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

УТВЕРЖДАЮ
Начальник отдела лицензирования и аккредитации
_____Чаленко К.Н.
«30» августа 2021 г.

Рабочая программа дисциплины
Машинное обучение

по профессионально-образовательной программе направление 01.03.05 "Статистика"
профиль 01.03.05.01 "Анализ больших данных"

Для набора 2021 года

Квалификация
Бакалавр

КАФЕДРА **Статистики, эконометрики и оценки рисков**

Распределение часов дисциплины по семестрам

Семестр (<Курс>.<Семестр на курсе>)	8 (4.2)		Итого	
Недель	14			
Вид занятий	УП	РП	УП	РП
Лекции	28	28	28	28
Лабораторные	28	28	28	28
Итого ауд.	56	56	56	56
Контактная работа	56	56	56	56
Сам. работа	88	88	88	88
Часы на контроль	36	36	36	36
Итого	180	180	180	180

ОСНОВАНИЕ

Учебный план утвержден учёным советом вуза от 30.08.2021 протокол № 1.

Программу составил(и): к.э.н., доцент, Кокина Е.П. _____

Зав. кафедрой: д.э.н., проф. Ниворожкина Л.И. _____

Методическим советом направления: к.э.н., доцент, Кислая И.А. _____

1. ЦЕЛИ ОСВОЕНИЯ ДИСЦИПЛИНЫ

1.1	Ознакомление студентов с теоретическими основами и основными принципами машинного обучения — а именно, с классами моделей (линейные, логические, нейросетевые), метриками, качествами и подходами к подготовке данных; формирование практических навыков работы с данными и решения прикладных задач анализа данных.

2. ТРЕБОВАНИЯ К РЕЗУЛЬТАТАМ ОСВОЕНИЯ ДИСЦИПЛИНЫ

ПК-6:Способен осуществлять поиск статистической информации, ее первичную обработку и подготовку для проведения аналитических исследований, в том числе с использованием технологий больших данных

В результате освоения дисциплины обучающийся должен:
Знать:
современные задачи, методы и модели анализа данных
Уметь:
применять алгоритмы машинного обучения для поиска статистической информации и первичной обработки данных
Владеть:
методиками и инструментарием для решения практических задач с использованием методов машинного обучения

3. СТРУКТУРА И СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Код занятия	Наименование разделов и тем /вид занятия/	Семестр / Курс	Часов	Компетен-ции	Литература
	Раздел 1. Введение в машинное обучение				
1.1	Тема "Введение в машинное обучение". Постановки основных классов задач в машинном обучении. Обучение с учителем (supervised learning): регрессия и классификация; обучение без учителя (unsupervised learning): кластеризация, снижение размерности; semi-supervised learning, рекомендательные системы, обработка текстов: тематическое моделирование, построение аннотаций, извлечение ответов на вопросы, машинный перевод; обработка изображений: порождение, преобразование; обучение представлений; обучение с подкреплением. Примеры задач. Виды данных: структурированные таблицы, тексты, изображения, звук, логи. Признаки. /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
1.2	Тема "Статистические оценки и проверка гипотез". Основные понятия математической статистики: статистические оценки (точечные и интервальные), их свойства, проверка гипотез. /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
1.3	Тема "Введение в машинное обучение". Постановки основных классов задач в машинном обучении. Обучение с учителем (supervised learning): регрессия и классификация; обучение без учителя (unsupervised learning): кластеризация, снижение размерности; semi-supervised learning, рекомендательные системы, обработка текстов: тематическое моделирование, построение аннотаций, извлечение ответов на вопросы, машинный перевод; обработка изображений: порождение, преобразование; обучение представлений; обучение с подкреплением. Примеры задач. Виды данных: структурированные таблицы, тексты, изображения, звук, логи. Признаки.(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
1.4	Тема "Статистические оценки и проверка гипотез". Основные понятия математической статистики: статистические оценки (точечные и интервальные), их свойства, проверка гипотез.(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3

1.5	Тема "Введение в машинное обучение". Постановки основных классов задач в машинном обучении. Обучение с учителем (supervised learning): регрессия и классификация; обучение без учителя (unsupervised learning): кластеризация, снижение размерности; semi-supervised learning, рекомендательные системы, обработка текстов: тематическое моделирование, построение аннотаций, извлечение ответов на вопросы, машинный перевод; обработка изображений: порождение, преобразование; обучение представлений; обучение с подкреплением. Примеры задач. Виды данных: структурированные таблицы, тексты, изображения, звук, логи. Признаки. /Ср/	8	4	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
1.6	Тема "Статистические оценки и проверка гипотез". Основные понятия математической статистики: статистические оценки (точечные и интервальные), их свойства, проверка гипотез. /Ср/	8	4	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
	Раздел 2. Линейные модели				
2.1	Тема "Машинное обучение как математическое моделирование". Статистические модели. Теоретико-вероятностная постановка задачи обучения с учителем. Минимизация ожидаемой ошибки. No free lunch theorem. Пример: задача регрессии, минимизация квадрата отклонения. Регрессионная функция: условное матожидание. Линейная регрессия и метод k ближайших соседей. Переобучение и недообучение. Разложение ошибки на шум, смещение и разброс. Проклятие размерности. Методы оценивания обобщающей способности, кросс-валидация. /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
2.2	Тема «Введение в линейные модели и задача регрессии». Градиентный спуск, методы оценивания градиента. Функции потерь. Метрики качества регрессии. Линейная регрессия, метод наименьших квадратов и максимизация правдоподобия. Теорема Гаусса—Маркова. Явный вид решения в методе наименьших квадратов. Ковариационная матрица для коэффициентов. Практические соображения: что делать с категориальными данными? Вычислительные соображения: точное решение vs градиентный спуск. Регуляризация. /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
2.3	Тема "Машинное обучение как математическое моделирование". Статистические модели. Теоретико-вероятностная постановка задачи обучения с учителем. Минимизация ожидаемой ошибки. No free lunch theorem. Пример: задача регрессии, минимизация квадрата отклонения. Регрессионная функция: условное матожидание. Линейная регрессия и метод k ближайших соседей. Переобучение и недообучение. Разложение ошибки на шум, смещение и разброс. Проклятие размерности. Методы оценивания обобщающей способности, кросс-валидация.(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
2.4	Тема «Введение в линейные модели и задача регрессии». Градиентный спуск, методы оценивания градиента. Функции потерь. Метрики качества регрессии. Линейная регрессия, метод наименьших квадратов и максимизация правдоподобия. Теорема Гаусса—Маркова. Явный вид решения в методе наименьших квадратов. Ковариационная матрица для коэффициентов. Практические соображения: что делать с категориальными данными? Вычислительные соображения: точное решение vs градиентный спуск. Регуляризация.(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3

2.5	Тема "Машинное обучение как математическое моделирование". Статистические модели. Теоретико-вероятностная постановка задачи обучения с учителем. Минимизация ожидаемой ошибки. No free lunch theorem. Пример: задача регрессии, минимизация квадрата отклонения. Регрессионная функция: условное матожидание. Линейная регрессия и метод k ближайших соседей. Переобучение и недообучение. Разложение ошибки на шум, смещение и разброс. Проклятие размерности. Методы оценивания обобщающей способности, кросс-валидация. /Ср/	8	6	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
2.6	Тема «Введение в линейные модели и задача регрессии». Градиентный спуск, методы оценивания градиента. Функции потерь. Метрики качества регрессии. Линейная регрессия, метод наименьших квадратов и максимизация правдоподобия. Теорема Гаусса—Маркова. Явный вид решения в методе наименьших квадратов. Ковариационная матрица для коэффициентов. Практические соображения: что делать с категориальными данными? Вычислительные соображения: точное решение vs градиентный спуск. Регуляризация. /Ср/	8	6	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
	Раздел 3. Признаковые представления				
3.1	Тема "Линейные модели и задача классификации". Задачи классификации. Общая постановка. 0-1 ошибка. Байесовский классификатор. Линейные методы для классификации. Логистическая регрессия, максимизация правдоподобия, кросс-энтропия. /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
3.2	Тема "Выбор и оценка моделей, работа с признаками". Кросс-валидация: тонкости (отбор переменных, переобучение на валидационное множество). Оценки ожидаемой ошибки для линейной регрессии: AIC и другие. L1 и L2 регуляризация. Методы отбора признаков. Метод главных компонент и singular spectrum analysis. Ядровые методы. Ядра и спрямляющие пространства, методы их построения. Операции в спрямляющих пространствах. /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
3.3	Тема "Линейные модели и задача классификации". Задачи классификации. Общая постановка. 0-1 ошибка. Байесовский классификатор. Линейные методы для классификации. Логистическая регрессия, максимизация правдоподобия, кросс-энтропия.(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
3.4	Тема "Выбор и оценка моделей, работа с признаками". Кросс-валидация: тонкости (отбор переменных, переобучение на валидационное множество). Оценки ожидаемой ошибки для линейной регрессии: AIC и другие. L1 и L2 регуляризация. Методы отбора признаков. Метод главных компонент и singular spectrum analysis. Ядровые методы. Ядра и спрямляющие пространства, методы их построения. Операции в спрямляющих пространствах.(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
3.5	Тема "Линейные модели и задача классификации". Задачи классификации. Общая постановка. 0-1 ошибка. Байесовский классификатор. Линейные методы для классификации. Логистическая регрессия, максимизация правдоподобия, кросс-энтропия. /Ср/	8	6	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
3.6	Тема "Выбор и оценка моделей, работа с признаками". Кросс-валидация: тонкости (отбор переменных, переобучение на валидационное множество). Оценки ожидаемой ошибки для линейной регрессии: AIC и другие. L1 и L2 регуляризация. Методы отбора признаков. Метод главных компонент и singular spectrum analysis. Ядровые методы. Ядра и спрямляющие пространства, методы их построения. Операции в спрямляющих пространствах. /Ср/	8	6	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
	Раздел 4. Решающие деревья и композиции				

4.1	Тема "Деревья и ансамбли". Ограничения линейных методов (пример: XOR). Решающие деревья. CART. Ансамбли. Бутстреп. Бэггинг. Случайный лес. /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
4.2	Тема "Бустинг" AdaBoost, градиентный бустинг. XGBoost. /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
4.3	Тема "Деревья и ансамбли". Ограничения линейных методов (пример: XOR). Решающие деревья. CART. Ансамбли. Бутстреп. Бэггинг. Случайный лес.(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
4.4	Тема "Бустинг" AdaBoost, градиентный бустинг. XGBoost.(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
4.5	Тема "Деревья и ансамбли". Ограничения линейных методов (пример: XOR). Решающие деревья. CART. Ансамбли. Бутстреп. Бэггинг. Случайный лес. /Ср/	8	6	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
4.6	Тема "Бустинг" AdaBoost, градиентный бустинг. XGBoost. /Ср/	8	6	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
	Раздел 5. Нейронные сети				
5.1	Тема "Признаковые представления для дискретных входных данных". Практические соображения. Кодирование категориальных данных. Пропущенные значения. Обработка текстов: bag of words, tf-idf, векторные эмбеддинги. /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
5.2	Тема "Введение в нейросети". Нейронные сети: общая архитектура. Реализация XOR с помощью трёх персептронов. Теорема об универсальной аппроксимации. Многослойные сети. Обратное распространение ошибки. Стохастический градиентный спуск. Проблемы: затухающие и взрывающиеся градиенты, невыпуклость функции потерь. /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
5.3	Тема "Современные нейросетевые архитектуры" Нейронные сети в обработке изображений. Фильтры. Сверточные слои. Нейронные сети и обучение представлений. Обработка последовательностей. Рекуррентные нейронные сети. /Лек/	8	4	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
5.4	Тема "Признаковые представления для дискретных входных данных". Практические соображения. Кодирование категориальных данных. Пропущенные значения. Обработка текстов: bag of words, tf-idf, векторные эмбеддинги.(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
5.5	Тема "Введение в нейросети". Нейронные сети: общая архитектура. Реализация XOR с помощью трёх персептронов. Теорема об универсальной аппроксимации. Многослойные сети. Обратное распространение ошибки. Стохастический градиентный спуск. Проблемы: затухающие и взрывающиеся градиенты, невыпуклость функции потерь.(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
5.6	Тема "Современные нейросетевые архитектуры" Нейронные сети в обработке изображений. Фильтры. Сверточные слои. Нейронные сети и обучение представлений. Обработка последовательностей. Рекуррентные нейронные сети. (С использованием языка программирования Python и ПО MS Office) /Лаб/	8	4	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3

5.7	Тема "Признаковые представления для дискретных входных данных". Практические соображения. Кодирование категориальных данных. Пропущенные значения. Обработка текстов: bag of words, tf-idf, векторные эмбеддинги. /Ср/	8	10	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
5.8	Тема "Введение в нейросети". Нейронные сети: общая архитектура. Реализация XOR с помощью трёх персептронов. Теорема об универсальной аппроксимации. Многослойные сети. Обратное распространение ошибки. Стохастический градиентный спуск. Проблемы: затухающие и взрывающиеся градиенты, невыпуклость функции потерь. /Ср/	8	10	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
5.9	Тема "Современные нейросетевые архитектуры" Нейронные сети в обработке изображений. Фильтры. Сверточные слои. Нейронные сети и обучение представлений. Обработка последовательностей. Рекуррентные нейронные сети. /Ср/	8	10	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
	Раздел 6. Кластеризация и методы снижения размерности				
6.1	Тема "Кластеризация". K-means. EM-алгоритм. Другие методы кластеризации: иерархическая кластеризация, DBSCAN, Affinity Propagation. /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
6.2	Тема "Снижение размерности". SVD-разложение. Метод главных компонент. t-SNE, UMAP /Лек/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
6.3	Тема "Кластеризация". K-means. EM-алгоритм. Другие методы кластеризации: иерархическая кластеризация, DBSCAN, Affinity Propagation.(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
6.4	Тема "Снижение размерности". SVD-разложение. Метод главных компонент. t-SNE, UMAP(С использованием языка программирования Python и ПО MS Office) /Лаб/	8	2	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
6.5	Тема "Кластеризация". K-means. EM-алгоритм. Другие методы кластеризации: иерархическая кластеризация, DBSCAN, Affinity Propagation. /Ср/	8	8	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
6.6	Тема "Снижение размерности". SVD-разложение. Метод главных компонент. t-SNE, UMAP /Ср/	8	6	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3
6.7	/Экзамен/	8	36	ПК-6	Л1.1 Л1.2 Л1.3 Л1.4Л2.1 Л2.2 Л2.3

4. ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

Структура и содержание фонда оценочных средств для проведения текущей и промежуточной аттестации представлены в Приложении 1 к рабочей программе дисциплины.

5. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

5.1. Основная литература

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л1.1	Местецкий Л. М.	Математические методы распознавания образов: курс лекций: курс лекций	Москва: Интернет-Университет Информационных Технологий (ИНТУИТ), 2008	https://biblioclub.ru/index.php?page=book&id=234163 неограниченный доступ для зарегистрированных пользователей

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л1.2	Айвазян, С. А., Мхитарян, В. С., Зехин, В. А.	Практикум по многомерным статистическим методам: учебное пособие	Москва: Евразийский открытый институт, Московский государственный университет экономики, статистики и информатики, 2003	http://www.iprbookshop.ru/10803.html неограниченный доступ для зарегистрированных пользователей
Л1.3	Алексеева Т. В., Амириди Ю. В., Дик В. В., Лужецкий М. Г.	Информационные аналитические системы: Учебник	Москва: Московский финансово-промышленный университет «Синергия», 2013	http://www.iprbookshop.ru/17015.html неограниченный доступ для зарегистрированных пользователей
Л1.4	Маккинли, Уэс, Слинкина, А.	Python и анализ данных	Саратов: Профобразование, 2019	http://www.iprbookshop.ru/88752.html неограниченный доступ для зарегистрированных пользователей

5.2. Дополнительная литература

	Авторы, составители	Заглавие	Издательство, год	Колич-во
Л2.1		Журнал "Вопросы статистики"	,	1
Л2.2		Введение в нейронные сети	Москва: Интернет-Университет Информационных Технологий (ИНТУИТ), 2016	http://www.iprbookshop.ru/52144.html неограниченный доступ для зарегистрированных пользователей
Л2.3	Шелудько В. М.	Язык программирования высокого уровня Python: функции, структуры данных, дополнительные модули: учебное пособие	Ростов-на-Дону, Таганрог: Южный федеральный университет, 2017	https://biblioclub.ru/index.php?page=book&id=500060 неограниченный доступ для зарегистрированных пользователей

5.3 Профессиональные базы данных и информационные справочные системы

1. База данных Центрального банка РФ http://cbr.ru/hd_base/
2. Базы данных Росстата https://gks.ru/databases
3. Центральная база статистических данных https://www.gks.ru/dbscripts/cbsd/dbinet.cgi
4. Базы данных Ростовстата https://rostov.gks.ru/folder/56777, https://rostov.gks.ru/folder/29957
5. Единая межведомственная информационно-статистическая система https://www.fedstat.ru/
6. База данных Российского мониторинга экономического положения и здоровья населения НИУ ВШЭ https://www.hse.ru/rlms
7. Базы данных ВЦИОМ https://wciom.ru/?id=79, https://wciom.ru/?id=1130
8. Консультант+

5.4. Перечень программного обеспечения

MS Office

5.5. Учебно-методические материалы для студентов с ограниченными возможностями здоровья

При необходимости по заявлению обучающегося с ограниченными возможностями здоровья учебно-методические материалы предоставляются в формах, адаптированных к ограничениям здоровья и восприятия информации. Для лиц с нарушениями зрения: в форме аудиофайла; в печатной форме увеличенным шрифтом. Для лиц с нарушениями слуха: в форме электронного документа; в печатной форме. Для лиц с нарушениями опорно-двигательного аппарата: в форме электронного документа; в печатной форме.

6. МАТЕРИАЛЬНО-ТЕХНИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ (МОДУЛЯ)

Помещения для проведения всех видов работ, предусмотренных учебным планом, укомплектованы необходимой специализированной учебной мебелью и техническими средствами обучения. Для проведения лекционных занятий используется демонстрационное оборудование. Лабораторные занятия проводятся в компьютерных классах, рабочие места в которых оборудованы необходимыми лицензионными программными средствами и выходом в Интернет.

7. МЕТОДИЧЕСКИЕ УКАЗАНИЯ ДЛЯ ОБУЧАЮЩИХСЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ (МОДУЛЯ)

Методические указания по освоению дисциплины представлены в Приложении 2 к рабочей программе дисциплины.

Приложение 1

ФОНД ОЦЕНОЧНЫХ СРЕДСТВ

1 . Описание показателей и критериев оценивания компетенций на различных этапах их формирования, описание шкал оценивания

1.1 Критерии оценивания компетенций

ЗУН, составляющие компетенцию	Показатели оценивания	Критерии оценивания	Средства оценивания
ПК-6 Способность осуществлять поиск статистической информации, ее первичную обработку и подготовку для проведения аналитических исследований, в том числе с использованием технологий больших данных			
Знать современные задачи, методы и модели анализа данных	Формулирует основные задачи, раскрывает сущность методов и моделей анализа данных	Полнота, содержательность и грамотность ответа.	Коллоквиум (вопросы 1-81) Экзаменационные вопросы и задания (1-23)
Уметь применять алгоритмы машинного обучения для поиска статистической информации и первичной обработки данных	Использует современное программное обеспечение для решения задач.	Правильность выбора и применения методов машинного обучения, а также алгоритмов языка программирования Python	Задания к лабораторным работам (1-10) Экзаменационные вопросы и задания (24-33)
Владеть методиками и инструментарием для решения практических задач с использованием методов машинного обучения	Использует прикладные методы машинного обучения. Применяет методы машинного обучения для анализа данных. Оценивает и формулирует выводы по результатам применения методов машинного обучения	Владение методами решения задач, верная интерпретация полученных результатов	Задания к лабораторным работам (1-10) Экзаменационные вопросы и задания (24-33)

1.2. Шкала оценивания

Текущий контроль успеваемости и промежуточная аттестация осуществляется в рамках накопительной балльно-рейтинговой системы в 100-балльной шкале:

84-100 баллов (оценка «отлично»)
67-83 баллов (оценка «хорошо»)
50-66 баллов (оценка «удовлетворительно»)
0-49 баллов (оценка «неудовлетворительно»)

2. Типовые контрольные задания или иные материалы, необходимые для оценки знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций в процессе освоения образовательной программы

Экзаменационные вопросы и задания

1. Условия и предпосылки формирования машинного обучения как самостоятельной научной дисциплины.
2. Основные термины и понятия машинного обучения.
3. Место машинного обучения среди других дисциплин (анализ данных, математическая статистика, базы данных, базы знаний и др.).
4. Классы задач, решаемые методами машинного обучения.
5. Задачи классификации данных - постановки задачи, примеры практических приложений.
6. Линейные методы классификации и регрессии: функционалы качества, методы настройки, особенности применения.
7. Метод опорных векторов как линейный классификатор - концептуальная схема и численные алгоритмы.
8. Реализация Метода опорных векторов в задаче классификации в случае линейной разделимости классов.
9. Реализация Метода опорных векторов в задаче классификации в случае линейной неразделимости классов.
10. Метод наименьших квадратов и его применение для восстановления регрессии.
11. Параметрические и непараметрические методы восстановления регрессии
12. Метрики качества алгоритм регрессии и классификации.
13. Оценивание качества алгоритмов. Отложенная выборка, ее недостатки. Оценка полного скользящего контроля. Кросс-валидация. Leave-one-out.
14. Деревья решений. Методы построения деревьев. Их регуляризация.
15. Композиции алгоритмов. Разложение ошибки на смещение и разброс.
16. Случайный лес, его особенности.
17. Градиентный бустинг, его особенности при использовании деревьев в качестве базовых алгоритмов.
18. Нейронные сети. Метод обратного распространения ошибок. Свёрточные сети.
19. Основные идеи использования искусственных нейронных сетей в задачах машинного обучения.
20. Реализация искусственных нейронных сетей для решения задач классификации.
21. Реализация искусственных нейронных сетей для решения задач распознавания образов.
22. Понятие обучения сети. Примеры.
23. Кластеризация. Алгоритм K-Means.
24. Пусть переменные x_1 и x_2 измерены на четырех объектах A, B, C и D:

Объекты	A	B	C	D
x_1	5	-1	1	-3

x_2	3	1	-2	-2

Необходимо классифицировать объекты на две группы методом k-средних.

25. Проанализируйте связь между полом работника и характером труда в сезонных отраслях:

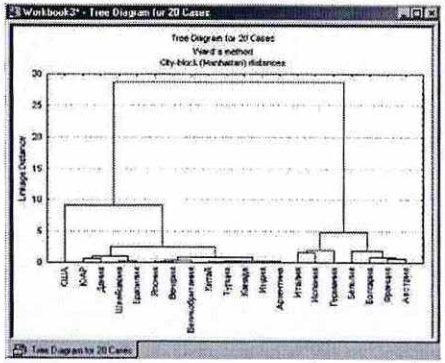
Пол	Численность занятых в отраслях		
	Сезонных	Не сезонных	Всего
Мужчины	187	265	452
Женщины	307	272	579
Всего	494	537	1031

26. Оцените значимость различий двух рынков сбыта бытовой техники. На первом рынке (число наблюдений – 5) средний уровень цены реализации составил 5 тыс. руб., а экспертная оценка качества обслуживания – 3,4 балла; на втором рынке (число наблюдений – 7) соответственно 7 тыс. руб. и 4,3 балла. Объединенная ковариационная матрица имеет вид: $\mathbf{S}=\begin{bmatrix} 9{,}3 & 0{,}26 \\ 0{,}26 & 2{,}0 \end{bmatrix}$.

27. Два эксперта проранжировали 10 предложенных им проектов по степени эффективности: X_1=(1; 2; 4; 6; 7; 3; 5; 8; 9; 10), X_2=(2; 3; 1; 4; 6; 5; 9; 7; 8; 10). Оцените степень согласованности мнений экспертов, вычислив ранговые коэффициенты корреляции Спирмена и Кендалла.

28. Имеется два набора данных $\mathbf{X}_1=\begin{pmatrix} 3 & 7 \\ 2 & 4 \\ 4 & 7 \end{pmatrix}$ и $\mathbf{X}_2=\begin{pmatrix} 6 & 9 \\ 5 & 7 \\ 4 & 8 \end{pmatrix}$, для которых $\bar{\mathbf{x}}_1=\begin{pmatrix} 3 \\ 6 \end{pmatrix}$, $\bar{\mathbf{x}}_2=\begin{pmatrix} 5 \\ 8 \end{pmatrix}$ и $\mathbf{\Sigma}=\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$. Вычислите линейную дискриминантную функцию. Классифицируйте наблюдение $\mathbf{x}_0=\begin{pmatrix} 2 & 7 \end{pmatrix}$.

29. По данным об импорте товаров 20-ти стран за 2000-2002гг. построена следующая дендограмма:



На основе анализа дендограммы классифицировать страны для K=5.

30. Компания не осуществляет инвестиционных вложений в ценные бумаги с дисперсией годовой доходности более чем 0,04. Выборка из 52 наблюдений по активу А показала, сто выборочная дисперсия ее доходности равна 0,045. Выяснить, допустимы ли для данной компании инвестиционные вложения в актив А на уровне значимости 0,05.

31. Для ковариационной матрицы трех переменных постройте факторную модель с одним фактором: найдите нагрузки фактора, значения общностей и специфических дисперсий:

$$\mathbf{\Sigma}=\begin{pmatrix} 1 & 0{,}4 & 0{,}9 \\ 0{,}4 & 1 & 0{,}7 \\ 0{,}9 & 0{,}7 & 1 \end{pmatrix}.$$

32. По таблице дисперсионного анализа найти значение коэффициента детерминации и сделать вывод о статистической значимости уравнения регрессии.

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	11918,3	5959,15	12,62	0,00445389
Остаток	17	8027,6	472,21		
Итого	19	19945,9			

33. По известной матрице факторных нагрузок $\begin{pmatrix} 0{,}76 & 0{,}42 \\ 0{,}45 & 0{,}21 \\ 0{,}65 & 0{,}37 \\ 0{,}38 & 0{,}19 \end{pmatrix}$ воспроизведите матрицу парных корреляций.

Критерии оценивания:

Максимальное количество баллов – 100.

Экзаменационный билет содержит два вопроса и одно задание.

- 84-100 баллов (оценка «отлично») – даны верные ответы на вопросы; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе. Задание выполнено в полном объеме, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов.
- 67-83 балла (оценка «хорошо») – даны верные ответы на вопросы, но с отдельными погрешностями и ошибками, уверенно исправленными после дополнительных вопросов; продемонстрировано наличие глубоких исчерпывающих / твердых и достаточно полных знаний, грамотное и логически стройное изложение материала при ответе. Задание выполнено в полном объеме с небольшими погрешностями, выбраны верные инструментальные методы и приемы решения, проведены верные расчеты, сделан полный, содержательный вывод по результатам проведенных расчетов, в расчетах и выводах содержатся незначительные ошибки.
- 50-66 баллов (оценка «удовлетворительно») – ответы на вопросы частично верны, продемонстрирована некоторая неточность ответов на дополнительные и наводящие

вопросы. Задание выполнено частично, частично выбраны верные инструментальные методы и приемы решения, проведены частичные расчеты, сделан вывод по результатам проведенных расчетов с отдельными, незначительными погрешностями.

- 0-49 баллов (оценка «неудовлетворительно) - Ответы на вопросы не верны, продемонстрирована неуверенность и неточность ответов на дополнительные и наводящие вопросы. Задание не выполнено или выполнено частично, частично выбраны необходимые инструментальные методы и приемы решения, расчеты не проведены или проведены частично, вывод по результатам проведенных расчетов не сделан или ошибочен.

Вопросы для коллоквиума

1. Основные понятия теории машинного обучения.
2. Задача обучения по прецедентам.
3. Объекты и признаки. Ответы и типы задач.
4. Модель алгоритмов и метод обучения. Этап обучения и этап применения.
5. Функционал качества. Примеры. Эмпирический риск. Переобучение, его возможные причины и способы его минимизации.
6. Сведение задачи обучения к задаче оптимизации. Примеры.
7. Обобщающая способность. Формализации понятия «обобщающая способность». Проблема обобщающей способности.
8. Этапы решения задач машинного обучения.
9. Типы задач, причисляемые к машинному обучению.
10. Понятие обучения с учителем и без. Примеры алгоритмов. Особенности.
11. Постановка задачи классификации. Примеры.
12. Постановка задачи регрессии. Примеры.
13. Постановка задачи ранжирования. Примеры.
14. Вероятностный подход в машинном обучении. Наивный Байесовский классификатор. Оценка его оптимальности и возможности применения на практике.
15. Возможность применения одних и тех же алгоритмов машинного обучения для задач разных типов. Примеры.
16. Основные алгоритмы, применяемые в задачах классификации.
17. Основные алгоритмы, применяемые в задачах регрессии.
18. Основные алгоритмы, применяемые в задачах кластеризации.
19. Основные алгоритмы, применяемые в задачах коллаборативной фильтрации.
20. Основные этапы решения практической задачи.
21. Основные проблемы исходных данных и способы их решения (пропуски, выбросы, несовместимые с алгоритмом типы данных).
22. Способы визуализации данных.
23. Методы минимизации числа признаков. Метод главных компонент (PCA).
24. Понятие интерпретируемости алгоритма и приминаемого им решения по отдельному прецеденту. Поясняющие примеры задач.
25. Функция потерь. Гладкие аппроксимации пороговой функции потерь. Основныевиды и их особенности. Примеры.
26. Нейронные сети. Структура сети.
27. Сжимающие нейронные сети. Цель. Структура и особенности.

28. Применение нейронных сетей для задач генерации признаков.
29. Глубокое обучение (Deep learning).
30. Два подхода к реализации многоклассового классификатора на основе бинарных классификаторов.
31. Статистический анализа данных. Примеры статистик и их интерпретация.
32. Сведения задачи регрессии к задаче классификации. Цель. Особенности.
33. Задача классификации на N пересекающихся классов. Примеры. Возможные подходы к решению.
34. Метрические методы классификации. Гипотеза компактности.
35. Обобщенный метрический классификатор.
36. Метод ближайшего соседа и его обобщения. Особенности, преимущества и недостатки.
37. Метод окна Парзена.
38. Метод потенциальных функций.
39. Пример алгоритма настройки весов объектов. Особенности, преимущества и недостатки.
40. Понятие отступа. Типы объектов в зависимости от отступа.
41. Понятие эталона. Два подхода к отбору эталонов.
42. Метрика. Примеры. Взвешенная метрика Минковского, ее анализ.
43. Полный скользящий контроль. Профиль компактности. Логические алгоритмы
44. Понятия закономерности, информативности и интерпретируемости.
45. Основные виды закономерностей.
46. Примеры критериев информативности. Возможные их недостатки.
47. Пример алгоритма поиска информативных закономерностей.
48. Бинарное решающее дерево.
49. Пример алгоритма построения бинарного решающего дерева (ID3). Его достоинства и недостатки.
50. Обработка пропусков в данных на стадии обучения и на стадии применения алгоритма.
51. Понятие редукции решающего дерева.Цель. Пример алгоритма.
52. Небрежные решающие деревья (ODT). Алгоритм обучения ODT.
53. Бинаризация вещественного признака. Цель. Способы разбиения области значений признака на зоны. Пример алгоритма. Линейные методы
54. Понятие оптимальной разделяющей гиперплоскости. Линейно разделимые и линейно неразделимые выборки. Обоснование кусочнолинейной функции потерь.
55. Типы объектов: периферийные, опорные (граничные и нарушители).
56. Линейный метод опорных векторов (linearSVM). Нелинейное обобщение SVM (kernelSVM). Ядра. Особенности метода проверки, является функция ядром или нет. Конструктивные методы синтеза ядер.
57. Примеры ядер.
58. Метод опорных векторов. Преимущества и недостатки.
59. Определение ROC-кривой и AUC. ХарактеристикиROC-кривой. Эффективный алгоритм вычисления AUCи его асимптотическая сложность. Коллаборативная фильтрация
60. Постановка задачи коллаборативной фильтрации. Примеры задач.
61. Определение корреляционных и латентных моделей. Сравнительный анализ возможности их применения к задачам BigData.
62. Корреляционные модели. Особенности. Примеры. Понятие холодного старта.
63. Латентные модели. Типы латентных моделей. Особенности. Обобщающая способность, методы отбора признаков
64. Внутренний и внешний критерии оценки качества обучения по прецедентам. Преимущества и недостатки.

65. Кросс-проверка (cross-validation). Полная кросс-проверка (CCV). Скользящий контроль (LOO). Кросс-проверка по k блокам, одинарная (kfoldCV) и многократная (t*k-foldCV).
66. Критерий непротиворечивости модели. Цель, преимущества и недостатки.
67. Понятие регуляризации. Цель. Примеры регуляризаторов.
68. Задача отбора признаков в логических закономерностях.
69. Задача отбора признаков. Алгоритм полного перебора. Особенности, преимущества и недостатки.
70. Задача отбора признаков. Алгоритмы Add, Del, Add-Del. Особенности, преимущества и недостатки.
71. Задача отбора признаков. Генетический алгоритм. Особенности, преимущества и недостатки. Композиция классификаторов
72. Бустинг. Основная идея. Бустинг для бинарной задачи классификации.
73. Переобучение при применении бустинга. Особенности в случаях построения комбинации простых и сложных алгоритмов.
74. Случайный лес (random forest). Особенности, преимущества и недостатки. Методы кластеризации
75. Постановка задачи кластеризации. Цели. Некорректность задачи кластеризации.
76. Основные типы кластерных структур. Особенности и схематичных чертежи.
77. Алгоритм выделения связных компонент.
78. Функции качества кластеризации
79. Визуализация кластерных структур: диаграмма вложения, дендрограмма.
80. EM-алгоритм. Цель.
81. Метод k-средних (k-means)

Критерии оценки:
Максимально количество баллов – 40.
Студент отвечает на четыре вопроса.
Максимальное количество баллов за один вопрос – 10.
10 баллов, если студент показал:
- глубокое и прочное усвоение программного материала,
- полные, последовательные, грамотные и логически излагаемые ответы,
- полные и глубокие ответы на дополнительные вопросы,

6-9 баллов, если студент показал:
- знание программного материала,
- грамотное изложение материала, без существенных неточностей в ответе на вопрос,
- умение правильно применить теоретические знания при выполнении практических задач,

1-5 баллов, если студент показал:
- усвоение основного материала,
- при ответе допускаются неточности,
- при ответе недостаточно правильные формулировки,
- нарушение последовательности в изложении программного материала,

0 баллов, если студент показал:
- незнание программного материала,
- при ответах возникают существенные ошибки.

Задания к лабораторным работам

Задание к лабораторной работе 1. Основы языка Python.

Цель: ознакомиться с основами языка Python, получить умения для выполнения дальнейших лабораторных работ.

Задания:

- изучить типизацию данных;
- научиться пользоваться циклами «for» и «while»;
- рассмотреть «ветвление» в Python;
- отработать задачи с использованием конструкции «try-except»;
- разобрать функции и пространства имён.

Задание к лабораторной работе 2. Проверка статистических гипотез

Цель: научиться применять алоритмы Python для проверки статистических гипотез.

Задания:

- тесты нормальности;
- корреляционные тесты;
- параметрические статистические проверки гипотез;
- непараметрические статистические проверки гипотез.

Задание к лабораторной работе 3. Классификация данных

Цель: научиться работать с данными при помощи визуальных инструментов и разобрать азы классификации при помощи построения простейшего классификатора со статичными параметрами.

Задания:

- научиться анализировать данные;
- сформировать понятие математических срезов;
- получить умения в работе с визуальными инструментами;
- построить классификатор на основе данных полученных при анализе;
- научиться калибровать нейросеть для получения более точных ответов.

Задание к лабораторной работе 4. Классификация методом "К-ближайших соседей"

Цель: изучить метод простейший метод классификации данных «К-ближайших соседей» и научиться производить оценку данных с помощью визуальных инструментов Python.

Задания:

- детально разобрать метод машинного обучения «К-ближайших соседей»;

- научиться работать с информацией;
- сформировать понятие математических срезов;
- получить умения в работе с визуальными инструментами;
- научиться калибровать нейросеть для получения более точных ответов.

Задание к лабораторной работе 5. Основы работы с Pandas.

Цель: научиться пользоваться библиотекой Pandas и её встроенными объектами для визуализации данных в датасетах.

Задания:

- получить умения по использованию библиотеки Pandas;
- сформировать понятия о DataFrame и Series;
- научиться строить графики с помощью scatter matrix (матрица рассеивания) и matplotlib.

Задание к лабораторной работе 6. Анализ данных с помощью Pandas

Цель: научиться пользоваться библиотекой Pandas и её встроенными объектами для анализа данных в датасетах.

Задания:

- получить умения по использованию библиотеки Pandas;
- научиться анализировать и обрабатывать данные с помощью Pandas;
- закрепить умения визуализации в Pandas.

Задание к лабораторной работе 7. Линейная регрессия

Цель: понять и научиться применять метод линейной регрессии в машинном обучении.

Задания:

- изучить модель линейного регрессора;
- произвести обучение модели;
- рассмотреть особенности данного метода машинного обучения;
- произвести предсказание на основе созданной модели.

Задание к лабораторной работе 8. Деревья решений.

Цель: познакомиться с методом машинного обучения, построенном на деревьях решений, а также научить строить сами деревья.

Задания:

- рассмотреть понятие дерева решений;
- рассмотреть варианты применения данной классификации;
- обучить модель на основе классов;
- отобразить дополнительный класс на модели и посмотреть

результат;

- рассмотреть плюсы и минусы данной модели.

Задание к лабораторной работе 9. Метод случайного леса.

Цель: сформировать понятие случайного леса, а также научиться использовать данную модель для решения задач.

Задания:

- рассмотреть понятие случайного леса;
- рассмотреть пример кода для решения простых задач;
- научить подбирать параметры модели для улучшения качества прогнозов модели.

Задание к лабораторной работе 10. Работа с OpenCV

Цель: изучить основы работы с машинным зрением и показать основные алгоритмы работы с ним.

Задания:

- разобрать импорт и просмотр изображения;
- разобрать кадрирование;
- научиться изменять размер изображения;
- научиться переворачивать изображение;
- рассмотреть способ преобразование изображения в черно-белое;
- научиться работать со сглаживанием и размытием;
- изучить метод распознавания лиц.

Критерии оценки.

Максимально число баллов – 60.

Максимальная оценка каждой лабораторной работы – 6 баллов.

- 3-6 баллов выставляется, если задача решена полностью, самостоятельно и рационально выбраны инструменты, в представленном решении обоснованно получены правильные ответы, проведен анализ, дана грамотная интерпретация полученных результатов, сделаны выводы, возможны отдельные логические и стилистические ошибки.
- 0-2 баллов выставляется, если решение неверно или отсутствует.

3. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

Процедуры оценивания включают в себя текущий контроль и промежуточную аттестацию.

Текущий контроль успеваемости проводится с использованием оценочных средств, представленных в п. 2 данного приложения. Результаты текущего контроля доводятся до сведения студентов до промежуточной аттестации.

Промежуточная аттестация проводится в форме экзамена.

Экзамен проводится по расписанию промежуточной аттестации в письменном виде. Экзаменационный билет содержит 2 теоретических вопроса. Проверка ответов и объявление результатов производится в день экзамена. Результаты аттестации заносятся в экзаменационную ведомость и зачетную книжку студента. Студенты, не прошедшие промежуточную аттестацию по графику сессии, должны ликвидировать задолженность в установленном порядке.

Приложение 2

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ОСВОЕНИЮ ДИСЦИПЛИНЫ

Учебным планом предусмотрены следующие виды занятий:

– лекционные занятия;

– лабораторные занятия.

В ходе лекционных занятий рассматриваются теоретические вопросы и практические примеры реализации методов машинного обучения, даются рекомендации для самостоятельной работы и подготовке к лабораторным занятиям.

В ходе лабораторных занятий развиваются навыки применения эконометрических методов для решения конкретных задач.

При подготовке к лабораторным занятиям студент должен:

– изучить рекомендованную учебную литературу;

– подготовить ответы на все вопросы по изучаемой теме.

В процессе подготовки к лабораторным занятиям студенты могут воспользоваться консультациями преподавателя.

Вопросы, не рассмотренные на лекционных и лабораторных занятиях, должны быть изучены студентами в ходе самостоятельной работы. Студент должен готовиться к предстоящему лабораторному занятию по всем, обозначенным в рабочей программе дисциплины вопросам.

При реализации различных видов учебной работы используются разнообразные (в т.ч. интерактивные) методы обучения, в частности, интерактивная доска для подготовки и проведения лекционных занятий. Для подготовки к занятиям, текущему контролю и промежуточной аттестации студенты могут воспользоваться электронно-библиотечными системами. Также обучающиеся могут взять на дом необходимую литературу на абонементе университетской библиотеки или воспользоваться читальными залами.