Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»

УТВЕРЖДАЮ
Директор Института магистратуры
_____Иванова Е.А.
«__27__» июня__ 2023___г.

Рабочая программа дисциплины
Машинное обучение: математические основы

Направление 01.04.02 Прикладная математика и информатика
магистерская программа 01.04.02.04 "Искусственный интеллект: математические модели и прикладные решения"

Для набора 2023 года

Квалификация
Магистр

Составитель(и) программы:

Сахарова Л.В., д.ф.-м.н проф, кафедры высшей фундаментальной и прикладной математики

I. ЦЕЛИ И ЗАДАЧИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Цели освоения дисциплины: понимание основных принципов и алгоритмов машинного обучения, знакомство с необходимым математическим аппаратом, знакомство с различными моделями машинного обучения, освоение современного программного обеспечения необходимого для построения моделей и анализа данных, получения опыта работы с искусственными и реальными наборами данных.

Задачи: освоение стандартных методов и моделей машинного обучения: метод ближайших соседей, линейная регрессия, метод опорных векторов, решающие деревья, случайный лес, градиентный бустинг, нейронные сети, алгоритмы кластеризации, обучение с подкреплением, приобретение навыков построения моделей машинного обучения и работы с современным программным обеспечением (Python, библиотеки scikit-learn, keras).

II. МЕСТО ДИСЦИПЛИНЫ В СТРУКТУРЕ ОП ВО

2.1. Учебная дисциплина «Машинное обучение: математические основы» (2 семестр, 1 курс) относится к части блока дисциплин (модулей), формируемой участниками образовательных отношений, и является дисциплиной по выбору.

2.2. Для изучения данной учебной дисциплины необходимы знания, умения и навыки, формируемые предшествующими дисциплинами: “Методы оптимизации для машинного обучения”, “Избранные вопросы теории вероятностей и математической статистики”, “Питон для анализа данных”, “Основы нейронных сетей”.

2.3. Знания, умения и навыки, полученные в ходе изучения данной дисциплины будут полезны при изучении последующих дисциплин “Анализ временных рядов”, “Глубокое обучение”, “Нейронные сети для мобильных приложений”, “Обучение с подкреплением и приложения” и могут использоваться для решения профессиональных задач в научно-исследовательской, научно-производственной и проектной деятельности, в частности, при подготовке выпускной квалификационной работы.

III. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Процесс изучения дисциплины направлен на формирование элементов следующих компетенций в соответствии с ФГОС ВО и ОП ВО по данному направлению подготовки:

Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы:

Шифр и формулировка компетенций (результаты освоения ОП)	**Индикаторы компетенций**	**Элементы компетенций, формируемые дисциплиной**
Профессиональные компетенции (ПК)		
ПК-3. Способен разрабатывать и применять методы и алгоритмы машинного обучения для решения задач	ПК-3.1. Ставит задачи по разработке или совершенствованию методов и алгоритмов для решения комплекса задач предметной области	ПК-3.1. З-1. Знает классы методов и алгоритмов машинного обучения ПК-3.1. У-1. Умеет ставить задачи и разрабатывать новые методы и алгоритмы машинного обучения ПК-3.1. Н-1. Имеет навыки работы со стандартными методами и моделями машинного обучения: метод ближайших соседей, линейная регрессия, метод опорных векторов, решающие деревья, случайный лес, градиентный бустинг, нейронные сети, алгоритмами кластеризации. Знает различные модели обучения: обучение с учителем, обучение без учителя, онлайн обучение, обучение с подкреплением.
	ПК-3.2. Руководит исследовательской группой по разработке или совершенствованию методов и алгоритмов для решения комплекса задач предметной области	ПК 3.2. З-1. Знает методы и критерии оценки качества моделей машинного обучения ПК 3.2. У-1. Умеет определять критерии и метрики оценки результатов моделирования при построении систем искусственного интеллекта в исследуемой области ПК-3.2. Н-1. Имеет навыки работы с различными функциями потерь. Владеет

		методами кросс-валидации и регуляризации, позволяющими оценивать истинный риск и производить настройку параметров моделей. Владеет методами предобработки данных, понижения размерности, кластеризации. Имеет навыки работы с библиотеками scikit-learn, keras. Имеет опыт работы с искусственными и реальными наборами данных.
ПК-4. Способен руководить проектами по созданию комплексных систем искусственного интеллекта	ПК-4.1. Руководит разработкой архитектуры комплексных систем искусственного интеллекта	ПК-4.1. З-1. Знает возможности современных инструментальных средств и систем программирования для решения задач машинного обучения ПК-4.1. У-1. Умеет проводить сравнительный анализ и осуществлять выбор инструментальных средств для решения задач машинного обучения ПК-4.1. Н-1. Программирование алгоритмов машинного обучения на языке Python, навыки работы с различными моделями машинного обучения из библиотек scikit-learn, keras.
	ПК-4.2. Осуществляет руководство созданием комплексных систем искусственного интеллекта с применением новых методов и алгоритмов машинного обучения	ПК-4.2. З-1. Знает функциональность современных инструментальных средств и систем программирования в области создания моделей и методов машинного обучения ПК-4.2. З-2. Знает принципы построения систем искусственного интеллекта, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта ПК-4.2. У-1. Умеет применять

		современные инструментальные средства и системы программирования для разработки новых методов и моделей машинного обучения ПК-4.2. У-2. Умеет руководить выполнением коллективной проектной деятельности для создания, поддержки и использования систем искусственного интеллекта ПК-4.2. Н-1. Имеет навыки работы с различными алгоритмами из библиотеки scikit-learn. Имеет опыт их применения и сравнительного анализа при работе с искусственными и реальными данными. ПК-4.2. Н-2. Умеет выбирать и настраивать алгоритмы машинного обучения для решения конкретных задач. Имеет опыт работы над простыми проектами (индивидуальными заданиями) в мини группах.

IV. СОДЕРЖАНИЕ И СТРУКТУРА ДИСЦИПЛИНЫ

Трудоемкость дисциплины составляет 5 зачетных единиц, 180 часов.

Из них 34 часа лекционных занятий, 52 часа практических занятий, 58 часов на самостоятельную работу в течение семестра и 36 часов на подготовку к экзамену.

Форма отчетности: экзамен (2 семестр)

4.1 Содержание дисциплины, структурированное по темам, с указанием видов учебных занятий и отведенного на них количества академических часов

№ п/п	Раздел дисциплины/темы	Семестр	Виды учебной работы, включая самостоятельную работу обучающихся и трудоемкость (в часах)				Формы текущего контроля успеваемости Форма промежуточной аттестации *(по семестрам)*
			Контактная работа преподавателя с обучающимися			Самостоятельная работа	
			Лекции	Практические занятия	Лабораторные занятия		
1	**Основы машинного обучения**		**10**		**14**	**16**	Контрольная работа, индивидуальные задания
1.1	Обучение с учителем, постановка задачи и методы ее анализа	**2**	2		2	3	
1.2	Язык Python и его модули. Примеры использования библиотеки scikit-learn	2	2		4	3	
1.3	Байесовский оптимальный классификатор	2	2		2	3	
1.4	Выпуклая оптимизация	2	2		2	3	
1.5	Градиентный спуск, стохастический градиентный спуск	2	2		4	4	
2	**Основные модели**	**2**	**14**		**20**	**22**	Контрольная работа, индивидуальные

№ п/п	Раздел дисциплины/темы	Семестр	Виды учебной работы, включая самостоятельную работу обучающихся и трудоемкость (в часах)				Формы текущего контроля успеваемости Форма промежуточной аттестации *(по семестрам)*
			Контактная работа преподавателя с обучающимися			Самостоятельная работа	
							задания
2.1	Метод ближайших соседей	2	2		3	3	
2.2	Линейная регрессия. Гребневая регрессия. Лассо	2	2		3	3	
2.3	Метод опорных векторов	2	2		3	3	
2.4	Решающие деревья. Случайный лес	2	2		3	3	
2.5	Градиентный бустинг	2	2		3	4	
2.6	Нейронные сети. Автоматическое дифференцирование. Обучение нейронных сетей	2	4		5	6	
3	**Другие модели обучения**	2	**10**		**18**	**20**	Контрольная работа, индивидуальные задания
3.1	Наивный байесовский подход. Линейный дискриминантный анализ. Гауссовские смеси	2	2		3	4	
3.2	Понижение размерности. Сингулярное разложение. Метод главных компонент	2	2		4	4	
3.3	Кластеризация	2	2		3	4	
3.4	Введение в онлайн обучение	2	2		3	4	
3.5	Введение в обучение с подкреплением	2	2		5	4	
	Подготовка к экзамену					**36**	
Итого часов			**34**		**52**	**94**	

4.2 План внеаудиторной самостоятельной работы обучающихся по дисциплине

Семестр	Название раздела, темы	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самостоятельной работы
		Вид самостоятельной работы	Сроки выполнения	Затраты времени (час.)		
2	Основы машинного обучения	Изучение лекций, учебной литературы и программного обеспечения	3 недели	9	Контрольная работа	Материалы лекций, рекомендованная учебная литература,
2	Основы машинного обучения	Изучение лекций, учебной литературы и программного обеспечения	2 недели	7	Индивидуальные задания	
2	Основные модели	Изучение лекций, учебной литературы и программного обеспечения	3 недели	12	Контрольная работа	
2	Основные модели	Изучение лекций, учебной литературы	4 недели	14	Индивидуальные задания	
2	Другие модели обучения	Изучение лекций, учебной литературы и программного обеспечения	3 недели	12	Контрольная работа	Материалы лекций, рекомендованная учебная литература
2	Другие модели обучения	Изучение лекций, учебной литературы и программного обеспечения	2 недели	8	Индивидуальные задания	

Семестр	Название раздела, темы	Самостоятельная работа обучающихся			Оценочное средство	Учебно-методическое обеспечение самостоятельной работы
		Вид самостоятельной работы	Сроки выполнения	Затраты времени (час.)		
Общая трудоемкость самостоятельной работы по дисциплине (час)				**58**		
Бюджет времени самостоятельной работы, предусмотренный учебным планом для данной дисциплины (час)				**58**		

4.3 Содержание учебного материала

1. Основы машинного обучения

1.1. Постановки задач машинного обучения. Обучение с учителем. Задачи классификации регрессии. Функции потерь. Эмпирический и истинный риск.

1.2. Недообучение и переобучение. Выбор модели, валидация и кросс-валидация. Регуляризация. Предобработка данных.

1.3. Язык Python и его модули: NumPy, Matplotlib, SciPy, Pandas. Библиотека scikit-learn.

1.4. Условное математическое ожидание. Условное распределение.

1.5. Байесовский оптимальный классификатор, байесовский риск.

1.6. Компромисс между смещением и дисперсией.

1.7. Выпуклая оптимизация. Двойственность.

1.8. Градиентный спуск. Стохастический градиентный спуск.

2. Основные модели

2.1. Метод ближайших соседей. Сходимость и проклятье размерности.

2.2. Линейная регрессия. Вероятностный подход. Связь с методом максимального правдоподобия для гауссовской модели. Байесовский подход. Гребневая регрессия. Лассо.

2.3. Логистическая регрессия. Метод максимального правдоподобия. Кросс-энтропия.

2.4. Перцептрон. Метод опорных векторов для линейно разделимой выборки. Случай неразделимой выборки. Условия оптимальности и опорные векторы. Двойственность.

2.5. Трюк с ядром. Свойства ядер. Гауссовское ядро.

2.6. Решающие деревья. Меры нечистоты для классификации и регрессии. Информационный выигрыш. Последовательное дробление пространства. Настройка параметров.

2.7. Бутстрап и бэггинг. Случайный лес. Out-of-Bag оценки. Важность признаков.

2.8. Градиентный бустинг. Метод наименьших квадратов, AdaBoost, XGBoost, TreeBoost.

2.9. Нейронные сети. Многослойный перцептрон.

2.10. Автоматическое дифференцирование. Прямой и обратный проход.

2.11. Обучение нейронных сетей. Стохастический градиентный спуск и его варианты. Инициализация. Нормализация.

3. Другие модели обучения

3.1. Наивный байесовский подход. Линейный дискриминантный анализ. Скрытые переменные и EM алгоритм. Гауссовские смеси.

3.2. Понижение размерности. Сингулярное разложение. Метод главных компонент.

3.3. Кластеризация. Метод -средних. Гауссовские смеси. Иерархическая кластеризация.

3.4. Введение в онлайн обучение. Алгоритм экспоненциально взвешенного среднего. Онлайн градиентный спуск. Онлайн перцептрон.

3.5. Введение в обучение с подкреплением. Управляемые марковские процессы. Уравнение Беллмана. Итерации по значению. Итерации по стратегиям. -обучение.

V. ОБРАЗОВАТЕЛЬНЫЕ ТЕХНОЛОГИИ

Традиционные лекции, обсуждение конкретных задач прикладного характера, индивидуальные задания, проекты. Дисциплина может быть реализована частично или полностью с использованием ЭИОС Университета (ЭО и ДОТ). Аудиторные занятия и другие формы контактной работы обучающихся с преподавателем могут проводиться с использованием платформ MicrosoftTeams, в том числе, в режиме онлайн-лекций и онлайн-семинаров.

VI. ОЦЕНОЧНЫЕ СРЕДСТВА ДЛЯ ТЕКУЩЕГО КОНТРОЛЯ И ПРОМЕЖУТОЧНОЙ АТТЕСТАЦИИ

Полный комплект контрольно-оценочных материалов (Фонд оценочных средств) оформляется в виде приложения к рабочей программе дисциплины.

VII. УЧЕБНО-МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

7.1. Основная литература.

Пролубников, А. В. Математические методы распознавания образов: учебное пособие: / А. В. Пролубников. – Омск : Омский государственный университет им. Ф.М. Достоевского, 2020. – 110 с. : ил. https://biblioclub.ru/index.php?page=book&id=614061

7.2. Дополнительная литература. Нет

7.3. Список авторских методических разработок. Нет

7.4. Периодические издания (при необходимости)

Machine learning (journal, Springer).

7.5. Перечень ресурсов сети Интернет, необходимых для освоения дисциплины

1. J.Watt and R.Borhani Machine Learning Refined: Notes, Exercises, and Jupyter notebooks
https://github.com/jermwatt/machine_learning_refined
2. A.Geron Machine Learning Notebooks
https://github.com/ageron/handson-ml2
3. A.Geron Deep Learning with TensorFlow 2 and Keras – Notebooks
https://github.com/ageron/tf2_course
4. Курс "Машинное обучение" на ФКН ВШЭ
https://github.com/esokolov/ml-course-hse

7.6. Программное обеспечение информационно-коммуникационных технологий

1. Microsift Windows
2. Microsift Office
3. Python (свободное ПО).
4. MicrosoftTeams

VIII. МАТЕРИАЛЬНО -ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

8.1. Учебно-лабораторное оборудование

При проведении дисциплины учащиеся должны быть обеспечены:
1. Лекционной аудиторией.
2. Аудиторией для лабораторных занятий с аппаратными и программными средствами в соответствии с реализуемой учебной тематикой.

8.2. Программные средства

1. Операционная система Microsoft Windows
2. Офисный пакет Microsoft Office
3. Средства для работы с языком Python (Anaconda, Jupyter Notebook).
4. MicrosoftTeams

IX. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Методические указания приведены в источниках, указанных в разделах 7.1, 7.5.

УЧЕБНАЯ КАРТА ДИСЦИПЛИНЫ

Машинное обучение: математические основы

Трудоемкость: 5 зач.ед.
Форма промежуточной аттестации: экзамен
Курс 1, семестр 2
Код и наименование направления подготовки (специальности): 01.04.02 «Прикладная математика и информатика» (академическая магистратура)
Магистерская программа: «Искусственный интеллект: математические модели и прикладные решения»

№	Виды контрольных мероприятий	Текущий контроль		Рубежный контроль *(при наличии)*
	Модуль 1. Основы машинного обучения	**20**		
1.1	Контрольная работа	10		
1.2	Индивидуальные задания	10		
	Модуль 2. Основные модели	**20**		
2.1	Контрольная работа	10		
2.2	Индивидуальные задания	10		
	Модуль 3. Другие модели обучения	**20**		
3.1	Контрольная работа	10		
3.2	Индивидуальные задания	10		
	Всего	**60**		
	Бонусные баллы	нет		
	Промежуточная аттестация *в форме экзамена*	до 40 баллов	Экзамен проводится в устной форме. Критерии оценки указаны в *Фонде оценочных средств.*	

Преподаватель:

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»
Факультет компьютерных технологий и защиты информации
Кафедра фундаментальной и прикладной математики

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

МАШИННОЕ ОБУЧЕНИЕ: МАТЕМАТИЧЕСКИЕ ОСНОВЫ

Код и наименование направления подготовки/специальности:
01.04.02 «Прикладная математика и информатика»

Уровень образования:
Магистратура

Магистерская программа:
«Искусственный интеллект: математические модели и прикладные решения»

Форма обучения:
Очная

Ростов-на-Дону, 2023

ПЕРЕЧЕНЬ КОМПЕТЕНЦИЙ, ФОРМИРУЕМЫХ ДИСЦИПЛИНОЙ
«Машинное обучение: математические основы»

Код компетенции	Формулировка компетенции
1	2
ПК	**ПРОФЕССИОНАЛЬНЫЕ КОМПЕТЕНЦИИ**
ПК-3	**Способен разрабатывать и применять методы и алгоритмы машинного обучения для решения задач**
ПК-4	**Способен руководить проектами по созданию комплексных систем искусственного интеллекта**

ПАСПОРТ ФОНДА ОЦЕНОЧНЫХ СРЕДСТВ ПО ДИСЦИПЛИНЕ
«Машинное обучение: математические основы»

№ п/п	*Контролируемые разделы дисциплины**	*Код контролируемой компетенции*	Наименование *оценочного средства***
1.	Основы машинного обучения	ПК-3	Контрольная работа, индивидуальные задания
2.	Основные модели	ПК-3, ПК-4	Контрольная работа, индивидуальные задания
3.	Другие модели обучения	ПК-3, ПК-4	Контрольная работа, индивидуальные задания

** Наименование раздела указывается в соответствии с рабочей программой дисциплины.*
***Наименование оценочного средства указывается в соответствии с учебной картой дисциплины.*

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»
Факультет компьютерных технологий и защиты информации
Кафедра фундаментальной и прикладной математики
Контрольные и индивидуальные задания по дисциплине
«Машинное обучение: математические основы»

Модуль 1. Основы машинного обучения

Теоретические задания

1.1. Найти производную по вектору $a \in \mathbb{R}^d$:

$$\frac{\partial}{\partial a}[a^T \exp(aa^T)a],$$

где $\exp(B)$ – матричная экспонента.

1.2. Пусть $X \sim N(\mu, \Sigma)$. Найти $\mathrm{E}\langle B(X-a), X-a\rangle$.

1.3. Пусть A, B – $p \times q$ матрицы и x – случайный $q \times 1$ вектор. Докажите, что

$$\mathrm{Cov}(Ax, Bx) = A\,\mathrm{Cov}(x)B.$$

1.4. Пусть X является гауссовским вектором и

$$\mathrm{E}X = \begin{pmatrix}10\\5\end{pmatrix}, \quad \mathrm{Cov}(X) = \begin{pmatrix}2 & 1\\1 & 1\end{pmatrix}.$$

Выпишите плотность X без использования матричных обозначений.

1.5. Для нормального случайного вектора $z \sim N(\mu, \Sigma)$, который имеет вид $z^T = (x^T, y^T)$, найти условную плотность $p(x|y)$.

1.6. Пусть (X, Y) — двумерный гауссовский вектор с $\mathrm{E}X = \mu_X$, $\mathrm{Var}(X) = \sigma_X^2$, $\mathrm{E}Y = \mu_Y$, $\mathrm{Var}(Y) = \sigma_Y^2$. Пусть коэффициент корреляции между X, Y равен ρ. Найти $\mathrm{E}(Y|X)$.

1.7. Предположим, что метка Y не зависит от признаков X. Докажите, что в таком случае дисперсия Y является нижней оценкой квадратичной ошибки любой модели.

1.8. Пусть $Y = f(X) + \varepsilon$, где признак X равномерно распределен на конечном множестве $\{1, \dots, K\}$, и $\varepsilon \sim N(0, \sigma^2)$. Найти смещение алгоритма

$$\hat{f}(x) = \frac{1}{K}\sum_{i=1}^{K} I_{\{x_i = x\}} y_i.$$

1.9. Рассмотрим задачу классификации точек $\mathcal{X} = \mathbb{R}^2$. Предположим, что истинная метка точки (x_1, x_2) совпадает с $\mathrm{sign}(x)$ (для определенности, $\mathrm{sign}(0) := 1$). Пусть точки распределены равномерно на окружности радиуса 1 с центром в 0. Пусть классификатор определяется прямой

$$x_1 = r\cos\theta, \quad x_2 = r\sin\theta, \quad r \in [0, \infty).$$

Здесь значение $\theta \in (0, \pi/2)$ фиксировано и $h(x) = 1$ в полуплоскости, содержащей $(1,0)$ и $h(x) = -1$ в полуплоскости, содержащей $(-1,0)$.

Найти ожидаемое число ошибок. Какова вероятность того, классификатор сделает хотя бы одну ошибку на выборке из n точек?

1.10. Пусть $L(y, y') = e^{-yy'}$. Найти

$$\operatorname*{argmin}_{b \in \mathbb{R}} \mathrm{E}L(y, b|x),$$

если $y \in \{-1,1\}$.

1.11. Рассмотрим стандартную гауссовскую модель в $\mathbb{R}^d$, где классы 0 и 1 равновероятны и их условные плотности имеют вид

$$N(\mu_0, I), \quad N(\mu_1, I), \quad \mu_0 = -\mu_1 = (a_1, \dots, a_d),$$

где $(a_1, \dots, a_d)$ — вектор параметров, $\| a \| = 1$.

Найти оптимальный байесовский классификатор и его ошибку. Найти байесовскую ошибку для подмножества признаков $X' = (X_{i_1}, \dots, X_{i_{d'}})$ в терминах соответствующих коэффициентов $a_0 = (a_{i_1}, \dots, a_{i_{d'}})$, $d' < d$. Найти оптимальный набор $d < d'$ признаков, если критерием отбора является байесовская ошибка.

1.12. Будем решать задачу линейной регрессии

$$Q(w) = \| Xw - y \|_2^2 \to \min_w$$

методом градиентного спуска:

$$w_t = w_{t-1} - \eta \nabla Q(w_{t-1}).$$

Для заданной итерации найти длину шага, при которой уменьшение функции будет наибольшим:

$$Q(w_{t-1} - \eta \nabla Q(w_{t-1})) \to \min_{\eta > 0}.$$

1.13. Привести пример функции на $\mathbb{R}$ такой, что для любого постоянного шага $\eta > 0$ метод градиентного спуска осциллирует около точки минимума (и не приближается к минимуму на расстояние меньше $\eta/4$) для некоторой начальной точки $x \in \mathbb{R}$.

1.14. Рассмотрим функцию $f(x, y) = x^2 + y^2/4$ и предположим, что метод градиентного спуска с постоянным шагом $\eta = 1/4$ стартует с точки $(1,1)$. Сходится ли он к точке минимума? Экспериментально найдите границу для η, выше которой начинаются осцилляции.

1.15. Пусть Y имеет показательное распределение и $EY = \theta$. Считая, что параметр является линейной функцией признаков $X \in R^d$, и используя метод максимального правдоподобия, предложить модель зависимости Y от X. Предложить практический метод определения коэффициентов модели с использованием метода градиентного спуска.

1.16. Найти симметричную матрицу X, наиболее близкую к матрице A по норме Фробениуса:

$$\sum_{i,j} (x_{ij} - a_{ij})^2 = ||X - A||_2^2 \to \min_X$$
$$X^T = X$$

Задания для программирования

1.17. Сгенерировать 10^4 точек, таким образом, что каждая точка с вероятностью $1/2$ получает метку 0 и порождается распредлением $N(\mu, I)$, $\mu = (a, \dots, a) \in \mathbb{R}^d$, и с той же вероятностью получает метку 0 и порождается распрделением $N(\mu, I)$, $\mu = (-a, \dots, -a) \in \mathbb{R}^d$.

Положить $a = \frac{2}{\sqrt{d}}$ и найти ошибку байесовского классификатора. Для $k = 3, 7, 15$ обучить k-NN классификатор для размерностей $1 \le d \le 500$ and оценить его качество с помощью 10 блоковой кросс-валидации, повторенной 10 раз. Для каждого k построить графики средней доли ошибок как функции d.

1.18. Для выборки (a_i, b_i) размера $n = 100$ из равномерного распределения на $[0,1]^2$ применить метод стохастического градиентного спуска с шагом $\eta = 1/2$ к функции

$$f(x, y) = \frac{1}{n}\sum_{i=1}^n f_i(x, y), \quad f_i(x, y) = (x - a_i)^2 + (y - b_i)^2.$$

Аналитически найти точку минимума (x^*, y^*) и выдать расстояние до нее для $T \in \{10,100,1000\}$ итераций. Сравнить с результатами для шага $\eta_t = 1/t$.

1.19. Для функции $f(x) = (x - 5)^2$ провести эксперименты со скоростью обучения метода градиентного спуска. Для малой скорости обучения сходимость должна быть монотонной. Градиенты и шаги велики, когда точка далека от оптимальной. Они становятся меньше, когда точка приближается к оптимуму. Для большой скорости обучения вы должны увидеть осцилляции и расходимость процесса.

Реализуйте методы Momentum, Nesterov, Adagrad, RMSProp, Adam methods используя Keras:

```
opt=keras.optimizers.SGD(lr=0.001, momentum=0.9)
opt=keras.optimizers.SGD(lr=0.001, momentum=0.9, nesterov=True)
opt=keras.optimizers.Adagrad(learning_rate=0.01)
opt=keras.optimizers.RMSprop(lr=0.01, rho=0.9)
opt=keras.optimizers.Adam(lr=0.001, beta_1=0.9, beta_2=0.999)
```

Сравните графики $(t, f(x_t))$ для (некоторых из) этих функций, где t - номер итерации, x_t - соответствующее значение независимой переменной.

Протестируйте методы градиентного спуска на одной из стандартных тестовых функций https://en.wikipedia.org/wiki/Test_functions_for_optimization

- Beale function
- Goldstein–Price function
- Booth function
- Matyas function
- Lévi function N.13

- Himmelblau's function
- Three-hump camel function
- McCormick function

Модуль 2. Основные модели

Теоретические задания

2.1. Пусть $\widehat{w}_R$, $\widehat{w}_L$ — оптимальные решения следующих задач:

- Ridge-регрессия: $Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell} (y_i - wx_i)^2 + \lambda w^2$;

- LASSO-регрессия: $Q(w) = \frac{1}{\ell}\sum_{i=1}^{\ell} (y_i - wx_i)^2 + \lambda |w|$;

Найти пределы

$$\lim_{\lambda\to 0}\widehat{w}_R, \quad \lim_{\lambda\to\infty}\widehat{w}_R, \quad \lim_{\lambda\to 0}\widehat{w}_L, \quad \lim_{\lambda\to\infty}\widehat{w}_L.$$

2.2. Для заданного набора данных (X_i, Y_i) найдем оптимальные параметры линейной модели (не предполагая, что она верна для реальных данных)

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i.$$

Докажите, что

$$\frac{1}{n}\sum_{i=1}^{n} \hat{Y}_i = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

2.2. Для двух одномерных нормальных распределений $p \sim N(\mu_1, \sigma_1)$, $q \sim N(\mu_2, \sigma_2)$ найдите дивергенцию Кульбака-Лейблера $\mathrm{KL}(p||q)$.

2.3. Пусть распределения p, q таковы, что

$$p(x,y) = p_1(x)p_2(y), \quad q(x,y) = q_1(x)q_2(y).$$

Докажите, что

$$\mathrm{KL}(q|p) = \mathrm{KL}(q_1||p_1) + \mathrm{KL}(q_2||p_2).$$

2.4. Пусть $p \sim N(0, I)$, $q \sim N(\mu, I)$ – d-мерные нормальные распределения. Докажите, что

$$\mathrm{KL}(q|p) = \frac{\|\mu\|^2}{2}.$$

2.5. Пусть даны выборка X, состоящая из 8 объектов, и классификатор $b(x)$, предсказывающий оценку принадлежности объекта положительному классу. Предсказания $b(x)$ и реальные метки объектов приведены ниже:

$$b(x_1) = 0.1, \quad y_1 = 1,$$
$$b(x_2) = 0.8, \quad y_2 = 1,$$
$$b(x_3) = 0.2, \quad y_3 = -1,$$
$$b(x_4) = 0.25, \quad y_4 = -1,$$
$$b(x_5) = 0.9, \quad y_5 = 1,$$
$$b(x_6) = 0.3, \quad y_6 = 1,$$
$$b(x_7) = 0.6, \quad y_7 = -1,$$
$$b(x_8) = 0.95, \quad y_8 = 1.$$

Постройте ROC-кривую и вычислите AUC-ROC для множества классификаторов, порожденных $b(x)$, на выборке X.

2.6. Покажите, что функции

$$K(x,x') = \cos(x - x'), \quad K(x,x') = \prod_{i=1}^{d} (1 + x_i z_i), \quad K(x,x') = \frac{1}{1+e^{-xx'}}$$

являются ядрами.

2.7. Предположим, что для задачи бинарной классификации и есть три алгоритма $b_1(x)$, $b_2(x)$, $b_3(x)$, каждый из которых ошибается с вероятностью p. Мы строим композицию

взвешенным голосованием: алгоритмам присвоены значимости w_1, w_2 и w_3. Взвешенный алгоритм относится к классу 0, если

$$\sum_{i=1}^{n} w_i I_{\{b_i(x)=0\}} \leq \sum_{i=1}^{n} w_i I_{\{b_i(x)=1\}}$$

и к классу 1 в противном случае. Какова вероятность ошибки такой композиции этих трех алгоритмов, если:

- $w_1 = 0.2, w_2 = 0.3, w_3 = 0.2$;
- $w_1 = 0.2, w_2 = 0.5, w_3 = 0.2$?

2.8. Рассмотрим задачу бинарной классификации, $Y = \{0,1\}$. Будем считать, что все алгоритмы из базового семейства $\mathcal{H}$ возвращают значения из отрезка $[0,1]$, которые можно интерпретировать как вероятности принадлежности объекта к классу 1. Выпишите формулы для поиска базового алгоритма h_k и коэффициента α_k в градиентном бустинге при использовании отрицательного логарифма правдоподобия в качестве функции потерь:

$$L(y, z) = -y\ln z - (1 - y)\ln(1 - z).$$

2.9. Пусть $(x_1, y_1), \dots, (x_m, y_m) \in \mathcal{X} \times \{-1,1\}$,

$$\Phi(u) = \begin{cases} (1+u)^2, & u \geq -1 \\ 0, & \text{иначе.} \end{cases}$$

Проверить, что функция

$$F(\alpha) = \sum_{i=1}^{m} \Phi(-y_i f(x_i)), \quad f = \sum_{t=1}^{T} \alpha_t h_t$$

выпукла и дифференцируема. Вывести и алгоритм бустинга для данной функции.

2.10. Рассмотрим полносвязную нейронную сеть, состоящую из (соответственно) входного слоя x_0, скрытого слоя x_1, скрытого слоя x_2, за которым следует выходной узел x_3. Предположим, что функцией активации является “стандартный” сигмоид. Пусть все веса инициализированы 0, и мы применяем метод стохастического градиентного спуска (с использованием обратного распространения) с фиксированной скоростью обучения. Покажите, что на каждом шаге все веса ребер в слое равны (важность случайной инициализации).

Задания для программирования

2.11. Сгенерировать выборку из 10^4 точек из равномерного распределения на гиперкубе $[0,1]^d$ (обозначим полученное множество точек через $\mathcal{X}$). Для $d \in \{1,2,3,5,10,20,50,100,500\}$ и евклидова расстояния ρ найти

- $\min \rho(x, \tilde{x})$: минимум по $x, \tilde{x} \in \mathcal{X}$
- $\overline{\rho(x, \tilde{x})}$: среднее по $x, \tilde{x} \in \mathcal{X}$
- $\max \rho(x, \tilde{x})$: максимум по $x, \tilde{x} \in \mathcal{X}$
- $\overline{d_{NN1}(x)}$: среднее расстояние до ближайшего соседа
- $\max d_{NN1}(x)$: максимальное расстояние до ближайшего соседа

Собрать результаты в таблицу и сопоставить с теоретическими выводами (проклятье размерности).

2.12. $X = R^d, Y = R$,

$$\| w \|_1 := \sum_{i=1}^{d} | w_i |, \quad \| w \|_2 := \sqrt{\sum_{i=1}^{d} | w_i |^2},$$

$$H = \{x \mapsto h(x) = \langle w, x \rangle + b\},$$

+ Обычная линейная регрессия

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m}(\langle w, x_i\rangle + b - y_i)^2 \to \min_{w,b}$$

+ Гребневая регрессия

$$L_S(h) + \lambda \parallel w \parallel_2^2 \to \min_{w,b}$$

+ Лассо

$$L_S(h) + \lambda \parallel w \parallel_1 \to \min_{w,b}$$

+ Эластичная сеть

$$L_S(h) + \lambda_1 \parallel w \parallel_2^2 + \lambda_2 \parallel w \parallel_1 \to \min_{w,b}$$

Пусть выборка определяется многочленом 3 порядка, возмущенным гауссовским шумом:

$$y_i = x_i^3 - 5x_i^2 + 3x_i + 1 + \xi_i, \quad \xi_i \sim N(0,1), \quad x_i \sim U(-1,5)$$

$i \in \{1, \dots, m\}$, $m = 30$.

a) Построить грайик невозмущенной функции и график рассеяния (x_i, y_i).

Обычная линейная регрессия. Рассмотреть расширенный набор признаков: $\mathbf{x} = (x, x^2, \dots, x^d)$ и соответствующую заджачу линейной регрессии:

$$L_S(h) = \frac{1}{m}\sum_{i=1}^{m}(\langle \mathbf{w}, \mathbf{x}_i\rangle + b - y_i)^2 \to \min_{\mathbf{w},b}$$

Класс гипотез

$$H_d = \{x \mapsto h(x) = \langle w, x\rangle + b\}$$

зависит от параметра d. Ошибки обучения:

$$\psi(d) = \inf_{\mathbf{w},b} L_S(h)$$

Средние ошибки кросс-валидации:

$$\overline{\psi}(d) = \frac{1}{k}\sum_{i=1}^{k}\psi_i(d), \quad \psi_i(d) = \inf_{\mathbf{w},b} L_{S\setminus S_i}(h).$$

b) Найти d с наименьшей ошибкой кросс-валидации.

c) Вычислить среднеквадратичную ошибку обучения. Сравнить ее со стандартным отклонением данных. Сравнить коэффициенты лучшей модели и исходные коэффициенты.

d) Построить графики полученных моделей для $d = 3$ и $d = 10$ и невозмущенную исходную функцию.

Гребневая регрессия.

e) Рассмотреть класс многочленов фиксированной большой степени и найти лучшее (в некотором диапазоне) значение параметра регуляризации α. Вычислить соответствующую среднеквадратичную ошибку обучения. Построить график зависимости ошибок обучения и кросс-валидации от параметра регуляризации.

f) Сравнить график лучшей модели и график исходной невозмущенной функции.

Лассо.

g) Повторить e) и f).

Эластичная сеть:

$$\frac{1}{2}L_S(h) + \alpha r|w|_1 + \frac{1}{2}\alpha(1-r)|w|_2^2.$$

h) Для фиксированного большого d найти лучшую комбинацию параметров α и r используя GridSearchCV.

i) Сравнить график лучшей модели и график исходной невозмущенной функции.

j) Сравнить коэффициенты лучших моделей Lasso и ElasticNet.

2.13. Загрузите набор данных «credit-g» с помощью ‘fetch_openml (‘ credit_g ’)’ (https://www.openml.org/d/31)

a) Определите, какие признаки являются непрерывными, а какие – категориальными.

b) Визуализируйте одномерное распределение каждого непрерывного признака и распределение целевого признака.

c) Разделите данные на обучающий и тестовый набор. Проведите предобработку данных без использования pipeline и проведите предварительную оценку LogisticRegression.

d) Используйте OneHotEncoder и pipeline для кодирования категориальных переменных. Оцените модели логистической регрессии, линейного метода опорных векторов и метода ближайших соседей с помощью кросс-валидации. Как влияет на результаты масштабирование непрерывных признаков с помощью StandardScaler?

e) Настройте параметры с помощью GridSearchCV. Улучшаются ли результаты? Оцените лучшую модель на тестовом наборе. Визуализируйте оценку качества как функцию параметров для всех трех моделей.

f) Измените стратегию перекрестной проверки с «стратифицированной k-кратной» на «k-кратную» с перемешиванием (shuffling). Меняются ли оптимальные параметры? Меняются ли они, если изменить базу генерации при перемешивании или при разбиении данных на обучающую и тестовую выборки?

g) Визуализируйте 20 наиболее важных коэффициентов для логистической регрессии и метода опорных векторов.

Модуль 3. Другие модели обучения

Теоретические задания

3.1. Рассмотрите смесь двух одномерных гауссовских распределений

$$p(x) = \pi_1 N(x|\mu_1, \sigma_1^2) + \pi_2 N(x|\mu_2, \sigma_2^2).$$

Пусть дана выборка $(x_1, \dots, x_n)$. Считая дисперсии σ_1^2, σ_2^2 известными, введите скрытые переменные и выведите формулы EM-алгоритма для настройки параметров π_1, π_2, μ_1, μ_2.

3.2. Наблюдается выборка бинарных значений $y = (y_1, \dots, y_n)$, $y_i \in \{0,1\}$. Все элементы выборки генерируются независимо, но известно, что в некоторый момент z меняется параметр распределения Бернулли:

$$\mathrm{P}(y_i = 1) = \begin{cases} P(y_i = 1) = \theta_1, & i < z, \\ P(y_i = 1) = \theta_2, & i \geq z \end{cases}$$

Вывести формулы для EM-алгоритма, где z – скрытая переменная, а θ_1, θ_2 – параметры распределений Бернулли.

3.3. Найти сингулярное разложение матрицы

$$A = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

3.4. Для набора данных, состоящего из всех точек декартова произведения $\{1,2,3,4\} \times \{1,2,3\}$ найти все возможные решения алгоритма кластеризации K-means (используя все возможные начальные точки) для $K = 2,3,4$.

3.5. Рассмотреть управляемый марковский с целевой функцией, соответствующей максимизации общего ожидаемого дохода (коэффициент дисконтирования равен 1). Пусть в конце каждого временного шага существует вероятность $\alpha > 0$ остановки процесса. Показать, что это равносильно использованию дисконтированного критерия на бесконечном горизонте.

3.6. Составить уравнение Беллмана для следующей задачи. Игрок имеет i фунтов и хочет увеличить свой капитал до N. Каждый раз он может поставить любое целое количество фунтов $j \leq i$. С вероятностью p он выигрывает и будет иметь $i + j$ рублей, в противном случае у него останется $i - j$ рублей. Игра заканчивается, когда его капитал достигнет N или 0. Максимизировать вероятность достижения N.

Задания для программирования

3.5. Пусть квадратная $m \times m$ матрица A имеет m независимых собственных векторов:

$$Av_i = \lambda_i v_i, \quad i = 1, \dots, m.$$

Тогда

$$AQ = Q\Lambda, \quad Q = (v_1, \dots, v_m), \quad \Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_m).$$

И может факторизвана следующим образом:

$$A = Q\Lambda Q^{-1}.$$

a) Создать случайную $n \times n$ (напр., $n = 4$) симметриччную матрицу A. Найти Q, Λ, используя ‘numpy.linalg.eig‘. Проверить, что $A = Q\Lambda Q^{-1}$.

b) Создать any $m \times n$ матрицу (напр., $m = 4, n = 3$). Используя ‘scipy.linalg.svd‘ найти U, Σ, V. Проверить, что $A = U\Sigma V^T$.

c) Создать случайную $m \times n$ матрицу X столбцы которой имеют нулевое среднее. Используя ‘sklearn.decomposition.PCA‘, найти C, V, X_s и долю объясненной дисперсии (напр., $m = 4, n = 3, s = 2$).

d) Применить метод главных компонент к набору данных ‘cancer‘ с $s = 2$. На плоскости (первая главная компонента, вторая главная компонента) показать спроектированные точки. Является ли этот набор данных приблизиженно линейным разделимым указанных координатах?

e) Примените ту же процедуру к набору данных ‘wine’.

f) Изучите пример Eigenfaces Vanderplas J. Python data science handbook (O’Reilly, 2016) https://jakevdp.github.io/PythonDataScienceHandbook/05.09-principal-component-analysis.html

g) Примените ту же процедуру к набору данных ‘MNIST’ dataset.

3.6. Рассмотреть задачу Frozen Lake (замерзшее озеро).

https://reinforcement-learning4.fun/2019/06/16/gym-tutorial-frozen-lake/
https://www.kaggle.com/sarjit07/reinforcement-learning-using-q-table-frozenlake

a) Решить ее с помощью метода динамического программирования, используя алгоритмы итераций по значению и по стратегиям.
b) Применить алгоритм Q-обучения.

Критерии оценивания

Максимальная оценка за задания каждого модуля составляет $M = 20$ баллов (отбор заданий осуществляется преподавателем). Студенту выставляется

M баллов, если работа выполнена полностью и оформлена в соответствии с предъявленными требованиями.

$0.75 * M$ баллов, если работа выполнена, но допущены неточности или ее оформление не соответствует предъявленным требованиям,

$0.5 * M$ баллов, если имеются существенные ошибки, но общая схема выполнения работы правильна,

$0.25 * M$ баллов, если выполнение работы было начато, но она выполнена лишь частично, существенные шаги не сделаны,

0 баллов, если работа не выполнена.

После суммирования всех набранных баллов, округление производится до ближайшего целого числа. Для допуска к экзамену необходимо набрать 38 баллов.

Вопросы к экзамену
по дисциплине **«Машинное обучение: математические основы»**

1. Постановки задач машинного обучения. Обучение с учителем. Задачи классификации регрессии. Функции потерь. Эмпирический и истинный риск.

2. Недообучение и переобучение. Выбор модели, валидация и кросс-валидация. Регуляризация.

3. Условное математическое ожидание. Условное распределение.

4. Байесовский оптимальный классификатор, байесовский риск.

5. Компромисс между смещением и дисперсией.

6. Выпуклая оптимизация. Двойственность.

7. Градиентный спуск. Стохастический градиентный спуск.

8. Метод ближайших соседей. Проклятье размерности.

9. Линейная регрессия. Вероятностный подход. Связь с методом максимального правдоподобия для гауссовской модели. Байесовский подход. Гребневая регрессия. Лассо.

10. Логистическая регрессия. Метод максимального правдоподобия. Кросс-энтропия.

11. Перцептрон. Метод опорных векторов для линейно разделимой выборки.

12. Метод опорных векторов для неразделимой выборки. Условия оптимальности и опорные векторы. Двойственность.

13. Трюк с ядром. Свойства ядер. Гауссовское ядро.

14. Решающие деревья. Меры нечистоты для классификации и регрессии. Информационный выигрыш. Последовательное дробление пространства.

15. Бутстрап и бэггинг. Случайный лес. Out-of-Bag оценки. Важность признаков.

16. Градиентный бустинг. Метод наименьших квадратов,

17. Градиентный бустинг. AdaBoost.

18. Нейронные сети. Многослойный перцептрон.

19. Автоматическое дифференцирование. Прямой и обратный проход.

20. Наивный байесовский подход. Линейный дискриминантный анализ.

21. Скрытые переменные и EM алгоритм. Гауссовские смеси.

22. Сингулярное разложение.

23. Понижение размерности. Метод главных компонент.

24. Кластеризация. Метод -средних.

25. Гауссовские смеси. Иерархическая кластеризация.

26. Онлайн обучение. Алгоритм экспоненциально взвешенного среднего.

27. Онлайн градиентный спуск. Онлайн перцептрон.

28. Управляемые марковские процессы. Уравнение Беллмана. Итерации по значению. Итерации по стратегиям.

29. Обучение с подкреплением. -обучение.

Форма экзаменационного билета

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования «Ростовский государственный экономический университет (РИНХ)»
Факультет компьютерных технологий и защиты информации
Кафедра фундаментальной и прикладной математики

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ №
По дисциплине «Машинное обучение: математические основы»
Направление подготовки 01.04.02 «Прикладная математика и информатика»

1 . Байесовский оптимальный классификатор, байесовский риск.
2. Кластеризация. Метод -средних.

Составитель

Заведующий кафедрой

Критерии оценивания

40 баллов – ответ полный и правильный; студент хорошо понимает дополнительные вопросы,

30 баллов – в ответе допущены две-три ошибки, исправленные после наводящих вопросов преподавателя,

22 балла – студент на идейном уровне понимает содержание материала понимает содержание материала, но затрудняется воспроизвести существенные технические детали,

10 баллов – студент понимает содержание поставленного в билете вопроса, но слабо ориентируется в содержании основного учебного материала, не может исправить сделанные ошибки при наводящих вопросах преподавателя,

0 баллов – ответ отсутствует.

Для успешной сдачи экзамена необходимо набрать 22 балла.