

Федеральное государственное образовательное бюджетное учреждение
высшего образования

**«ФИНАНСОВЫЙ УНИВЕРСИТЕТ ПРИ ПРАВИТЕЛЬСТВЕ
РОССИЙСКОЙ ФЕДЕРАЦИИ»
(Финансовый университет)**

**Кафедра искусственного интеллекта
Факультет информационных технологий и анализа больших данных**

Документ подписан усиленной неквалифицированной электронной подписью
Организация: Финансовый университет при Правительстве РФ
Утверждено: Проректор по учебной и методической работе Е.А. Каменева
Сертификат: LU4fisCyb/cAld9+jtegGN30DXO3U/YI
Дата: 23.09.2025 г.

М.В. Коротеев, Б. Али

Машинное обучение

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки:

02.03.01 - Математика и компьютерные науки,

Образовательная программа «Цифровая аналитика и математическая оценка
рисков»

Профиль:

«Цифровая аналитика и математическая оценка рисков»

Рекомендовано

*Факультет информационных технологий и анализа больших данных
(протокол № 2 от 18.11.2025 г.)*

Одобрено

*Кафедра искусственного интеллекта
(протокол № 3 от 05.11.2025 г.)*

© Москва 2025

СОДЕРЖАНИЕ

1.	Наименование дисциплины	3
2.	Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине	3
3.	Место дисциплины в структуре образовательной программы	6
4.	Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся	6
5.	Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий	7
5.1.	Содержание дисциплины	7
5.2.	Учебно-тематический план	9
5.3.	Содержание семинаров	10
6.	Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине	12
6.1.	Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы	12
6.2.	Перечень вопросов, заданий, тем для подготовки к текущему контролю	14
7.	Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине	17
8.	Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины	27
9.	Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины	28
10.	Методические указания для обучающихся по освоению дисциплины	29
11.	Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем	30
12.	Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине	30

1. Наименование дисциплины

«Машинное обучение».

2. Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине

Код компетенции	Наименование компетенции	Индикаторы достижения компетенции	Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции
ПКП-5	Способен применять компьютерные технологии для анализа данных	Демонстрирует знания в области компьютерных технологий, используемых для анализа данных	знать: основные понятия и этапы машинного обучения, разные модели, метрики оценки качества, принципы работы библиотек Python для анализа данных (pandas, numpy, scikit-learn, matplotlib, seaborn), а также полный цикл модели: сбор, очистка, разведочный анализ, обучение, валидация, развертывание уметь: применять библиотеки Python для предобработки данных, строить и сравнивать модели классификации/регрессии, проводить кросс-валидацию и подбор гиперпараметров, визуализировать данные и результаты, оценивать качество моделей с использованием метрик и интерпретировать полученные результаты
		Выбирает компьютерные технологии в зависимости от специфики решаемых задач	знать: специфику задач анализа данных (классификация, регрессия, кластеризация, снижение размерности), сильные и слабые стороны алгоритмов машинного обучения (линейные модели, деревья решений, ансамбли, нейронные сети, k-NN, SVM) уметь: анализировать постановку задачи, определять её тип и ограничения, выбирать подходящую модель, обосновывать выбор с учётом специфики данных и цели
		Использует компьютерные	знать: широкий спектр

		технологии для решения прикладных задач в области анализа данных	компьютерных технологий и инструментов, доступных для анализа данных уметь: применять методы машинного обучения и статистического анализа для решения конкретных прикладных задач, таких как прогнозирование, классификация и кластеризация данных; адаптировать решения на основе обратной связи и изменений в условиях задачи, чтобы обеспечивать актуальность результатов
ОПК-2	Способен проводить под научным руководством исследование на основе существующих методов в конкретной области профессиональной деятельности	Владеет инструментами и навыками подготовки аналитических обзорных научных работ и списка источников, используемых в проводимом исследовании	знать: требования и стандарты для оформления списка источников и цитирования в академических работах в соответствии с выбранным стилем ГОСТ уметь: эффективно искать и выбирать научные источники, связанные с темой машинного обучения; анализировать источники, выделять в них ключевые концепции, методы и результаты, чтобы лучше понимать современные тенденции в машинном обучении; создавать четкие и информативные аннотации для источников, чтобы быстро понимать их содержание и применимость к исследованию.
		Выбирает методики решения научных задач для достижения поставленной цели и применяет их на практике.	знать: различные методики и подходы к решению задач в области машинного обучения, включая обучение с учителем, обучение без учителя, исследование данных и другие; обработку данных для применения выбранных методик, включая предобработку и анализ данных; уметь: адаптировать выбранные методики под конкретную задачу и контекст своего исследования в области машинного обучения
		Проводит исследования под научным руководством в конкретной области	знать: методы и инструменты машинного обучения для сбора, очистки, анализа и интерпретации

		профессиональной деятельности	данных в рамках проводимых исследований уметь: применять модели машинного обучения для решения научных задач; оценивать результаты исследований по метрикам и интерпретируемости, формулировать обобщенные выводы и предлагать рекомендации для практической деятельности или дальнейших исследований;
--	--	----------------------------------	--

3. Место дисциплины в структуре образовательной программы

Дисциплина «Машинное обучение» относится к «Профилю "Цифровая аналитика и математическая оценка рисков"».

4. Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся

Очная форма обучения

Вид учебной работы по дисциплине	Всего (в з/е и часах)	Семестр 3 (в часах)	Семестр 4 (в часах)
Общая трудоёмкость дисциплины	6/216	108	108
Контактная работа- Аудиторные занятия	100	50	50
Лекции	32	16	16
Семинары, практические занятия	68	34	34
Самостоятельная работа	116	58	58
Вид текущего контроля	Контрольная работа, Контрольная работа	Контрольная работа	Контрольная работа
Вид промежуточной аттестации	Зачет, Экзамен	Зачет	Экзамен

5. Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий

5.1. Содержание дисциплины

Тема 1. Введение в машинное обучение

Основные понятия машинного обучения. Связь с другими дисциплинами. Контекст машинного обучения как дисциплины - анализ данных и искусственный интеллект. Сферы применения машинного обучения. Типы задач машинного обучения - обучение с учителем и без учителя. Структура данных для машинного обучения. Инструментальные средства машинного обучения. Понятие модели машинного обучения.

Тема 2. Регрессия

Постановка задачи регрессии. Линейная регрессия с одной переменной - функция гипотезы, функция ошибки, метод градиентного спуска. Регрессия с несколькими переменными - множественная линейная регрессия, нормализация признаков, полиномиальная регрессия. Практическое построение регрессии - загрузка и представление данных, реализация метода градиентного спуска, оценка качества регрессии, подбор скорости обучения, знакомство с библиотекой sklearn.

Тема 3. Классификация

Постановка задачи классификации как задачи машинного обучения. Отличия от задачи регрессии. Структура данных для классификации. Логистическая регрессия - функция гипотезы, граница принятия решений, функция ошибки логистической регрессии, градиентный спуск для логистической регрессии, многоклассовая классификация, алгоритм “один против всех”.

Тема 4. Методы обучения с учителем

Универсальность методов обучения с учителем, общая постановка задачи. Линейные модели - линейная и логистическая регрессии - как единая модель. Полиномиальные модели. Метод опорных векторов, ядра - линейное, гауссово, другие. Перцептрон. Деревья решений. К ближайшим соседям. Наивная байесовская модель. Достоинства и недостатки разных типов моделей, их сравнительная характеристика, применимость. Применение этих моделей для решения задач классификации и регрессии.

Тема 5. Диагностика систем машинного обучения

Метрики эффективности машинного обучения - сравнение с функциями ошибки. Типичные метрики эффективности для моделей регрессии - MAE, MSE, RMSE, MSLE, MAPE и другие. Метрики эффективности для моделей классификации - accuracy, precision, recall, F1, ROC, PR и другие. Недообучение и переобучение. Проблема bias-variance. Оценка сложности моделей. Обобщающая способность моделей, тестовый набор, кривые обучения. Методы борьбы с недо- и переобучением. Регуляризация. Задача выбора модели - кросс-валидация, гиперпараметры моделей, поиск по сетке, валидационный набор.

Тема 6. Предварительный анализ и обработка данных

Сбор данных для моделей обучения с учителем - реляционная форма данных, понятие чистых данных, оценка источников и объемов данных. Описательный (предварительный) анализ данных (EDA) - анализ репрезентативности, шкалы и типы, визуализация, проблема несбалансированности, обнаружение корреляций, аномалий в данных. Очистка и преобразование данных - удаление лишних признаков, удаление непоказательных объектов, заполнение отсутствующих значений, создание суррогатных признаков, преобразование шкал, воспроизводимость преобразования данных.

Тема 7. Задачи обучения без учителя

Задача кластеризации - постановка задачи, структура датасета, результат и интерпретация. Метод К средних - формализация, гиперпараметры, применимость. Другие методы кластеризации - DBSCAN, иерархическая, агломеративная кластеризация. Задача обнаружения аномалий. Задача понижения размерности - метод главных компонент, метод независимых компонент. Обучение с подкреплением.

Тема 8. Практическое использование моделей машинного обучения

Стохастический и пакетный градиентный спуск. Отбор признаков. Частичное обучение с учителем. Ансамблирование моделей - беггинг, бустинг, стакинг, случайный лес, XGBoost, CatBoost. Конвейеризация моделей машинного обучения. Основные этапы проекта по машинному обучению. Визуализация, интерпретация, представление и анализ результатов машинного обучения. Работа с разными типами данных - преобразование графической информации, методы векторизации текста.

5.2. Учебно-тематический план

№ п/п	Наименование тем(разделов) дисциплины	Трудоемкость в часах				
		Всего	Контактная работа - Аудиторная работа			Самостоя тельная работа
			Общая, в т.ч.:	Лекции	Семинары, практическ ие занятия	
1	Введение в машинное обучение	24	6	2	4	18
2	Регрессия	28	10	2	8	18
3	Классификация	24	6	2	4	18
4	Методы обучения с учителем	30	18	6	12	12
5	Диагностика систем машинного обучения	30	20	8	12	10
6	Предварительный анализ и обработка данных	30	18	6	12	12
7	Задачи обучения без учителя	26	12	4	8	14
8	Практическое использование моделей машинного обучения	24	10	2	8	14
	Итого	216	100	32	68	116

5.3. Содержание семинаров

Наименование тем (разделов) дисциплины	Перечень вопросов для обсуждения на семинарах, практических занятиях	Формы проведения занятий
Введение в машинное обучение	Входной контроль. Изучение технологического стека анализа данных, построенного на базе языка программирования Python. Знакомство с библиотеками numpy, pandas, matplotlib, seaborn. Выполнение задания по простому статистическому анализу данных инструментальными средствами, включающему продвинутое владение соответствующими инструментами.	Решение и обсуждение задач
Регрессия	Построение модели регрессии методом градиентного спуска своими руками. Знакомство с основами обучения параметров модели. Построение модели множественной регрессии своими руками. Построение модели библиотечными средствами. Сравнение результатов. Реализация модели регрессии на примере, близком к реальному (датасет california housing). Интерпретация результатов моделирования.	Решение и обсуждение задач
Классификация	Реализация модели классификации (логистической регрессии) своими руками наподобие линейной регрессии. Использование библиотечных функций.	Решение и обсуждение задач
Методы обучения с учителем	Построение модели классификации на простых данных. Построение нескольких моделей, сравнение их эффективности. Построение модели на сложных данных с нелинейными и слабыми зависимостями. Оценка эффективности, выбор модели. Построение модели регрессии на сложных данных. Построение разных типов регрессоров. Оценка эффективности, выбор модели.	Решение и обсуждение задач
Диагностика систем машинного обучения	Оценка эффективности модели регрессии с помощью разных метрик. Оценка эффективности классификации с помощью разных метрик. Построение диагностических кривых (PR, ROC) для бинарной классификации. Диагностика недо-и переобучения в модели классификации. Построение кривых обучения, их интерпретация. Тестовый набор данных. Демонстрация необходимости кросс-валидации. Использование кросс-валидированных оценок качества модели.	Решение и обсуждение задач

	Демонстрация оптимизации гиперпараметров модели, использование валидационного набора для выбора модели. Поиск модели от простых к сложным.	
Предварительный анализ и обработка данных	Сбор и интеграция данных из разных источников. Основные численные характеристики датасета, анализ на чистоту. Нормализация признаков и заполнение отсутствующих значений на искусственном примере. Преобразование категориальных признаков в численные. Краткий алгоритм анализа данных на реальном примере. Интерпретация результатов. Отбор и инжиниринг признаков на примере, приближенном к реальному (датасет titanic).	Решение и обсуждение задач
Задачи обучения без учителя	Решение задачи кластеризации на искусственных данных. Метод К-средних. Метод локтя. Обнаружение аномалий на искусственных данных. Обнаружение аномалий как задача предварительного анализа данных. Понижение размерности на искусственных данных. Интерпретация результата. Понижение размерности для визуализации данных.	Решение и обсуждение задач
Практическое использование моделей машинного обучения	Использование простых ансамблей моделей. Оптимизация гиперпараметров моделей. Визуализация, интерпретация и представление результатов машинного обучения исходя из бизнес-задачи моделирования. Построение конвейера машинного обучения. Построение модели классификации текстов. Знакомство с разными методами векторизации текстов. Построение модели классификации изображений. Представление изображений в виде вектора.	Решение и обсуждение задач

6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

6.1. Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы

Наименование тем (разделов) дисциплины	Перечень вопросов, отводимых на самостоятельное освоение	Формы внеаудиторной самостоятельной работы
Введение в машинное обучение	Изучение истории машинного обучения как дисциплины, связи с другими областями знания. Изучение базовых библиотек анализа данных на языке программирования Python - numpy, pandas, matplotlib, seaborn по официальной документации (при необходимости).	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Регрессия	Знакомство с нормированием данных, адаптивной скоростью обучения в ручной реализации метода градиентного спуска. Векторизация вычислений в модели множественной регрессии. Использование множественной регрессии для моделирования полиномиальных признаков и нелинейных зависимостей (корень, экспонента, синус, логарифм, обратная пропорциональность). Визуализация исходных данных и результатов моделирования в реальном случае. Анализ и представление результатов моделирования и совершение предсказания. Статистический анализ достоверности предсказания.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Классификация	Реализация алгоритма “один против всех”. Оценка эффективности классификации. Сравнение двух реализаций.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Методы обучения с учителем	Визуализация результатов классификации. Построение и анализ отчета от классификации. Оценка скорости обучения и работы разных моделей машинного обучения на большом объеме данных. Визуализация результатов классификации, интерпретация полученных результатов. Совершение предсказания и оценка его надежности.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Диагностика систем машинного обучения	Автоматизация построения модели и подсчета выбранных метрик. Построение сводного отчета. Обоснование выбора метрик эффективности	Работа с учебной литературой. Разбор вопросов по теме

	исходя из поставленной задачи. Визуализация диагностических кривых для множественной классификации. Использование регуляризованных моделей - ridge, lasso, elastic net. Демонстрация позитивного влияния регуляризации на точность модели. Построение и робастная оценка эффективности модели на данных, близких к реальным. Использование кросс-валидации. Построение модели и ее последующая оптимизация. Построение ряда простых моделей, диагностика и выбор класса модели, оптимизация и оценка качества.	занятия. Выполнение домашних заданий к каждому занятию.
Предварительный анализ и обработка данных	Анализ распределения каждого атрибута и связь с целевой переменной. Визуализация. Оценка влияния нормализации признаков на качество обучения. Заполнение отсутствующих значений на реальном примере. Группировка данных и продвинутые алгоритмы решкалирования. Полный алгоритм анализа данных на реальном примере. Анализ влияния на моделирование. Алгоритмический отбор признаков, отбор по итогам оценки важности признаков по итогам построения простых моделей.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Задачи обучения без учителя	Решение задачи кластеризации на реальных данных. Интерпретация результата. Обнаружение аномалий в реальных данных. Интерпретация результатов. Понижение размерности для задачи снижения объема данных. Понижение размерности как инжиниринг признаков.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Практическое использование моделей машинного обучения	Сложные ансамбли моделей - XGBoost, CatBoost. Оптимизация гиперпараметров, выбор модели. Использование различных типов графики, инфографики для оптимального представления результатов моделирования. Знакомство с разными задачами обработки текста - рубрикация, перевод, анализ тональности, генерация. Знакомство с разными задачами анализа изображений - идентификация объекта, обнаружение объекта, распознавание объекта.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.

6.2. Перечень вопросов, заданий, тем для подготовки к текущему контролю

Примерные задания контрольных работ

1. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Описать основные количественные характеристики датасета. Визуализировать распределение каждого признака, входящего в датасет. Визуализировать совместное распределение каждого признака и целевой переменной. Сделать вывод о линейной разделимости классов.
2. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Построить модель парной классификации каждого признака и целевой переменной. Оценить метрики эффективности четырех моделей классификации. Сделать вывод о значимости каждого признака для классификации.
3. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Описать основные количественные характеристики датасета. Построить модель классификации объектов по всем четырем признакам, используя метод деревьев решений. Построить отчет о классификации, оценить основные метрики классификации - полноту, точность, F-метрику. Сделать вывод об эффективности классификации.
4. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Описать основные количественные характеристики датасета. Построить модель классификации объектов по всем четырем признакам, используя метод логистической регрессии. Построить отчет о классификации, оценить основные метрики классификации - полноту, точность, F-метрику. Сделать вывод об эффективности классификации.
5. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Описать основные количественные характеристики датасета. Построить модель классификации объектов по всем четырем признакам, используя метод опорных векторов. Построить отчет о классификации, оценить основные метрики классификации - полноту, точность, F-метрику. Сделать вывод об эффективности классификации.
6. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Описать основные количественные характеристики датасета. Построить модель классификации объектов по всем четырем признакам, используя метод многослойного перцептрона. Построить отчет о классификации, оценить основные метрики

классификации - полноту, точность, F-метрику. Сделать вывод об эффективности классификации.

7. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Построить модель классификации объектов по всем четырем признакам, используя любой метод классификации. Оптимизировать гиперпараметры модели при помощи поиска по сетке. Сделать вывод об улучшении эффективности классификации.

8. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Построить модель классификации объектов по всем четырем признакам, используя любой метод. Оценить эффективность модели при помощи перекрестной проверки (кросс-валидации). Сделать вывод об эффективности классификации.

9. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Понизить размерность датасета до двух измерений. Визуализировать получившийся набор данных учитывая значение целевой переменной. Сделать вывод о линейной разделимости классов.

10. Загрузить встроенный в библиотеку sklearn датасет “Бостон”. Описать основные количественные характеристики датасета. Охарактеризовать шкалу измерения каждого признака (вид шкалы, минимальное, максимальное значение, количество значений). Визуализировать совместное распределение каждого признака и целевой переменной. Сделать вывод о значимости каждого признака для регрессии.

11. Загрузить встроенный в библиотеку sklearn датасет “Бостон”. Построить модель множественной регрессии. Оценить метрики эффективности регрессии. Сделать вывод об эффективности получившейся модели.

12. Загрузить встроенный в библиотеку sklearn датасет “Бостон”. Построить модель регрессии по всем признакам, используя любой метод регрессии. Оптимизировать гиперпараметры модели при помощи поиска по сетке. Сделать вывод об улучшении эффективности регрессии.

13. Загрузить встроенный в библиотеку sklearn датасет “Бостон”. Построить модель регрессии по всем признакам, используя любой метод. Оценить эффективность модели при помощи перекрестной проверки (кросс-валидации). Сделать вывод об эффективности регрессии.

14. Загрузить встроенный в библиотеку sklearn датасет “Бостон”. Понизить размерность датасета до двух измерений. Визуализировать получившийся набор данных учитывая значение целевой переменной. Сделать вывод.

15. Загрузить встроенный в библиотеку sklearn датасет “Цифры”. Построить модель классификации объектов, используя метод многослойного перцептрона. Построить отчет о классификации, оценить основные метрики классификации - полноту, точность, F-метрику. Сделать вывод об эффективности классификации.

16. Загрузить встроенный в библиотеку sklearn датасет “Цифры”. Построить модель классификации объектов, используя любой метод классификации. Оптимизировать гиперпараметры модели при помощи поиска по сетке. Сделать вывод об улучшении эффективности классификации.

17. Загрузить встроенный в библиотеку sklearn датасет “Цифры”. Построить модель классификации объектов, используя любой метод. Оценить эффективность модели при помощи перекрестной проверки (кросс-валидации). Сделать вывод об эффективности классификации.

18. Загрузить встроенный в библиотеку sklearn датасет “Цифры”. Понизить размерность датасета до двух измерений. Визуализировать получившийся набор данных учитывая значение целевой переменной. Сделать вывод о линейной делимости классов.

Критерии балльной оценки различных форм текущего контроля успеваемости содержатся в соответствующих методических рекомендациях Кафедры искусственного интеллекта Факультета информационных технологий и анализа больших данных.

7. Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине

Перечень компетенций с указанием индикаторов их достижения в процессе освоения образовательной программы содержится в разделе 2. *Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине.*

Типовые контрольные задания или иные материалы, необходимые для оценки индикаторов достижения компетенций, умений и знаний

ПКП-5 Способен применять компьютерные технологии для анализа данных

1) Демонстрирует знания в области компьютерных технологий, используемых для анализа данных

Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции

Знать: основные понятия и этапы машинного обучения, разные модели, метрики оценки качества, принципы работы библиотек Python для анализа данных (pandas, numpy, scikit-learn, matplotlib, seaborn), а также полный цикл модели: сбор, очистка, разведочный анализ, обучение, валидация, развертывание

Уметь: применять библиотеки Python для предобработки данных, строить и сравнивать модели классификации/регрессии, проводить кросс-валидацию и подбор гиперпараметров, визуализировать данные и результаты, оценивать качество моделей с использованием метрик и интерпретировать полученные результаты

Типовые контрольные задания

- Предобработка датасета в pandas.
- Обучение и сравнение 3 моделей scikit-learn с метриками.
- Кросс-валидация + GridSearchCV.
- Визуализация + выводы.

2) Выбирает компьютерные технологии в зависимости от специфики решаемых задач

Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции

Знать: специфику задач анализа данных (классификация, регрессия, кластеризация, снижение размерности), сильные и слабые стороны алгоритмов машинного обучения (линейные модели, деревья решений, ансамбли, нейронные сети, k-NN, SVM)

Уметь: анализировать постановку задачи, определять её тип и ограничения, выбирать подходящую модель, обосновывать выбор с учётом специфики данных и цели

Типовые контрольные задания

- Анализ задачи + обоснование выбора 2–3 модели.
- Сравнение выбранных моделей на одном датасете (метрики + вывод).

3) Использует компьютерные технологии для решения прикладных задач в области анализа данных

Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции

Знать: широкий спектр компьютерных технологий и инструментов, доступных для анализа данных

Уметь: применять методы машинного обучения и статистического анализа для решения конкретных прикладных задач, таких как прогнозирование, классификация и кластеризация данных; адаптировать решения на основе обратной связи и изменений в условиях задачи, чтобы обеспечивать актуальность результатов

Типовые контрольные задания

- Формулировка задачи + загрузка/подготовка реального датасета.
- End-to-end пайплайн с Pipeline и сохранением модели.
- Применение модели к новым данным + предсказание.
- Краткий отчёт с бизнес-интерпретацией.

ОПК-2 Способен проводить под научным руководством исследование на основе существующих методов в конкретной области профессиональной деятельности

1) Владеет инструментами и навыками подготовки аналитических обзорных научных работ и списка источников, используемых в проводимом исследовании

Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции

Знать: требования и стандарты для оформления списка источников и цитирования в академических работах в соответствии с выбранным стилем ГОСТ

Уметь: эффективно искать и выбирать научные источники, связанные с темой машинного обучения; анализировать источники, выделять в них ключевые концепции, методы и результаты, чтобы лучше понимать современные тенденции в машинном обучении; создавать четкие и информативные аннотации для источников, чтобы быстро понимать их содержание и применимость к исследованию.

Типовые контрольные задания

Написание аннотаций (80–120 слов) к 3 выбранным статьям с выделением: цель, метод, ключевой результат, применимость к исследованию.

2) Выбирает методики решения научных задач для достижения поставленной цели и применяет их на практике.

Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции

Знать: различные методики и подходы к решению задач в области машинного обучения, включая обучение с учителем, обучение без учителя, исследование данных и другие; обработку данных для применения выбранных методик, включая предобработку и анализ данных;

Уметь: адаптировать выбранные методики под конкретную задачу и контекст своего исследования в области машинного обучения

Типовые контрольные задания

Обоснование выбора методики по задаче «Прогнозирование оттока клиентов» сравнить 3 подхода (логистическая регрессия, Random Forest, XGBoost) с указанием критериев выбора (размер данных, интерпретируемость, метрика AUC) в таблице.

3) Проводит исследования под научным руководством в конкретной области профессиональной деятельности

Результаты обучения (умения и знания), соотнесенные с индикаторами достижения компетенции

Знать: методы и инструменты машинного обучения для сбора, очистки, анализа и интерпретации данных в рамках проводимых исследований

Уметь: применять модели машинного обучения для решения научных задач; оценивать результаты исследований по метрикам и интерпретируемости, формулировать обобщенные выводы и предлагать рекомендации для практической деятельности или дальнейших исследований;

Типовые контрольные задания

- Сбор и подготовка данных: загрузить датасет, выполнить очистку, разведочный анализ с визуализациями.
- Применение модели: обучить несколько моделей машинного обучения, вывести метрики.
- Оценка и выводы: сравнительная таблица метрик, интерпретация топ-5 признаков, формулировка выводов.

Примеры практико-ориентированных заданий

Практическая часть представляет собой задачу по анализу конкретного датасета. Для каждого задания будет необходимо:

1. Загрузить датасет в Python.
2. Описать набор данных и решаемую задачу.
3. Выделить целевую переменную и факторные переменные.
4. Удалить ненужные данные, проанализировать отсутствующие значения.
5. Прокомментировать количественные параметры датасета.
6. Разбить выборку на обучающую и тестовую.
7. Работа по вариантам.

Вариант 1. Очистка данных и обучение моделей.

Данный вариант предполагает фокусировку на обучении нескольких видов моделей обучения с учителем. В зависимости от набора данных, может предполагаться задача классификации и регрессии. Необходимо после минимальной подготовки датасета к обучению обучить несколько моделей и сравнить их эффективность.

Вариант 2. Описательный анализ и визуализация данных.

Данный вариант предполагает фокусировку на исследовании данных и визуализации. При решении этого варианта следует провести как можно более подробный описательный анализ данных с использованием максимального спектра средств визуализации. При этом следует делать значимые выводы об обнаруженных в данных закономерностях.

Вариант 3. Построение модели и оптимизация гиперпараметров.

Данный вариант предполагает фокусировку на процессе улучшения эффективности модели обучения с учителем. Студенту следует подготовить датасет к обучению, обучить одну из моделей с учителем со значениями гиперпараметров по умолчанию, получить значение эффективности. После этого вручную или автоматически подобрать значения гиперпараметров таким образом, чтобы получить максимальный прирост эффективности.

Вариант 4. Выбор признаков.

Данный вариант предполагает фокусировку на улучшении модели путем ввода новых признаков в модель. Следует подготовить модель к обучению, обучить модель и зафиксировать начальный уровень эффективности. Затем следует исследовать влияние исключения существующих и введения новых признаков в модель на эффективность. Как вариант можно рассматривать введение полиномиальных признаков. Следует стремиться к максимальному увеличению эффективности модели.

Вариант 5. Исследование влияния обучения без учителя на эффективность обучения.

Данный вариант предполагает фокусировку на использовании методов обучения без учителя для ускорения или повышения эффективности обучения с учителем. Следует подготовить модель к обучению, обучить модель и зафиксировать начальный уровень эффективности. Затем следует попробовать применить понижение размерности, обнаружение аномалий или кластеризацию (в любой комбинации) для трансформации исходного датасета. В конце работы следует сделать значимый вывод об изменении скорости и эффективности обучения с учителем.

Примерные вопросы для подготовки к зачету, экзамену

Примерные вопросы для подготовки к зачету

1. Понятие машинного обучения. Отличие машинного обучения от других областей программирования.
2. Классификация задач машинного обучения. Примеры задач из различных классов.
3. Основные понятия машинного обучения: набора данных, объекты, признаки, атрибуты, модели, параметры.
4. Структура и представление данных для машинного обучения.
5. Инструментальные средства машинного обучения.

6. Задача регрессии: постановка, математическая формализация.
7. Метод градиентного спуска для парной линейной регрессии.
8. Понятие функции ошибки: требования, использование, примеры.
9. Множественная и нелинейная регрессии.
10. Нормализация признаков в задачах регрессии.
11. Задача классификации: постановка, математическая формализация.
12. Метод градиентного спуска для задач классификации.
13. Логистическая регрессия в задачах классификации.
14. Множественная и многоклассовая классификация. Алгоритм “один против всех”.
15. Метод опорных векторов в задачах классификации.
16. Понятие ядра и виды ядер в методе опорных векторов.
17. Метод решающих деревьев в задачах классификации.
18. Метод k ближайших соседей в задачах классификации.
19. Однослойный перцептрон в задачах классификации.

Примерные вопросы для подготовки к экзамену

1. Понятие машинного обучения. Отличие машинного обучения от других областей программирования.
2. Классификация задач машинного обучения. Примеры задач из различных классов.
3. Основные понятия машинного обучения: набора данных, объекты, признаки, атрибуты, модели, параметры.
4. Структура и представление данных для машинного обучения.
5. Инструментальные средства машинного обучения.
6. Задача регрессии: постановка, математическая формализация.
7. Метод градиентного спуска для парной линейной регрессии.
8. Понятие функции ошибки: требования, использование, примеры.
9. Множественная и нелинейная регрессии.

10. Нормализация признаков в задачах регрессии.
11. Задача классификации: постановка, математическая формализация.
12. Метод градиентного спуска для задач классификации.
13. Логистическая регрессия в задачах классификации.
14. Множественная и многоклассовая классификация. Алгоритм “один против всех”.
15. Метод опорных векторов в задачах классификации.
16. Понятие ядра и виды ядер в методе опорных векторов.
17. Метод решающих деревьев в задачах классификации.
18. Метод k ближайших соседей в задачах классификации.
19. Однослойный перцептрон в задачах классификации.
20. Метрики эффективности и функции ошибки: назначение, примеры, различия.
21. Понятие набора данных (датасета) в машинном обучении. Требования, представление. Признаки и объекты.
22. Шкалы измерения признаков. Виды шкал, их характеристика.
23. Понятие чистых данных. Определение, очистка данных.
24. Основные этапы проекта по машинному обучению.
25. Предварительный анализ данных: задачи, методы, цели.
26. Проблема отсутствующих данных: причины, исследование, пути решения.
27. Проблема несбалансированных классов: исследование, пути решения.
28. Понятие параметров и гиперпараметров модели. Обучение параметров и гиперпараметров. Поиск по сетке.
29. Понятие недо- и переобучения. Определение, пути решения.
30. Диагностика модели машинного обучения. Методы, цели.
31. Проблема выбора модели машинного обучения. Сравнение моделей.
32. Измерение эффективности работы моделей машинного обучения. Метрики эффективности.

33. Метрики эффективности моделей классификации. Виды, характеристика, выбор.
34. Метрики эффективности моделей регрессии. Виды, характеристика, выбор.
35. Перекрестная проверка (кросс-валидация). Назначение, схема работы.
36. Конвейеры в библиотеке sklearn. Назначение, использование.
37. Использование методов визуализации данных для предварительного анализа.
38. Исследование коррелированности признаков: методы, цели, выводы.
39. Решкалирование данных. Виды, назначение, применение. Нормализация и стандартизация данных.
40. Преобразование категориальных признаков в числовые.
41. Методы визуализации данных для машинного обучения.
42. Задача выбора модели. Оценка эффективности, валидационный набор.
43. Кривые обучения для диагностики моделей машинного обучения.
44. Регуляризация моделей машинного обучения. Назначение, виды, формализация.
45. Проблема сбора и интеграции данных для машинного обучения.
46. Понятие чистых данных и требования к данным.
47. Основные задачи описательного анализа данных.
48. Полиномиальные модели машинного обучения.
49. Основные виды преобразования данных для подготовки к машинному обучению.
50. Задача выбора признаков в машинном обучении.
51. Задачи обучения без учителя: общая характеристика, особенности, примеры.
52. Задача кластеризации. Формализация, применение, примеры, общая характеристика. Метрики качества кластеризации.
53. Алгоритм кластеризации K-средних.
54. Иерархическая (агломеративная) кластеризация.
55. Плотностные алгоритмы кластеризации. DBSCAN.
56. Задача понижения размерности в машинном обучении. Метод главных компонент.

57. Алгоритм t-SNE. Общая характеристика, применение, особенности.
58. Задача обнаружения аномалий в машинном обучении. Использование многомерного гауссова распределения.
59. Алгоритм DBSCAN для задач обнаружения аномалий.
60. Ансамбли моделей машинного обучения.
61. Случайный лес как ансамблевая модель.
62. Стекинг как вид ансамблирования моделей. Общая характеристика, особенности.
63. Беггинг как вид ансамблирования моделей. Общая характеристика, особенности.
64. Бустинг как вид ансамблирования моделей. Общая характеристика, особенности.
65. Алгоритм градиентного бустинга. Общая характеристика. LightGBM, XGBoost, CatBoost.

Примерные задания для подготовки к экзамену

1. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Несколькоми способами, в том числе графическим, убедиться в отсутствии пропущенных значений.
2. Загрузить встроенный в библиотеку sklearn датасет “Диабет”. Визуализировать распределение четырех любых признаков, входящих в датасет. Сделать содержательные выводы по полученным данным.
3. Загрузить встроенный в библиотеку sklearn датасет “Рак груди”. Построить модель бинарной классификации любым методом. Вывести несколько первых теоретических и эмпирических значений целевой переменной. Сделать выводы по полученным результатам.
4. Загрузить встроенный в библиотеку sklearn датасет “Вина”. Построить линейную модель обучения с учителем, вывести и проинтерпретировать коэффициенты линейной модели. Коэффициенты должны выводиться вместе с названием соответствующего признака.
5. Загрузить встроенный в библиотеку sklearn датасет “Калифорния”. Построить модель регрессии любым методом. Оптимизировать гиперпараметры модели при помощи поиска по сетке. Сделать выводы.

6. Загрузить встроенный в библиотеку sklearn датасет “Ирисы”. Построить модель множественной классификации любым методом. Оценить ее эффективность при помощи кросс-валидации. Сделать выводы.
7. Загрузить встроенный в библиотеку sklearn датасет “Диабет”. Построить модель регрессии по методу опорных векторов с линейным ядром. Оценить ее эффективность по метрикам r^2 , mae , rmse , mape . Сделать выводы о применимости модели.
8. Загрузить встроенный в библиотеку sklearn датасет “Рак груди”. Построить модель бинарной линейной классификации. Задать значения аргументов конструктора объекта модели, отличающиеся от значений по умолчанию. Пояснить смысл каждого аргумента.
9. Загрузить встроенный в библиотеку sklearn датасет “Вина”. Построить модель множественной классификации по методу опорных векторов с полиномиальным ядром. Оценить ее эффективность по метрикам accuracy , precision , recall , f 1. Сделать выводы о применимости модели.
10. Загрузить встроенный в библиотеку sklearn датасет “Калифорния”. Построить модель регрессии с регуляризацией. Задать значения аргументов конструктора объекта модели, отличающиеся от значений по умолчанию. Пояснить смысл каждого аргумента.

Пример экзаменационного билета

Экзаменационный билет №

1. Метрики эффективности моделей классификации. Виды, характеристика, выбор. **(20 баллов)**
2. Загрузить встроенный в библиотеку sklearn датасет “Калифорния”. Построить модель регрессии любым методом. Оптимизировать гиперпараметры модели при помощи поиска по сетке. Сделать выводы. **(40 баллов)**

8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины

Основная литература:

1. Коротеев, М.В. Практикум по машинному обучению на Python. (2 часть) : учебное пособие по дисциплине "Машинное обучение" для студентов, обучающихся по направлению: "Прикладная информатика", всех профилей (программы подготовки бакалавров) / М.В. Коротеев, В.А. Одинцова ; Финуниверситет, Кафедра анализа данных и машинного обучения Факультета информационных технологий и анализа больших данных Москва : Финуниверситет, 2024 1 файл (1,87 Мб) : ил. Свободный доступ из сети Интернет (чтение) <http://elib.fa.ru/fbook/books139159.pdf> (дата обращения 22.05.2024) + Текст : электронный. [БИК ID: RU\FA\books\139159]
2. Колдаев, Виктор Дмитриевич Структуры и алгоритмы обработки данных : Учебное пособие / Московский институт электронной техники 1 Москва : Издательский Центр РИОР, 2021 296 с. (Высшее образование: Бакалавриат) ВО - Бакалавриат <https://znanium.com/catalog/document?id=398591> ISBN 978-5-369-01264-2 ISBN 978-5-16-101275-8 (электр. издание) ISBN 978-5-16-009012-2 (ISBN соиздателя) . [БИК ID: RU\infra-m\znanium\bibl\1230215]
3. Коротеев, М. В. Практикум по машинному обучению на Python [Электронный ресурс] : учебное пособие по дисциплине: «машинное обучение» для студентов, обучающихся по направлению: 09.03.03 «прикладная информатика», всех профилей (программы подготовки бакалавров), институт онлайн-образования / Коротеев М. В.,Одинцова В. А.,Плешакова Е. С. Москва : Финансовый университет, 2023 161 с. Книга из коллекции Финансовый университет - Информатика <https://e.lanbook.com/book/431102>. [БИК ID: RU-LAN-BOOK-431102]

Дополнительная литература:

1. Златопольский, Дмитрий Михайлович Программирование: типовые задачи, алгоритмы, методы : Учебное пособие / Московский городской педагогический университет 5 Москва : Лаборатория знаний, 2025 226 с. ВО - Бакалавриат <https://znanium.ru/catalog/document?id=459643> ISBN 978-5-93208-832-6. [БИК ID: RU\infra-m\znanium\bibl\2191862]

9. Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины

1. •Электронная библиотека Финансового университета (ЭБ) <http://elib.fa.ru/>
2. Электронно-библиотечная система BOOK.RU <http://www.book.ru>
3. •Электронно-библиотечная система «Университетская библиотека ОНЛАЙН» <http://biblioclub.ru/>
4. •Электронно-библиотечная система Znanium <http://www.znanium.ru/>
5. •Электронно-библиотечная система издательства «ЮРАЙТ» <https://urait.ru/>
6. •Электронно-библиотечная система издательства Проспект <http://ebs.prospekt.org/books>
7. •Электронно-библиотечная система издательства Лань <https://e.lanbook.com/>
8. •Деловая онлайн-библиотека Alpina Digital <http://lib.alpinadigital.ru/>
9. Научная электронная библиотека eLibrary.ru <http://elibrary.ru>
10. •Национальная электронная библиотека <http://нэб.рф/>
11. Финансовая справочная система «Финансовый директор» <http://www.1fd.ru/>
12. PyLru 1.0.9 <https://pypi.python.org/pypi/pylru>
13. <http://pandas.pydata.org/>
14. Python Documentation <http://python.org/doc/>
15. Python Standard Library <https://docs.python.org/2/library/>

10. Методические указания для обучающихся по освоению дисциплины

При изучении теоретического материала необходимо опираться на рабочую программу дисциплины, материалы лекций и литературу из основного списка. Кроме этого, необходимо активно работать с Интернет-источниками и пособиями других авторов, помогающими усвоить материал отдельных разделов программы.

Необходимо конспектировать лекции, пометая сложные и непонятные моменты с тем, чтобы задать вопросы лектору в конце лекции или же на консультации.

При подготовке к семинарским занятиям необходимо изучить вопросы, вынесенные на самостоятельное изучение, так как семинарские занятия предполагают их обсуждение и дискуссию по теме; кроме того, задания для самостоятельной работы необходимы для того, чтобы успешно выполнить самостоятельные задания на семинарах.

Индивидуальные задания для работы на компьютере, файлы с выполненными заданиями необходимо хранить в личной сетевой папке в компьютерной сети вуза.

Для выполнения кейса задачи, рекомендуется использовать язык программирования Python и специализированные библиотеки для анализа данных и машинного обучения.

Студент должен прислать рабочий notebook с решением кейса задачи, комментариями кода и аналитическими выводами до конца экзамена на электронную почту преподавателя.

Критерии оценки практической экзаменационной работы:

1. Структурированность отчета. В работе должна прослеживаться четкая структура - подготовительный этап, анализ данных, построение простых моделей, сравнение и анализ моделей, выводы, построение моделей с учетом выводов, итоговый результат.
2. Наличие выводов. Работа должна содержать текстовые замечания, поясняющие каждый шаг работы студента: что делается, зачем и какую информацию это нам дает. Оценивается полнота и адекватность выводов.
3. Визуализация. Работа должна демонстрировать навыки студента визуализировать информацию. Особенно на этапах описательного анализа и анализа обучаемости модели. Оценивается разнообразие, наглядность и информативность визуализации.
4. Использование метрик эффективности. Оценивается разнообразие и адекватность задаче примененных метрик эффективности (включая время обучения) а также полнота сравнения и правильность выводов из сравнения моделей по разным метрикам.

5. Валидность результатов. Студент должен продемонстрировать умение оценивать достоверность измерения метрик моделей и повышать ее с использованием перекрестной проверки (кросс-валидации). Использование k-fold cross validation является предпочтительным методом измерения эффективности модели. Если происходит выбор модели, то ее итоговая эффективность должна измеряться на чистом наборе данных.

11. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем

Комплект лицензионного программного обеспечения:

1. Python 3.8
2. Anaconda
3. Jupyter Notebook (Windows 10)
4. Kaspersky

Современные профессиональные базы данных и информационные справочные системы:

1. Информационно-правовая система «Гарант»
2. Информационно-правовая система «Консультант Плюс»
3. Электронная энциклопедия: <http://ru.wikipedia.org/wiki/Wiki>
4. Система комплексного раскрытия информации «СКРИН» - <http://www.skrin.ru/>

Сертифицированные программные и аппаратные средства защиты информации:

1. не предусмотрены

12. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

1. **Компьютерный класс** для проведения учебных занятий, предусмотренных программой, в том числе групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенный оборудованием и техническими

средствами обучения: мебель аудиторная (столы, стулья, доска аудиторная), персональные компьютеры, набор демонстрационного оборудования (проектор, экран)

2. Учебная аудитория для проведения учебных занятий, предусмотренных программой, в том числе групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, оснащенная оборудованием и техническими средствами обучения: мебель аудиторная (столы, стулья, доска аудиторная), набор демонстрационного оборудования (проектор, экран)