивалентное определение выпуклой функции. Функция называется **выпуклой функций**, если ее область определения выпуклое множество и ее

надграфик $\{(x,y): x \in D, f(x) \leq y\}$– выпуклое множество. Здесь D– область определения функции.

Функция $f(x)$ называется **вогнутой**, если $-f(x)$– выпуклая функция.

Пусть $f_1(x),\ldots,f_m(x)$– последовательность выпуклых функций. Показать, что функции $\sum_{i=1}^{m}\alpha_i f_i(x)\,,\alpha_i \geq 0$, и$max\{f_1(x),\ldots,f_m(x)\}$ – выпуклые функции.

Функция $f(x)=x^2$является выпуклой, поэтому функция $f(x)=\sum_{i=1}^{m}\alpha_i x_i^2\,,\alpha_i \geq 0$ является выпуклой. Функции $f_1(x)=x, f_2(x)=-x$ являются выпуклыми, поэтому функция$|x| = max\{x,-x\}$ является выпуклой.

Рассмотрим произвольное семейство выпуклых функций $\{f_\alpha(x)\}_{\alpha\in A}$, функция $f(x)=\max\limits_{\alpha\in A} f_\alpha(x)$ является выпуклой.

Максимальное собственное число симметричной матрицы A является выпуклой функцией матрицы. Действительно, максимальное собственное число симметричной матрицы удовлетворяет равенству Рэлея $\lambda(A)=\max\limits_{x}\frac{(Ax,x)}{(x,x)}_{max}$. Семейство функций $f_x(A)=\frac{(Ax,x)}{(x,x)}$– семейство линейных функций, следовательно, семейство выпуклых функций. Поэтому $\lambda(A)_{max}$– выпуклая функция как максимум выпуклых функций.

Рассмотрим аффинное преобразование $Ax + b$ и выпуклую функцию $f(y)$. Показать, что суперпозиция $f(Ax + b)$ является выпуклой функцией.

Рассмотрим прямую линию $y = x + tv$, пересекающую область определения выпуклой функции $f(y)$, x принадлежит области определения, $v \in R^n$, t принадлежит интервалу $[t_i, t_s], t_i = inf\{t: x + tv \in D\}, t_s = sup\{t: x + tv \in D\}$. Доказать, что функция $\phi(t) = f(x + tv)$ является выпуклой функцией от одной переменной на интервале $[t_i, t_s]$.

Функция называется **строго выпуклой**, если

$$f(\lambda x + (1 - \lambda)y) < \lambda f(x) + (1 - \lambda)f(y),$$

для всех $x \neq y$ из области определения функции, и всех $\lambda \in (0,1)$

Первый критерий для дифференцируемых функций звучит следующим образом. Дифференцируемая функция –выпуклая функция тогда и только тогда, когда для любых x и y из области определения функции выполняется неравенство:

$$f(y) - f(x) \geq (f'(x), x - y).$$

В неравенстве $f'(x)$– вектор частных производных или градиент. Рис.4 является иллюстрацией первого критерия.

Лекция 7

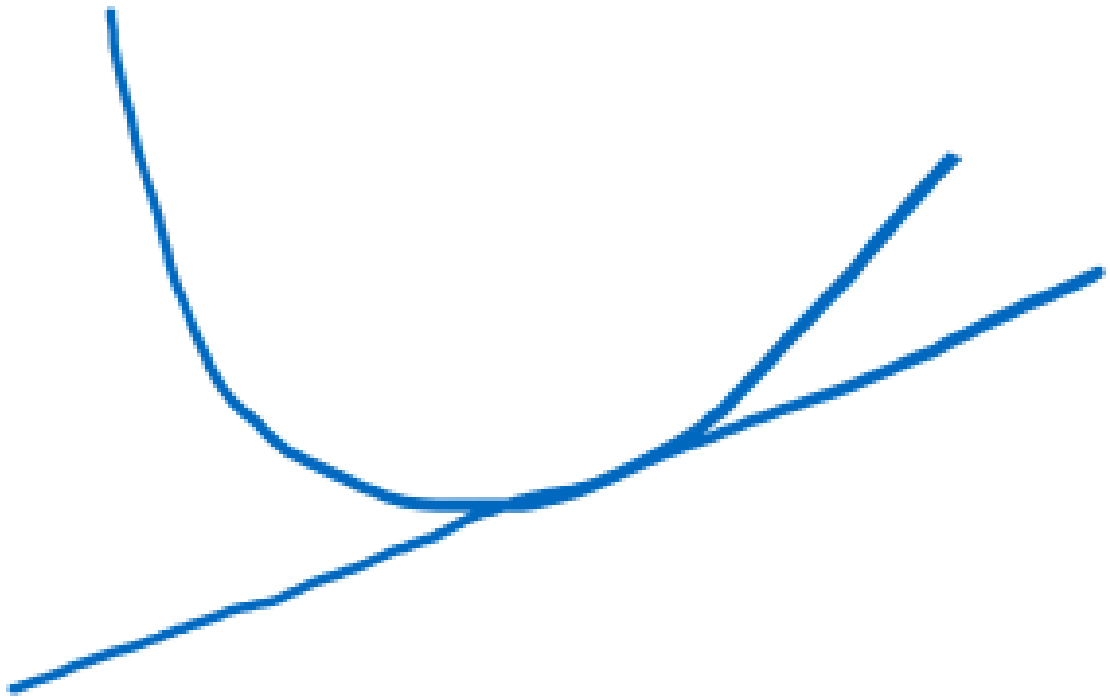


График выпуклой функции и касательная

Доказательство.

Необходимость.Начнем с неравенства, которое является определением выпуклой функции$f\big(x + \lambda(y - x)\big) \leq (1 - \lambda)f(x) + \lambda f(y).$

От правой и левой части отнимем $f(x)$, в результате получим неравенство

$$f\big(x + \lambda(y - x)\big) - f(x) \leq \lambda\big(f(y) - f(x)\big).$$

Разделим левую и правую часть неравенства на положительное λ

$$\frac{f\big(x+\lambda(y-x)\big)-f(x)}{\lambda} \leq f(y) - f(x).$$

Вычислим предел от левой и правой частей неравенства

$$\lim_{\lambda \downarrow 0} \frac{f\left(x+\lambda(y-x)\right)-f(x)}{\lambda} \leq f(y)-f(x).$$

Из дифференцируемости функции следует окончательное неравенство

$$(f'(x), y-x) \leq f(y)-f(x).$$

Достаточность.Начнем с неравенств

$$f(y)-f(\lambda x+(1-\lambda)y) \geq \lambda(f'(\lambda x+(1-\lambda)y), y-x), f(x)-f(\lambda x+(1-\lambda)y) \geq -(1-\lambda)(f'(\lambda x+(1-\lambda)y), y-x).$$

Левую и правую часть первого неравенства умножим на $1-\lambda$, левую и правую часть второго неравенства умножим на λ, $\lambda \in [0,1]$. Результаты сложим

$$\lambda f(x) + (1-\lambda) f(y) - f(\lambda x + (1-\lambda) y) \geq 0.$$

Данное неравенство доказывает, что функция $f(x)$– выпуклая функция

Критерий второго порядка звучит следующим образом. Функция $f(x)$– выпуклая функция тогда и только тогда, когда матрица вторых производных неотрицательно определена для всех значений аргумента из области определения функции.

Доказательство.

Необходимость.Воспользуемся формулой Тейлора второго порядка

$$f(y) = f(x) + (f'(x), y - x) + \frac{1}{2}(f''(x)(y - x), y - x) + o(\|y - x\|^2)$$

Здесь и далее, пока это не будет оговорено особо, $\|x\|$ – евклидова норма. В формуле Тейлора остаточный член обладает свойством $\lim\limits_{y\to x}\frac{o(\|y-x\|^2)}{\|y-x\|^2}=0$. Положим $y=x+\theta h, \|h\|=1, \theta>0$ (при достаточно малом h точка y будет принадлежать области определения функции, если область определения – открытое множество). Воспользуемся первым критерием ив результате получим неравенство:

$$\frac{\theta^2}{2}(f''(x)h,h)+o(\theta^2)\geq 0.$$

Разделим обе части неравенства на положительное число θ и перейдем к пределу при $\theta \downarrow 0$. В результате получим неравенство

$$(f''(x)h, h) \geq 0,$$

из которого следует неотрицательная определенность матрицы.

Достаточность.Воспользуемся формулой Тейлора первого порядка с остаточным членом в форме Лагранжа

$$f(y) = f(x) + (f'(x), y - x) + \frac{1}{2}\Big(f''(x)\big(x + \alpha(y - x)\big), x + \alpha(y - x)\Big),$$

где α– некоторое число из интервала [0,1]. Из неотрицательной определенности матрицы вторых производных и первого критерия следует выпуклость функции $f(x)$

Локальные и глобальные минимумы функции

Критерий локального минимума. Точка x_0является точкой локального минимума функции тогда и только тогда, когда вектор частных производных$f'(x_0) = 0$ и матрица частных производных второго порядка $f''(x_0)$– неотрицательно определена, то есть $(f''(x_0)h, h) \geq 0, \forall h.$

Доказательство.

Необходимость.Выберем $\alpha > 0$таким образом, чтобы точки $y_1 = x_0 + \alpha h$ и $y_2 = x_0 - \alpha h$ принадлежали окрестности точки x_0. Для этих точек применим формулу Тейлора первого порядка$f(y) - f(x) = (f'(x_0), y - x) + o(\|y - x\|)$. В результате получим одновременное выполнение двух неравенств

$$(f'(x_0), h) \geq 0 \text{ и } (f'(x_0), h) \leq 0$$

одновременно для любого h. Из одновременного выполнения двух неравенств следует справедливость равенства $(f'(x_0), h) = 0$ для всех h. Отсюда вытекает равенство нулю вектора частных производных$f'(x_0) = 0$.

Для точки y_1 применим формулу Тейлора второго порядка $f(y) - f(x) = (f'(x_0), y - x) + \frac{1}{2}(f''(x_0)(y - x), y - x) + o(\|y - x\|^2)$, в результате получим неравенство

$$\frac{\alpha^2}{2}(f''(x_0)h, h) + o(\alpha^2) \geq 0.$$

Разделив обе части неравенства на $\alpha > 0$ и, вычислив предел при $\alpha \downarrow 0$ от обеих частей, получим неравенство

$$(f''(x_0)h, h) \geq 0, \text{ для всех } h,$$

которое означает неотрицательную определенность матрицы вторых частных производных.

Достаточность. Для yиз окрестности точки x_0 применим формулу Тейлора первого порядка с остаточным членом в форме Лагранжа $f(y) - f(x_0) = \frac{1}{2}\left(f''\left(x_0 + \beta(y - x_0)\right)(y - x_0), y - x_0\right)$. Из неотрицательной определенности матрицы вторых частных производных следует неравенство$f(y) - f(x_0) \geq 0$, которое устанавливает, что x_0 является точкой локального минимума

Критерий минимума выпуклой функции. Точка x_0 является точкой минимума выпуклой функции $f(x)$ на выпуклом множестве S тогда и только тогда, когда для всех $y \in S$ выполняется неравенство

$$(f'(x), y - x_0) \geq 0.$$

Доказательство.

Необходимость.Используем формулу Тейлора первого порядка

$$f(x_0 + \alpha h) - f(x_0) = \alpha(f'(x_0), h) + o(\alpha), \|h\| = 1, \alpha > 0.$$

Для вывода неравенства

$$\alpha(f'(x_0), h) + o(\alpha) \geq 0$$

разделим левую и правую часть неравенства на α и вычислим от обеих частей предел при $\alpha \downarrow 0$. В результате получим неравенство

$$(f'(x), y - x_0) \geq 0.$$

Достаточность. Неравенство $f(y) - f(x_0) \geq 0$ следует из первого критерия выпуклости функции $f(y) - f(x_0) \geq (f'(x_0), y - x_0)$.

Неравенство выполняется для всех $y \in S$, поэтому точка x_0 является точкой минимума функции $f(x)$

Метод проекции субградиента

Под проекцией точки y на множество S – $P_S(y)$ – понимается решение задачи

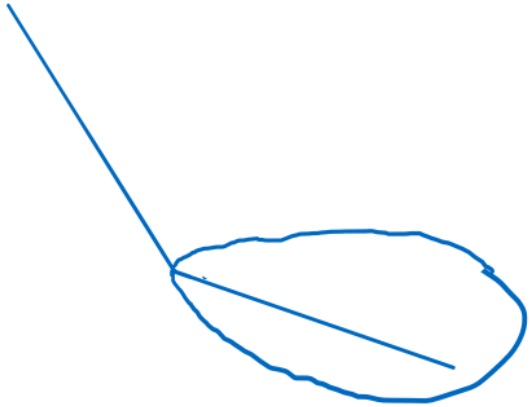
$$\min_{x \in S} \|y - x\|$$

Теорема о проекции точки. Если множество S – выпуклое и замкнутое множество, то проекция точки y – $P_S(y)$существует, причем проекция – единственная.

Проекция характеризуется неравенством тупого угла (см. рис.5):

$$\big(y - P_s(y), x - P_s(y)\big) \leq 0, \forall x \in S,$$

которое непосредственно выводится из критерия оптимальности.



Неравенство тупого угла.

Теорема. Проекция является нерасширяющим отображением, то есть

$$\|P_S(y) - P_S(x)\| \leq \|y - x\|.$$

Для доказательства дважды воспользуемся неравенством тупого угла:

$\left(x - P_S(x), P_S(y) - P_S(x)\right) \leq 0, \left(y - P_S(y), P_S(x) - P_S(y)\right) \leq 0.$

Складывая эти неравенства, получим неравенство:

$$\left(y - x - \left(P_S(y) - P_S(x)\right), P_S(y) - P_S(x)\right) \geq 0.$$

Рассмотрим два вектора, для которых выполняется неравенство: $(a - b, b) \geq 0$. С помощью несложных преобразований $\|a\|^2 = \|a - b + b\|^2 = \|a - b\|^2 + 2(a - b, b) + \|b\|^2 \geq \|b\|^2$ устанавливается, что $\|a\|^2 \geq \|b\|^2$. Теперь, чтобы доказать теорему, достаточно положить $a = y - x, b = P_S(y) - P_S(x)$

Метод проекции градиента заключается в генерации последовательности точек:

$$x_{t+1} = P_S\left(x_t - h_t f'(x_t)\right),$$

h_t– выбранная подходящим способом последовательность шагов.

При $S = R^n$

$$x_{t+1} = x_t - h_t f'(x_t)$$

Неравенство должно выполняться для всех y из области определения функции. Множество всех субградиентов называется **субдифференциалом** функции –$\partial f(x)$

Определение.Если $\partial f(x) \neq \emptyset$, то функция субдифференцируемая в точке (x).

Выпуклая функция $y = |x|$– недифференцируемая функция при $x = 0$

В качестве упражнения доказать, что $\partial|x| =$
$$\begin{cases} -1, x < 0 \\ [-1,1], x = 0 \\ 1, x > 0 \end{cases}$$

Рассмотрим функцию, которая является максимумом семейства выпуклых функций

$F(x) = max\{f_1(x), f_2(x), \ldots, f_m(x)\}$. Если каждая функция семейства – субдифференцируемая функция в точке x, функции семейства – непрерывные функции в точке x, то функция $F(x)$– субдифференцируемая функция в точке x, причем субдифференциал этой функции – выпуклая оболочкаобъединения активной части субдифференциалов функций семейства:$\partial F(x) = conv \bigcup_{i \in A(x)} \partial f_i(x), A(x) = \{j: f_j(x) = F(x)\}$

Теорема Если функция $f(x)$ удовлетворяет условию Липшица: $|f(y) - f(x)| \leq L\|y - x\|$, то субградиент равномерно ограничен этой константой: $\|g(x)\| \leq L$.

Доказательство Выберем $y = x + \varepsilon \frac{g}{\|g\|}, \varepsilon > 0$ и подставим в определяющее субградиент неравенство:

$$L\varepsilon \geq |f(y) - f(x)| \geq f(y) - f(x) \geq (g, y - x) = \varepsilon \|g\|.$$

Разделим на $\varepsilon > 0$ и получим необходимое неравенство

Теорема 2. Если субградиент функции $f(x)$ равномерно ограничен константой L: $\|g\| \leq L$, то функция удовлетворяет условию Липшица:

$$|f(y) - f(x)| \leq L\|y - x\|$$

Доказательство. Используем определение субградиента в точке x и неравенство Коши:

$$f(x) - f(y) \leq (-g, y - x) \leq |(-g, y - x)| \leq \|g\|\|y - x\| \leq L\|y - x\|.$$

Аналогичным образом используем определение субградиента в точке y:

$$f(y) - f(x) \leq L\|y - x\|$$

Отсюда и из предыдущего неравенства следует доказательство теоремы

Формула генерации последовательности точек:

$$x_{t+1} = P_S(x_t - h_t g_t), g_t \in \partial f(x_t)$$

Теорема. Пусть выпуклая функция $f(x)$ удовлетворяет условию Липшица: $|f(x) - f(y)| \leq L\|x - y\|$, тогда для

последовательности x_t, полученной проекцией субградиента, справедливо неравенство:

$$f_T^* - f^* \leq \frac{\|x_1 - x_*\|^2 + L^2 \sum_{t=1}^{T} h_t^2}{2 \sum_{t=1}^{T} h_t}.$$

В этом неравенстве $f_T^* = \min\limits_{1 \leq t \leq T} f(x_t), f^* = \min\limits_{x \in S} f(x), x^* = arg \min\limits_{x \in S} f(x)$

Доказательство.Используем обозначение $r_t^2 = \|x_t - x^*\|^2$. Рассмотрим разность $r_{t+1}^2 - r_t^2 = \|P_s(x_t - h_t g_t) - x^*\|^2 - \|x_t - x^*\|^2 \leq \|x_t - x^* - h_t g_t\|^2 - \|x_t - x^*\|^2 = -2h_t(x_t - x^*, g_t) + h_t^2 \|g_t\|^2$

Из определения субградиента следует неравенство $2h_t\big(f(x_t)-f(x^*)\big)\le r_t^2-r_{t+1}^2+h_t^2\|g_t\|^2$.
Просуммируем левую и правую часть этого неравенства и в результате получим неравенство: $2\sum_{t=1}^{T}h_t\left(f(x_t)-f(x^*)\right)\le r_1^2+\sum_{t=1}^{T}h_t^2\|g_t\|^2$. Отсюда несложно получить доказательство теоремы.

Поскольку оптимальное значение x^* нам неизвестно, то мы предположим, что $\|x_1-x_*\|\le R$. При постоянном шаге $h_t=h$, неравенство приобретает вид:

$$f_T^*-f^*\le\frac{R^2+L^2Th^2}{2Th}.$$

Минимальное значение правой части неравенства достигается при шаге $h = \sqrt{\frac{R^2}{TL^2}}$. Для такого шага выполняется неравенство: $f_T^* - f^* \leq \frac{RL}{\sqrt{T}}$.

Шаг h зависит от числа итерацийTи от трудно определяемых констант Rи L. Хорошимсвойством оценки является ее стремление к нулю при T, стремящемся к бесконечности. Задав погрешностьε, можно определить число итераций $T \geq \frac{R^2 L^2}{\varepsilon^2}$

Основная задача – разработка стратегии вычисления последовательности решений x_t, для которой отношение сожаления к периоду

$\frac{R_T(x_1,\dots,x_T;u)}{T} = \frac{1}{T}\left(\sum_{t=1}^{T} l_t(x_t) - \sum_{t=1}^{T} l_t(u)\right) \longrightarrow 0$ при $T \longrightarrow 0$.

Стратегия «Следуй за лидером»:

$$x_{t+1} = arg\min_{x} \sum_{i=1}^{t} l_i(x)$$

Алгоритм типа субградиентного спуска:

$$x_{t+1} = P_v(x_t - h_t g_t), g_t \in \partial l_t(x_t)$$

Оценим сожаление для выпуклых потерь, удовлетворяющих условию Липшица

Обозначим через:

$$r_t = \|x_t - u\|^2$$

Справедливы неравенства:

$$r_{t+1}^2 - r_t^2 \le \|x_t - u - \eta_t g_t\|^2 - \|x_t - u\|^2$$
$$= -2\eta_t (g_t, x_t - u) + \eta_t^2 \|g_t\|^2$$

$$r_{t+1}^2 - r_t^2 \le 2\eta_t \big(l_t(u) - l_t(x_t)\big) + \eta_t^2 \|g_t\|^2$$

Оценка сожаления имеет вид:

$$R(x_1, x_2, \ldots, u) \leq \frac{1}{2}\sum_{t=1}^{T}\frac{r_t^2 - r_{t+1}^2}{\square_t} + \frac{L^2}{2}\sum_{t=1}^{T}\square_t$$

При постоянном шаге:

$$R(x_1, x_2, \ldots, u) \leq \frac{D}{2\square} + \frac{L^2 T\square}{2}$$

Для шага:

$$\square = \sqrt{\frac{D^2}{L^2 T}}$$

Сожаление:

$$R(x_1, x_2, \ldots, u) \leq LD\sqrt{T}$$

Задача стохастического программирования:

$$\min_{x \in V} F(x) = \min_{x \in V} Ef(x, \xi)$$

Метод проекции стохастического градиента:

$$x_{t+1} = P_V\left(x_t - \square_{t+1} g_{\xi_{t+1}}(x_t)\right)$$

$g(x, \xi)$– субградиент функции$f(x, \xi)$

Справедливо неравенство:

$$EF\left(\sum_{t=1}^{T} \nu_t x_t\right) - F(x_*) \leq \frac{R^2 + L^2 \sum_{t=1}^{T} ?_t^2}{2 \sum_{t=1}^{T} ?_t}$$

Для постоянного шага неравенство имеет следующий вид:

$$EF\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - F(x_*) \leq \frac{R^2 + L^2 T h^2}{2Th}$$

Для $□_* = \frac{R}{L\sqrt{T}}$

$$EF\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - F(x_*) \leq \frac{RL}{\sqrt{T}}$$

Отсюда

$EF\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right)$ стремится к $F(x_*)$ со скоростью $\frac{1}{\sqrt{T}}$

Альтернативный способ вычисления условного минимума регрессии заключается в следующем.

Выбирается достаточно большое значение для T и рассматривается приближенная задача:

$$\min_{x \in V} F_e(x) = \min_{x \in V} \frac{1}{T} \sum_{i=1}^{T} f(x, \xi_i).$$

Отметим, что задача решается для фиксированного набора значений

Теорема. Если дисперсия случайной величины $f(x, \xi)$ равномерно ограничена на множестве V: $\sup_{x \in V} Df(x, \xi) \leq C$, то

$$P\left(\left|F(x) - \frac{1}{T} \sum_{t=1}^{T} f(x, \xi)\right| \leq \varepsilon\right) \geq \sigma, \text{ для } T \geq \frac{C}{(1-\sigma)\varepsilon^2}.$$

Доказательство опирается непосредственно на неравенство Чебышева: $P(|\eta - E\eta| \leq \varepsilon) \geq 1 - \frac{D\eta}{\varepsilon^2}$.

Среднее $E\frac{1}{T}\sum_{i=1}^{T} f(x, \xi_i) = F(x)$,

дисперсия$D\frac{1}{T}\sum_{i=1}^{T} f(x, \xi_i) = \frac{Df(x,\xi)}{T} \leq \frac{C}{T}$

Рассмотрим случайную величину θ, равновероятно распределенную на множестве значений $\{1, \ldots, T\}$, и случайный вектор $g_\theta(x)$. Математическое ожидание

случайного вектора $Eg_\theta(x) = \frac{1}{T}\sum_{i=1}^{T} g_i(x)$. Применив теорему, получим, что субградиент функции $F_e(x)$–$g_e(x)$ равен $Eg_\theta(x)$. Следовательно, $g_\theta(x)$–стохастический субградиент функции $F_e(x)$

Метод проекции стохастического субградиента заключается в генерации последовательности:

$$x_{t+1} = P_v\left(x_t - h_{t+1}g_{\theta_{t+1}}(x_t)\right)$$

Для T итераций при постоянном шаге $h^* = \frac{R}{L\sqrt{T}}$ получим оценку

$$EF_e\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - F_e(x_*) \leq \frac{RL}{\sqrt{T}}$$

Рассмотрим задачу бинарной классификации, точнее, задачу обучения для решающего правила вида: $d(x) = \begin{cases} 1, (l,x) \geq 0 \\ 2, (l,x) < 0 \end{cases}$. Дана обучающая выборка $V = \{(x_i, y_i)\}$ объема N. Элементы $y_i \in \{-1,1\}$. Предполагается, что множества $V_1 = (x_i : y_i = 1)$ и $V_2 = (x_i : y_i = -1)$линейно не разделяются,то есть не существует

гиперплоскости $(l, x) = 0$, что $(l, x) \geq 0$ для всех $x \in V_1$ и $(l, x) < 0$для всех $x \in V_2$. Рассмотрим штраф $max\{1 - y(l, x), 0\}$,с помощью которого определяется качество решающего правила для обучающей выборки:

$$F(l) = \frac{1}{N} \sum_{i=1}^{N} max\{1 - y_i(l, x_i), 0\}$$

Таким образом, задача обучения заключается в вычислении минимума функции$F(l)$.

Для использования метода спуска по стохастическому субградиенту требуется вычислить субградиент функции $f_i(l) = max\{1 - y_i(l, x_i), 0\}$. Для этого найдем

субградиент функции от одной переменной. По определению субградиента $g_\phi(x) = \begin{cases} 1, x > 0 \\ \alpha, x = 0\,, \alpha \in [0,1] \\ 0, x < 0 \end{cases}$.

Отсюда субградиент функции $g_i(l) =$

$$\begin{cases} -y_i x_i, 1 - y_i(l, x_i) > 0 \\ \alpha(-y_i x_i), 1 - y_i(l, x_i) = 0. \\ 0, 1 - y_i(l, x_i) < 0 \end{cases}$$

Знание субградиента позволяет для решения задачи обучения применить описанную выше процедуру спуска по стохастическому субградиенту

Задача обучения в общей постановке

Рассматривая задачу обучения в более общей постановке, мы предположим, что у нас имеется решающее правило, точнее класс решающих правил $\varphi_l(x)$, параметризованных вектором l. Мы можем рассмотреть случайный вектор $\xi = \begin{pmatrix} x \\ y \end{pmatrix}$ и переопределить вектор параметров: $l := \begin{pmatrix} l \\ -1 \end{pmatrix}$. Функция $f(l, \xi) = [(l, \xi)]^2$. Стохастический градиент $g_\xi(l) = (l, \xi)\xi$. Проекция вектора z на множество V вычисляется достаточно просто: $\left(P_V(z)\right)_i = \begin{cases} z_i, i \neq d+1 \\ -1, i = d+1 \end{cases}$. Эта

задача непосредственно связана с задачей линейного прогноза ненаблюдаемой случайной величины y по наблюдаемым случайным величинам x.

Рассмотренный в этом разделе подход к обучению называется методом минимизации эмпирического риска. Основной смысл этого метода заключается в замене целевой функции задачи обучения $R(x) = Ef(x, \xi)$ на функцию $R_e(x) = \frac{1}{T}\sum_{t=1}^{T} f(x, \xi_t)$.Эта замена связана с тем, что закон распределения случайной величины ξ при решении задачи обучения неизвестен

Определение. Функция $f(x)$ называется μ**-сильно выпуклой**, если для нее существует оценка снизу разности $f(y)-f(x)$:

$$f(y)-f(x)\geq(g,y-x)+\frac{\mu}{2}\|y-x\|^2.$$

Является очевидным утверждение.

Теорема. Пусть $f(x)$– функция выпуклая и дважды дифференцируемая. Если выполняется неравенство $(f''(x)y,y)\geq\mu\|y\|^2$, для всех$x$, то она$\mu$-сильно выпуклая функция.

Лекция 9

Online субградиентный спуск для сильно выпуклых функций

Рассмотрим основную задачу, которая изучается в этом пособии. Имеется последовательность штрафов $l_t(x)$. Функции $l_t(x)$определены на выпуклом и замкнутом множестве V, на котором они являются сильно выпуклыми функциями:

$$l_t(y) - l_t(x) \geq (g_t, y - x) + \frac{\mu_t}{2}\|y - x\|$$

Рассмотрим online субградиентный спуск, порождающий последовательность стратегий:

$$x_{t+1} = P_V\left(x_t - h_t g_t\right).$$

Применив методику, изложенную ранее можем доказать следующую оценку для разности $l_t(x_t) - l_t(u)$:

$$l_t(x_t) - l_t(u) \le \left(\frac{1}{2h_t} - \frac{\mu_t}{2}\right)\|x_t - u\|^2 - \frac{1}{2h_t}\|x_{t+1} - u\|^2 + \frac{h_t}{2}\|g_t\|^2$$

Выберем шаг таким образом, чтобы

$$\frac{1}{2h_t} - \frac{\mu_t}{2} = \frac{1}{2h_{t-1}},$$

Положив

$$h_1 = \frac{1}{\mu_1}$$

Применив метод математической индукции нетрудно доказать формулу для шага:

$$h_t = \frac{1}{\mu_1 + \mu_2 + \ldots + \mu_t}.$$

С этим шагом оценка разности будет иметь вид:

$$l_t(x_t) - l_t(u) \le \frac{1}{h_{t-1}}\|x_t - u\|^2 - \frac{1}{2h_t}\|x_{t+1} - u\|^2 + \frac{h_t}{2}\|g_t\|^2, t = 2,\ldots,T; l_1(x_1) - l_1(u) \le$$

$$\le -\frac{\mu_1}{2}\|x_2 - u\|^2 + \frac{1}{2\mu_1}\|g_1\|^2.$$

Отсюда несложно получить оценку для сожаления:

$$R(u)=\sum_{t=1}^{T}\left(l_t(x_t)-l(u)\right)\le-\frac{\mu_1}{2}\|x_2-u\|^2+\frac{1}{2}\sum_{i=2}^{T}\left(\frac{1}{h_{i-1}}\|x_i-u\|^2-\frac{1}{h_i}\|x_{i+1}-u\|^2\right)+$$

$$+\frac{1}{2}\sum_{i=1}^{T}h_i\|g_i\|^2$$

После несложных преобразований получаем оценку для сожаления:

$$R(u)=\sum_{t=1}^{T}\left(l_t(x_t)-l(u)\right)\le-\frac{1}{2h_T}\|x_{T+1}-u\|^2+\frac{1}{2}\sum_{i=1}^{T}h_i\|g_i\|^2\le\frac{1}{2}\sum_{i=1}^{T}h_i\|g_i\|^2$$

Предположим, что $\mu_t\ge\mu>0$ и потери удовлетворяют условию Липшица с общей константой Липшица $-L$. При этих предположениях оценка сожаления будет выглядеть следующим образом:

$$R(u)=\sum_{t=1}^{T}\left(l_t(x_t)-l_t(u)\right)\le\frac{L^2}{2\mu}\left(1+\ln T\right).$$

Данная оценка сожаления для сильно выпуклых потерь лучше, чем оценка для выпуклых потерь, полученная ранее. Поскольку предыдущая оценка имела порядок роста $O(\sqrt{T})$, а данная оценка имеет порядок роста $O(\ln T)$. Но в обоих случаях алгоритм online субградиентного спуска является сублиненйным алгоритмом.

Лекция 10

Мартингалы

Стохастический базис $\langle \Omega, (F_t)_{t=0}^{T}, F, P \rangle$.

Последовательность σ - алгебр F_t удовлетворяет условиям:

$$F_0 = \sigma(\Omega, \bar{\Omega})$$– тривиальная алгебра,

$$F_t \subseteq F_{t+1}, F_t \subseteq F$$

X_t называется **адаптированной последовательностью**

Случайная последовательность является **предсказуемой последовательностью** , если X_t - измеримая случайная величина относительно σ - алгебры F_{t-1}, для всех t

Адаптированная последовательность X_t называется **мартингалом**, если выполняются условия:

$$E|X_t| < \infty,$$

$$E(X_t/F_{t-1}) = X_{t-1}$$

Рассмотрим мартингал с конечным числом членов $-T < \infty$, для мартингала справедливо равенство: $X_t = E(X_T/F_t)$ и мартингал$_{X_t}$ является равномерно интегрируемой последовательностью. Можно поступить иначе. Рассмотрим абсолютно интегрируемую случайную величину ξ. Построим случайную последовательность X_t следующим образом:

$$X_t = E(\xi/F_t)$$

данная последовательность является мартингалом

Риск, эмпирический риск и сожаление

Вернемся к исходной задаче минимизации риска

$$\min F(l) = \min Ef(l, \xi)$$

Допустим, мы применяем некоторый алгоритм, который генерирует решения $l_1, \ldots, l_t, \ldots$ Относительно решений, можно утверждать, что они предсказуемые относительно потока σ-алгебр $F_t = \sigma(\xi_1, \xi_2, \ldots, \xi_t)$

Сожаление, связанное с рассматриваемой задачей

$$R(l) = \sum_{i=1}^{T} f(l_i, \xi_i) - \sum_{i=1}^{T} f(l, \xi_i)$$

Рассмотрим последовательность

$$Z_t : Z_0 = 0, \Delta Z_t = F(l_t) - f(l_t, \xi_t)$$

Потребуем, чтобы случайные величины ΔZ_t были абсолютно интегрируемые

Условное математическое ожидание

$$E(\Delta Z_t / F_{t-1}) = E(F(l_t) - f(l_t, \xi_t) / F_{t-1})$$

Далее

$$E(\Delta Z_t / F_{t-1}) = F(l_t) - Ef(l_t, \xi_t) = 0$$

Следовательно, последовательность Z_t является мартингалом

Потребуем равномерную ограниченность для приращений:

$$\left|\Delta Z_t\right| \le C$$

тогда

$$P\left(Z_T < \varepsilon\right) \ge 1 - \exp\left(-\frac{\varepsilon^2}{2C^2T}\right)$$

Приравняв

$$\exp\left(-\frac{\varepsilon^2}{2C^2T}\right) = \delta,$$

получим равенство:

$$\varepsilon = C\sqrt{2T \ln \frac{1}{\delta}}$$

Отсюда

$$P\left(Z_T < C\sqrt{2T \ln \frac{1}{\delta}} \right) \geq 1 - \delta$$

В результате получим

$$P\left(\sum_{i=1}^{T} \left(F(l_i) - f(l_i, \xi_i) \right) < C\sqrt{2T \ln \frac{1}{\delta}} \right) \geq 1 - \delta$$

Воспользуемся определением сожаления

$$\sum_{i=1}^{T} f(l_i, \xi_i) = R(u) + \sum_{i=1}^{T} f(u, \xi_i)$$

и из предыдущего неравенства получим

$$P\left(\sum_{i=1}^{T} F(l_i) \le R(u) + \sum_{i=1}^{T} f(u, \xi_i) + C\sqrt{2T \ln \frac{1}{\delta}} \right) \ge 1 - \delta$$

Рассмотрим последовательность

$$\bar{Z}_t = \sum_{i=1}^{t} \left(F(u) - f(u, \xi_i) \right),$$

эта последовательность имеет независимые приращения

$$\Delta Z_t = F(u) - f(u, \xi_i)$$

Математическое ожидание

$$E\Delta Z_t = 0$$

Поэтому последовательность $\overline{Z}_t$ –мартингал.

При том же предположении о равномерной ограниченности приращений получаем неравенство

$$P\left(\sum_{i=1}^{T} f(u,\xi_i) \le TF(u) + C\sqrt{2T\ln\frac{1}{\delta}}\right) \ge 1-\delta$$

Таким образом

Лекция 10

$$P\left(\frac{1}{T}\sum_{i=1}^{T}F(l_i)\le\frac{R(u)}{T}+F(u)+2C\sqrt{\frac{2\ln\frac{1}{\delta}}{T}}\right)\ge 1-2\delta$$

Пусть

$$F(u^*)=\min F(u)$$

тогда

$$P\left(\frac{1}{T}\sum_{i=1}^{T}F(l_i)\le\frac{R(u^*)}{T}+F(u^*)+2C\sqrt{\frac{2\ln\frac{1}{\delta}}{T}}\right)\ge 1-2\delta$$

Если $F(l)$– выпуклая функция, то выполняется неравенство:

$$P\left(F\left(\frac{1}{T}\sum_{i=1}^{T} l_i \right) \le \frac{R(u^*)}{T} + F(u^*) + 2C\sqrt{\frac{2\ln\frac{1}{\delta}}{T}} \right) \ge 1 - 2\delta .$$

Если

$$\lim_{T\to\infty} \frac{R(u)}{T} = 0$$

то, выбрав достаточно большое значение для T и взяв в качестве решения среднее

$$\bar{u} = \frac{1}{T}\sum_{i=1}^{T} l_i$$

мы можем с вероятностью $1-2\delta$ оказаться сколь угодно близко от оптимального значения $F\left(u^*\right)$