

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: erzT/EqFVUdkz/Jg4dxg9ph+l47DPPRW
Дата: 22.09.2025 г.

В.А. Одинцова, М.В. Коротеев

Машинное обучение

Рабочая программа дисциплины

для студентов, обучающихся по направлению подготовки:

09.03.04 - Программная инженерия,

Образовательная программа «Разработка и внедрение информационно -
аналитических систем»

Рекомендовано

*Факультет информационных технологий и анализа больших данных
(протокол № 03 от 16.12.2025 г.)*

Одобрено

*Кафедра искусственного интеллекта
(протокол № 4 от 05.12.2025 г.)*

© Москва 2026

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	4
4.	Объем дисциплины (модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся	4
5.	Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий	5
5.1.	Содержание дисциплины	5
5.2.	Учебно-тематический план	9
5.3.	Содержание семинаров	10
6.	Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине	13
6.1.	Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы	13
6.2.	Перечень вопросов, заданий, тем для подготовки к текущему контролю	15
7.	Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине	21
8.	Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины	29
9.	Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины	29
10.	Методические указания для обучающихся по освоению дисциплины	31
11.	Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных справочных систем	31
12.	Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине	32

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
ОПК-1	Способен применять естественнонаучные и общеинженерные знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности	Демонстрирует знания основных методов математического анализа и моделирования, применяет их на практике для решения задач профессиональной деятельности	знать: основные понятия машинного обучения и интеллектуальной обработки данных, их области применения и сравнительные характеристики уметь: применять основные методы обучения с учителем и без учителя для решения типовых и практико-ориентированных задач машинного обучения и анализа данных
		Проводит теоретические исследования по выбранной области профессиональной деятельности	знать: основы машинного обучения и математики, связанной с этой областью уметь: создавать и настраивать модели машинного обучения. Анализировать данные и применять их в моделях. Критически оценивать методы машинного обучения. Эффективно коммуницировать результаты своих исследований
		Проводит численные эксперименты на основе математических или информационных методов, делает выводы и обосновывает их	знать: основы машинного обучения и статистики; методы и модели машинного обучения уметь: планировать и проводить численные эксперименты; анализировать данные и извлекать выводы из численных экспериментов; обосновывать полученные результаты на основе математических и статистических методов; эффективно коммуницировать результаты своих исследований;

3. Место дисциплины в структуре образовательной программы

Дисциплина «Машинное обучение» относится к «Общепрофессиональному циклу».

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	10/360	180	180
Контактная работа- Аудиторные занятия	136	68	68
Лекции	32	16	16
Семинары, практические занятия	104	52	52
Самостоятельная работа	224	112	112
Вид текущего контроля	Контрольная работа, Проектная работа	Контрольная работа	Проектная 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. Основные понятия и инструменты машинного обучения

Введение в машинное обучение. Определение машинного обучения и его место в структуре искусственного интеллекта, отличие от классического программирования. Математические дисциплины в основе машинного обучения. Метод градиентного спуска. Краткая история развития машинного обучения и науки о данных. Данные в машинном обучении. Типы задач машинного обучения: обучение с учителем (регрессия, классификация), обучение без учителя (кластеризация, снижение размерности). Примеры применения в индустрии. Формальная постановка задачи обучения с учителем: матрица объектов-признаков X , вектор ответов y , функция потерь. Типы признаков. Этапы построения проекта по машинному обучению.

Инструменты для изучения на практике классического машинного обучения. Библиотеки Python: NumPy, Pandas, Matplotlib, Seaborn и Scikit-learn.

Тема 2. Линейная регрессия

Задача регрессии. Модель линейной регрессии. Функция потерь MSE. Аналитическое решение методом наименьших квадратов. Нормальное уравнение. Реализация методом градиентного спуска.

Тема 3. Линейная бинарная классификация

Задача классификации. Бинарная классификация. Сведение к задаче регрессии путём интерпретации выхода, как вероятности. Сигмоидная функция и её свойства. Модель логистической регрессии. Реализация методом градиентного спуска. Порог принятия решения. Преобразование вероятности в метку класса.

Тема 4. Масштабирование признаков. Применение конвейера

Проблема разных масштабов признаков. Влияние на скорость схождения градиентного спуска. Методы масштабирования: StandardScaler, MinMaxScaler. Сравнение методов и критерии выбора. Правильное применение масштабирования без «утечки данных» (data leakage). Значимость признаков в линейных моделях. Конвейеры (Pipeline) в scikit-learn. Концепция последовательной обработки данных.

Тема 5. Метрики качества регрессии и бинарной классификации

Метрики качества регрессии. Коэффициент детерминации R^2 . Диапазон значений. Недостаток коэффициента детерминации. Ошибки регрессии: MAE, MSE, RMSE, MAPE.

Метрики качества бинарной классификации. Матрица ошибок в бинарной классификации (четыре исхода: TP, TN, FP, FN). Ошибки первого и второго рода. Метрики на основе матрицы ошибок: Accuracy, Precision, Recall, Specificity. F1-score. ROC-анализ. ROC-кривая: построение и интерпретация. Метрика ROC-AUC. Precision-Recall анализ. PR-кривая.

Тема 6. Выделение данных для тестирования

Обучающая способность модели. Отложенная выборка и кросс-валидация. Типы разбиений для кросс-валидации: K-Fold, Stratified K-Fold, Leave-One-Out.

Тема 7. Линейная многоклассовая классификация. Метрики качества многоклассовой классификации

Постановка задачи многоклассовой классификации. Стратегии сведения к бинарной классификации: One-vs-Rest (OvR) и One-vs-One (OvO). Сравнение подходов. Многоклассовая логистическая регрессия. Softmax-функция. Кросс-энтропия для многих классов.

Матрица ошибок многоклассовой классификации. Метрики качества многоклассовой классификации: Precision, Recall, F1 для каждого класса. Усреднение метрик: macro, micro, weighted. ROC-анализ для многоклассовой классификации. ROC-кривые методом OvR. Micro-average и macro-average ROC.

Тема 8. Дисбаланс классов

Проблема дисбаланса классов и его диагностика. Степени дисбаланса. Выбор правильных метрик. Борьба с дисбалансом классов: методы на уровне данных (Undersampling, Oversampling, SMOTE), методы на уровне алгоритма (взвешивание классов, настройка порога классификации). Комбинирование методов.

Тема 9. Полиномиальные модели

Нелинейные зависимости в данных. Полиномиальные признаки. Степень полинома и количество новых признаков. Полиномиальная регрессия, как линейная модель на расширенном пространстве признаков. Полиномиальная классификация, как

логистическая регрессия с полиномиальными признаками. Нелинейные разделяющие границы.

Тема 10. Недообучение и переобучение

Недообучение (Underfitting) и переобучение (Overfitting) модели. Разрыв между значениями метрик на обучающих и тестовых данных. Понятие сложности модели. Смещение и разброс (Bias-Variance). Декомпозиция ошибки обобщения. Высокое смещение (bias) при недообучении и высокий разброс (variance) при переобучении. Диагностика при помощи построения валидационных кривых (зависимость метрики от гиперпараметра модели) и кривых обучения (зависимость метрики от объёма данных).

Методы борьбы с переобучением. Выбор оптимального значения гиперпараметра сложности модели путём нахождения компромисса bias-variance.

Обобщённые линейные модели. L2-регуляризация (Ridge), L1-регуляризация (Lasso) и комбинация L1 и L2 (Elastic-Net). Уменьшение степени переобученности модели путём выбора коэффициента регуляризации. Проблема мультиколлинеарности.

Методы борьбы с недообучением модели. Подбор наилучших гиперпараметров модели. Различие между параметрами и гиперпараметрами. Выделение валидационного набора. Поиск по сетке (Grid Search) и рандомизированный поиск по сетке (Random Search).

Тема 11. Основные модели классического машинного обучения

Метод k ближайших соседей (k-NN). Принцип классификации по соседям. Выбор числа соседей и метрики расстояния. Проклятие размерности.

Метод опорных векторов (SVM). Максимизация зазора между классами. Линейный SVM. Параметр регуляризации C. Ядровый трюк. Ядра: линейное, полиномиальное, RBF. Параметр gamma для RBF-ядра.

Наивный байесовский классификатор. Теорема Байеса в классификации. Виды: Gaussian, Multinomial, Bernoulli.

Деревья решений (Decision Trees). Структура дерева. Алгоритм построения дерева. Критерии разбиения (Gini, Entropy, MSE). Визуализация графа дерева решений. Важность признаков.

Ансамблевые методы: Bagging и bootstrap-агрегирование: случайный лес (Random Forest) и Boosting: градиентный бустинг (GradientBoosting). Обзор оптимизированных

реализаций градиентного бустинга XGBoost, LightGBM, CatBoost. Out-of-bag оценка качества.

Сравнение моделей. Критерии выбора модели.

Тема 12. Предварительный анализ и обработка данных

Разведочный анализ данных (EDA). Изучение структуры данных. Описательная статистика. Визуализация распределений и корреляций.

Работа с пропущенными значениями. Типы пропусков: MCAR, MAR, MNAR. Удаление и импутация. Одномерная и многомерная импутация пропусков.

Обработка категориальных признаков. Label Encoding для порядковых признаков и One-Hot Encoding для номинальных признаков. Обработка высококординальных признаков (с очень большим числом уникальных значений).

Обнаружение и обработка выбросов. Типы выбросов. Одномерные методы обнаружения: IQR, Z-score. Методы для многомерных данных: Isolation Forest, Local Outlier Factor. Стратегии обработки: удаление, обрезка, замена на границы, трансформация (логарифмическое преобразование). Создание индикатора выброса.

Создание новых признаков (Feature Engineering). Математические преобразования. Бинаризация и дискретизация.

Отбор признаков (Feature Selection). Фильтрующие, обёрточные и встроенные методы.

Тема 13. Задачи обучения без учителя

Введение в обучение без учителя. Типы задач: кластеризация (Clustering), снижение размерности (Dimensionality Reduction) и другие (поиск аномалий).

Кластеризация методом k-средних (K-Means). Алгоритм. Инициализация центроидов. Выбор числа кластеров: метод локтя, метод силуэта.

Плотностная кластеризация (DBSCAN). Три типа точек. Параметры модели: радиус окрестности и минимальное число точек в окрестности.

Иерархическая кластеризация, дендрограммы.

Задача снижения размерности. Метод главных компонент (PCA). Проекция на направления максимальной дисперсии. Выбор числа компонент.

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	Основные понятия и инструменты машинного обучения	27	11	3	8	16
2	Линейная регрессия	13	5	1	4	8
3	Линейная бинарная классификация	9	3	1	2	6
4	Масштабирование признаков. Применение конвейера	13	5	1	4	8
5	Метрики качества регрессии и бинарной классификации	22	8	2	6	14
6	Выделение данных для тестирования	14	6	2	4	8
7	Линейная многоклассовая классификация. Метрики качества многоклассовой классификации	19	7	1	6	12
8	Дисбаланс классов	15	5	1	4	10
9	Полиномиальные модели	9	3	1	2	6
10	Недообучение и переобучение	33	13	3	10	20
11	Основные модели классического машинного обучения	72	28	8	20	44
12	Предварительный анализ и обработка данных	72	26	4	22	46
13	Задачи обучения без учителя	42	16	4	12	26
	Итого	360	136	32	104	224

5.3. Содержание семинаров

Наименование тем (разделов) дисциплины	Перечень вопросов для обсуждения на семинарах, практических занятиях	Формы проведения занятий
Основные понятия и инструменты машинного обучения	Обзор библиотек: numpy, pandas. Обзор библиотеки Matplotlib для построения 2D и 3D-графиков. Обзор библиотеки Seaborn. Реализация метода градиентного спуска своими руками.	Решение и обсуждение задач
Линейная регрессия	Построение модели линейной регрессии своими руками с использованием аналитического решения и методом градиентного спуска. Построение модели библиотечными средствами.	Решение и обсуждение задач
Линейная бинарная классификация	Реализация модели логистической регрессии методом градиентного спуска своими руками и библиотечными средствами.	Решение и обсуждение задач
Масштабирование признаков. Применение конвейера	Масштабирование данных методами стандартной и min-max нормализации. Применение конвейера. Значимость признаков.	Решение и обсуждение задач
Метрики качества регрессии и бинарной классификации	Вычисление основных метрик качества задачи регрессии (коэффициент детерминации, RMSE, MAE и прочие) своими руками и с использованием библиотечных функций для модели линейной регрессии. Вычисление основных метрик качества задачи бинарной классификации (матрица ошибок, Accuracy, Precision, Recall, F1-score) своими руками и с использованием библиотечных функций для модели логистической регрессии с двумя классами. Построение и интерпретация ROC-кривой. Вычисление метрики ROC-AUC.	Решение и обсуждение задач
Выделение данных для тестирования	Выделение тестовой выборки методом деления на две части (hold-out). Обучение модели на обучающих данных, а предсказание на тестовых данных.	Решение и обсуждение задач
Линейная многоклассовая классификация. Метрики качества многоклассовой классификации	Реализация двух подходов в задаче многоклассовой классификации: One-vs-Rest и Softmax с использованием библиотечных функций. Построение матрицы ошибок многоклассовой классификации. Нахождение метрик качества многоклассовой классификации: Precision, Recall, F1 для каждого класса. Усреднение метрик: macro, micro, weighted. ROC-анализ для многоклассовой классификации. Определение по метрикам какой подход оказался	Решение и обсуждение задач

	эффективнее для конкретных данных.	
Дисбаланс классов	Осуществление борьбы с дисбалансом классов разными методами: undersampling, oversampling, добавлением синтетических данных SMOTE, методом на уровне алгоритма (Class Weighting). Сравнение методов по метрике ROC-AUC и выбор наилучшего для конкретных данных. Использование библиотеки imblearn.	Решение и обсуждение задач
Полиномиальные модели	Полиномиальные данные. Построение полиномиальных моделей регрессии и классификации, как линейных моделей на полиномиальных данных. Визуализация нелинейной линии регрессии в задаче регрессии и нелинейной границы разделения на классы в задаче классификации.	Решение и обсуждение задач
Недообучение и переобучение	Диагностика переобучения полиномиальных моделей регрессии и классификации. Построение валидационных кривых и кривых обучения. Борьба с переобучением подбором наилучшего гиперпараметра сложности модели. Регуляризация в задаче регрессии и классификации: Ridge, Lasso, Elastic-Net. Уменьшение степени переобученности модели путём выбора коэффициента регуляризации. Борьба с недообучением логистической регрессии путём подбора гиперпараметров модели методом поиска по сетке и рандомизированного поиска по сетке.	Решение и обсуждение задач
Основные модели классического машинного обучения	Модель k-ближайших соседей для задачи регрессии и классификации. Выбор числа соседей и метрики расстояния. Метод опорных векторов для задачи регрессии и классификации. Наивный байесовский классификатор. Различные типы: Gaussian, Multinomial, Bernoulli. Дерево решений для задачи регрессии и классификации. Модель случайного леса для задачи регрессии и классификации. Модель градиентного бустинга.	Решение и обсуждение задач
Предварительный анализ и обработка данных	Разведочный анализ данных (EDA). Импутация пропусков одномерными методами. Использование класса SimpleImputer. Импутация пропусков многомерными методами. Использование класса KNNImputer. Обработка выбросов (Outliers). Методы обнаружения: IQR, Z-score, Isolation Forest, LOF. Стратегии борьбы: удаление, замена. Кодирование категориальных переменных. Порядковые и номинальные признаки. Создание новых признаков (Feature Engineering). Математические преобразования. Бинаризация и дискретизация. Отбор признаков	Решение и обсуждение задач

	(Feature Selection). Удаление признаков с низкой дисперсией (VarianceThreshold). Статистические тесты (SelectKBest). Удаление мультиколлинеарных признаков. Важность признаков в деревьях. Комплексная предобработка данных с использованием Pipeline. Использование класса ColumnTransformer.	
Задачи обучения без учителя	Кластеризация методом k-средних (K-Means). Выбор числа кластеров: метод локтя. Метрика силуэт. Плотностная кластеризация (DBSCAN). Иерархическая кластеризация. Дендрограмма. Задача снижения размерности. Метод главных компонент (PCA). Выбор числа компонент.	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
Основные понятия и инструменты машинного обучения	Знакомство с понятием обучение с подкреплением (reinforcement).	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Линейная регрессия	Реализация модели линейной регрессии методом стохастического градиентного спуска и мини-батчи.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Линейная бинарная классификация	Реализация модели логистической регрессии методом стохастического градиентного спуска и мини-батчи.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Масштабирование признаков. Применение конвейера	Отработка изученного материала на других датасетах. Изучение способов масштабирования: RobustScaler, MaxAbsScaler, Normalizer.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Метрики качества регрессии и бинарной классификации	Отработка изученного материала на других датасетах. Изучение дополнительных метрик: Adjusted R ² для моделей регрессии с разным числом признаков, F-beta мера для разного баланса precision/recall в задаче классификации. Precision-Recall кривая и AUC-PR.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Выделение данных для тестирования	Отработка изученного материала на других датасетах. Изучение способов разбиения данных Repeated K-Fold, GroupKFold.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.

Линейная многоклассовая классификация. Метрики качества многоклассовой классификации	Отработка изученного материала на других датасетах. Знакомство с понятием многозначной классификации (multi-label).	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Дисбаланс классов	Продвинутые методы undersampling: TomekLinks.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Полиномиальные модели	Отработка изученного материала на других датасетах. Сплайны как альтернатива полиномам.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Недообучение и переобучение	Отработка изученного материала на других датасетах. Ранняя остановка (early stopping) при борьбе с переобучением.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Основные модели классического машинного обучения	Отработка изученного материала на других датасетах. Стекинг (stacking) и блендинг моделей.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Предварительный анализ и обработка данных	Отработка изученного материала на других датасетах. Отбор признаков при помощи L1-регуляризации.	Работа с учебной литературой. Разбор вопросов по теме занятия. Выполнение домашних заданий к каждому занятию.
Задачи обучения без учителя	Отработка изученного материала на других датасетах. Детекция аномалий: One-Class SVM.	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 контрольной работы

1. Загрузите предложенный вам датасет из библиотеки `sklearn.datasets.fetch_openml`. Выведите описание загруженного датасета. Обозначьте матрицу признаков за X , а целевую переменную за y . Выведите их размеры. Убедитесь, что в данных нет пропусков и все признаки числовые. Сделайте значения целевой переменной числовыми. Выведите количество объектов каждого класса.
2. Разделите данные на обучающую и тестовую части с перемешиванием, выделив в тестовую часть 25% объектов, а также зафиксируйте `random_state=0` для воспроизводимости разбиения.
3. Выполните масштабирование признаков методом `min-max` нормализации.
4. Постройте на масштабированных данных модель логистической регрессии с параметрами по умолчанию из библиотеки `sklearn`. Выведите полученные коэффициенты гиперплоскости.
5. Сделайте предсказание на масштабированных тестовых данных для модели из библиотеки `sklearn`. Вычислите метрики: `accuracy` и `F1`-меру (позитивным классом считайте единицу). Сделайте вывод о качестве работы модели логистической регрессии для этих данных.
6. Постройте полиномиальную модель второй степени для задачи классификации, используя модель логистической регрессии с параметрами по умолчанию из библиотеки `sklearn`. Вычислите метрики качества: `accuracy` и `F1`-меру (позитивным классом считайте единицу). Сделайте вывод о том, какая из моделей линейная или полиномиальная описывает данные лучше.

Примерный вариант проектной работы

(Проектная работа выполняется в Jupyter Notebook с комментариями к коду)

Тема: Комплексная предобработка данных

Инструменты: Python, pandas, sklearn, matplotlib

Датасеты: Adult Income (классификация) или Ames Housing (регрессия) или свой на выбор

Задание 1. Разведочный анализ (EDA)

- Описательные статистики (mean, std, квантили, частоты)
- Визуализация: гистограммы, boxplot, barplot
- Анализ пропусков (missingno) + гипотезы о природе
- Корреляционная матрица, выявление мультиколлинеарности

Задание 2. Обработка пропусков

Выбрать стратегию для каждого признака (удаление / mean / median / mode / KNN). Обосновать выбор.

Задание 3. Обработка выбросов

- Обнаружение: IQR, Z-score, Isolation Forest (минимум 2 метода)
- Стратегия: удаление / winsorization / log-преобразование / сохранение

Задание 4. Кодирование категорий

Label Encoding (порядковые), One-Hot (низкая кардинальность), Target/Frequency (высокая кардинальность). Обосновать выбор.

Задание 5. Feature Engineering

Создать минимум 2 новых признака: математические преобразования, агрегации, бинаризация, взаимодействия.

Задание 6. Отбор признаков

Применить минимум 2 метода: SelectKBest, RFE, Feature Importance, VarianceThreshold.

Задание 7. Pipeline

Собрать весь процесс в sklearn.pipeline.Pipeline + ColumnTransformer.

Задание 8. Сравнение результатов

- Baseline: модель на сырых данных (минимальная предобработка)

- После обработки: модель после Pipeline
- Метрики: Accuracy, F1, ROC-AUC (классификация) или MAE, RMSE, R^2 (регрессия)

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

Дополнительная информация

Примерный вариант тестирования

Вопрос 1. Какая функция используется в логистической регрессии для преобразования выхода линейной модели в вероятность?

- A) ReLU (Rectified Linear Unit)
- B) Softmax
- C) Сигмоида (логистическая функция)
- D) Гиперболический тангенс (tanh)
- E) Ступенчатая функция Хевисайда

Вопрос 2. Какой метод масштабирования признаков преобразует данные так, чтобы среднее значение стало равно 0, а стандартное отклонение — 1?

- A) MinMaxScaler
- B) StandardScaler
- C) RobustScaler
- D) Normalizer
- E) MaxAbsScaler

Вопрос 3. Какая метрика качества классификации рассчитывается как среднее гармоническое между Precision и Recall?

- A) Accuracy

- B) Specificity
- C) ROC-AUC
- D) F1-score
- E) Log Loss

Вопрос 4. Какой метод кросс-валидации рекомендуется использовать для задачи классификации с несбалансированными классами?

- A) K-Fold
- B) Leave-One-Out (LOO)
- C) Stratified K-Fold
- D) Hold-Out
- E) Shuffle Split

Вопрос 5. Что из перечисленного является методом борьбы с переобучением модели?

- A) Увеличение степени полинома
- B) Удаление регуляризации
- C) L2-регуляризация (Ridge)
- D) Уменьшение объёма обучающей выборки
- E) Увеличение числа признаков без отбора

Вопрос 6. Какую основную идею реализует алгоритм градиентного бустинга?

- A) Обучение деревьев на случайных подвыборках данных параллельно
- B) Последовательное обучение моделей на остатках (ошибках) предыдущих моделей
- C) Усреднение предсказаний независимых деревьев максимальной глубины
- D) Выбор лучшей модели из ансамбля по валидационной выборке
- E) Случайный выбор признаков для каждого узла дерева

Вопрос 7. Какое ключевое предположение делает наивный байесовский классификатор?

- A) Все признаки имеют нормальное распределение
- B) Признаки условно независимы при заданном классе
- C) Число признаков должно быть меньше числа примеров
- D) Классы должны быть сбалансированы
- E) Признаки должны быть масштабированы

Вопрос 8. Какой тип пропусков (missing data) возникает, когда вероятность пропуска зависит от самого пропущенного значения?

- A) MCAR (Missing Completely At Random)
- B) MAR (Missing At Random)
- C) MNAR (Missing Not At Random)
- D) MSAR (Missing Systematically At Random)
- E) MRD (Missing by Random Distribution)

Вопрос 9. Какой метод обнаружения выбросов основан на построении случайных деревьев и измерении длины пути изоляции точки?

- A) Z-score
- B) IQR (межквартильный размах)
- C) Local Outlier Factor (LOF)
- D) Isolation Forest
- E) DBSCAN

Вопрос 10. На что проецирует данные метод главных компонент (PCA)?

- A) На направления минимальной дисперсии
- B) На направления максимальной дисперсии (главные компоненты)
- C) На центроиды кластеров

D) На случайные ортогональные направления

E) На направления максимальной корреляции с целевой переменной

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

ОПК-1 Способен применять естественнонаучные и общетехнические знания, методы математического анализа и моделирования, теоретического и экспериментального исследования в профессиональной деятельности

1) Демонстрирует знания основных методов математического анализа и моделирования, применяет их на практике для решения задач профессиональной деятельности

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

Знать: основные понятия машинного обучения и интеллектуальной обработки данных, их области применения и сравнительные характеристики

Уметь: применять основные методы обучения с учителем и без учителя для решения типовых и практико-ориентированных задач машинного обучения и анализа данных

Типовые контрольные задания

Метод градиентного спуска для парной линейной регрессии.

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.

2) Проводит теоретические исследования по выбранной области профессиональной деятельности

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

Знать: основы машинного обучения и математики, связанной с этой областью

Уметь: создавать и настраивать модели машинного обучения. Анализировать данные и применять их в моделях. Критически оценивать методы машинного обучения. Эффективно коммуницировать результаты своих исследований

Типовые контрольные задания

Предоставлен набор данных о стоимости недвижимости.

Задача - разработать и обучить модель машинного обучения для прогнозирования цен на недвижимость на основе различных атрибутов, таких как площадь, количество комнат, район и др.

Для этого необходимо выполнить следующие шаги:

1. Подготовка данных
2. Разработка модели
3. Оценка модели
4. Настройка модели
5. Визуализация данных и результатов
6. Отчет и коммуникация.

3) Проводит численные эксперименты на основе математических или информационных методов, делает выводы и обосновывает их

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

Знать: основы машинного обучения и статистики; методы и модели машинного обучения

Уметь: планировать и проводить численные эксперименты; анализировать данные и извлекать выводы из численных экспериментов; обосновывать полученные результаты на основе математических и статистических методов; эффективно коммуницировать результаты своих исследований;

Типовые контрольные задания

Предоставлен набор данных о стоимости недвижимости.

Задача - разработать и обучить модель машинного обучения для прогнозирования цен на недвижимость на основе различных атрибутов, таких как площадь, количество комнат, район и др.

Для этого необходимо выполнить следующие шаги:

1. Подготовка данных
2. Разработка модели
3. Оценка модели
4. Настройка модели
5. Визуализация данных и результатов
6. Отчет и коммуникация.

Примеры практико-ориентированных заданий

Вариант 1. Загрузите встроенный набор данных Iris. Разделите данные на обучающую и тестовую выборки в соотношении 70/30. Обучите три модели многоклассовой классификации: логистическую регрессию с Softmax, метод k ближайших соседей (k-NN) и дерево решений. Рассчитайте macro-F1 для каждой модели и сделайте вывод о том, какая из них оказалась эффективнее.

Вариант 2. Загрузите встроенный набор данных Wine о физико-химических свойствах вин. Разделите данные на обучающую и тестовую выборки в соотношении 80/20. Обучите модель случайного леса и градиентного бустинга для предсказания качества вина. Извлеките и визуализируйте важность признаков для обеих моделей в виде горизонтальных столбчатых диаграмм. Сравните ранжирование признаков по важности между моделями и сделайте вывод о ключевых факторах, влияющих на качество вина.

Вариант 3. Вам предоставлен подготовленный набор данных с признаками разных масштабов. Разделите данные на обучающую и тестовую выборки в соотношении 80/20. Обучите модель k ближайших соседей (k-NN) на трёх вариантах данных: без масштабирования, с StandardScaler и с MinMaxScaler. Для каждого варианта выполните 5-fold кросс-валидацию и рассчитайте среднее значение Ассигасу и его стандартное отклонение. Визуализируйте результаты и объясните влияние масштабирования на качество модели k-NN.

Примерные вопросы для подготовки к экзамену, зачету

Примерные вопросы для подготовки к зачету

Тема 1. Основные понятия и инструменты машинного обучения

1. Определение машинного обучения, его место в структуре искусственного интеллекта и отличие от классического программирования
2. Метод градиентного спуска: принцип работы, скорость обучения, условия сходимости
3. Типы задач машинного обучения
4. Формальная постановка задачи обучения с учителем
5. Этапы построения проекта по машинному обучению
6. Для чего используются библиотеки Python: NumPy, Pandas, Matplotlib, Seaborn, Scikit-learn

Тема 2. Линейная регрессия

7. Постановка задачи регрессии и модель линейной регрессии
8. Функция потерь MSE и аналитическое решение методом наименьших квадратов (нормальное уравнение)
9. Реализация линейной регрессии методом градиентного спуска

Тема 3. Линейная бинарная классификация

10. Постановка задачи бинарной классификации и сведение к задаче регрессии
11. Сигмоидная функция: определение, свойства, применение
12. Модель логистической регрессии и функция потерь (бинарная кросс-энтропия)
13. Порог принятия решения и преобразование вероятности в метку класса

Тема 4. Масштабирование признаков

14. Проблема разных масштабов признаков и влияние на сходимость градиентного спуска
15. Методы масштабирования StandardScaler и MinMaxScaler: принципы работы и критерии выбора
16. Правильное применение масштабирования без утечки данных (data leakage)
17. Значимость признаков в линейных моделях после масштабирования

Тема 5. Метрики качества регрессии и бинарной классификации

- 18. Коэффициент детерминации R^2 : определение, диапазон значений, недостатки
- 19. Метрики ошибок регрессии: MAE, MSE, RMSE, MAPE
- 20. Матрица ошибок в бинарной классификации (TP, TN, FP, FN) и ошибки первого и второго рода
- 21. Метрики Accuracy, Precision, Recall, Specificity: определения и интерпретация
- 22. F1-score как гармоническое среднее Precision и Recall
- 23. ROC-анализ: построение ROC-кривой и метрика ROC-AUC
- 24. Precision-Recall анализ и PR-кривая

Тема 6. Выделение данных для тестирования

- 25. Обучающая способность модели и проблема оценки качества на обучающих данных
- 26. Отложенная выборка (hold-out): принцип и ограничения
- 27. Кросс-валидация: принцип и преимущества перед отложенной выборкой
- 28. Типы разбиений для кросс-валидации: K-Fold, Stratified K-Fold, Leave-One-Out

Тема 7. Линейная многоклассовая классификация

- 29. Стратегии One-vs-Rest (OvR) и One-vs-One (OvO) для многоклассовой классификации
- 30. Многоклассовая логистическая регрессия, Softmax-функция и кросс-энтропия
- 31. Матрица ошибок многоклассовой классификации
- 32. Метрики Precision, Recall, F1 для многоклассовой задачи и способы усреднения (macro, micro, weighted)
- 33. ROC-анализ для многоклассовой классификации: micro-average и macro-average

Тема 8. Дисбаланс классов

- 34. Проблема дисбаланса классов: диагностика и степени дисбаланса
- 35. Выбор правильных метрик при дисбалансе классов

36. Методы борьбы с дисбалансом на уровне данных: Undersampling, Oversampling, SMOTE

37. Методы борьбы с дисбалансом на уровне алгоритма: взвешивание классов, настройка порога

Тема 9. Полиномиальные модели

38. Построение полиномиальных признаков

39. Полиномиальная модель

Тема 10. Недообучение и переобучение

40. Недообучение (Underfitting) и переобучение (Overfitting): определения, признаки, причины

41. Компромисс bias-variance при выборе сложности модели

42. Валидационные кривые и кривые обучения: построение и интерпретация

43. L2-регуляризация (Ridge)

44. L1-регуляризация (Lasso)

45. Комбинация L1 и L2 регуляризации (Elastic-Net)

46. Проблема мультиколлинеарности и методы её решения

47. Методы борьбы с переобучением модели

48. Методы борьбы с недообучением модели

49. Различие между параметрами и гиперпараметрами модели

50. Поиск по сетке (Grid Search) и рандомизированный поиск (Random Search)

Примерные вопросы для подготовки к экзамену

Тема 11. Основные модели классического машинного обучения

1. Метод k ближайших соседей (k-NN): принцип классификации и выбор гиперпараметров

2. Проклятие размерности и его влияние на метод k-NN

3. Метод опорных векторов (SVM): принцип максимизации зазора и параметр регуляризации C

4. Ядровый трюк в SVM и типы ядер (линейное, полиномиальное, RBF), параметр γ
5. Наивный байесовский классификатор: теорема Байеса и виды (Gaussian, Multinomial, Bernoulli)
6. Деревья решений: структура, алгоритм построения, критерии разбиения (Gini, Entropy, MSE)
7. Визуализация дерева решений и важность признаков
8. Ансамблевые методы: общая идея объединения моделей
9. Bagging и bootstrap-агрегирование
10. Случайный лес (Random Forest): принцип работы и out-of-bag оценка
11. Boosting: общий принцип последовательного построения ансамбля
12. Градиентный бустинг: алгоритм и оптимизированные реализации (XGBoost, LightGBM, CatBoost)
13. Выбор наилучшей модели

Тема 12. Предварительный анализ и обработка данных

14. Разведочный анализ данных (EDA): цели, этапы, описательная статистика, визуализация
15. Типы пропущенных значений: MCAR, MAR, MNAR
16. Методы работы с пропусками: удаление и импутация (одномерная, многомерная)
17. Label Encoding для порядковых и One-Hot Encoding для номинальных признаков
18. Обработка высокоразмерных категориальных признаков
19. Типы выбросов и методы обнаружения (IQR, Z-score, Isolation Forest, Local Outlier Factor)
20. Стратегии обработки выбросов: удаление, обрезка, замена, трансформация
21. Создание новых признаков (Feature Engineering): математические преобразования, бинаризация, дискретизация
22. Отбор признаков (Feature Selection): фильтрующие, обёрточные и встроенные методы

Тема 13. Основные задачи обучения без учителя

- 23. Введение в обучение без учителя: типы задач и области применения
- 24. Кластеризация методом k-средних (K-Means): алгоритм, инициализация центроидов, выбор числа кластеров
- 25. Плотностная кластеризация DBSCAN: три типа точек и параметры модели
- 26. Иерархическая кластеризация и построение дендрограмм
- 27. Метод главных компонент (PCA): принцип и выбор числа компонент

Пример экзаменационного билета

- 1. Представление графической информации в моделях машинного обучения. (20 баллов)
- 2. Метрики эффективности и функции ошибки: назначение, примеры, различия. (20 баллов)
- 3. Вам предоставлен датасет с "чистыми" данными. Разделите данные на обучающую и тестовую выборки в соотношении 80/20. Обучите три модели классификации с параметрами по умолчанию: логистическую регрессию, случайный лес (Random Forest) и градиентный бустинг. Проведите ROC-анализ: постройте ROC-кривые всех моделей на одном графике, рассчитайте ROC-AUC для каждой модели. Сравните результаты и сделайте вывод о лучшей модели. (20 баллов)

Практическая часть выполняется в Jupyter Notebook с комментариями к коду.

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

Основная литература:

1. Коротеев, М.В. Основы машинного обучения на Python : Учебник / М.В. Коротеев Электрон. дан. Москва : КноРус, 2025 431 с. Режим доступа: book.ru Internet access <https://book.ru/book/957785> ISBN 978-5-406-14728-3. [БИК ID: RU\bookru\bibl\957785]
2. Одинцова, В.А. Сборник лабораторных работ по дисциплине "Машинное обучение" : для студентов обучающихся по направлениям: "Прикладная информатика", "Программная инженерия" всех профилей (программы подготовки бакалавров) / В.А. Одинцова Москва : Финуниверситет, 2025 1 файл (3,62 Мб) : ил. Свободный доступ из сети Интернет (чтение) <http://elib.fa.ru/fbook/books144393.pdf> (дата обращения 26.06.2025) ISBN 978-5-7942-1458-1 + Текст : электронный. [БИК ID: RU\FA\books\144393]

Дополнительная литература:

1. Златопольский, Д.М. Программирование: типовые задачи алгоритмы методы — 5-е изд. : Не указано / Д.М. Златопольский Электрон. дан. Москва : Лаборатория знаний, 2025 225 с. Режим доступа: book.ru Internet access <https://book.ru/book/957129> ISBN 978-5-93208-832-6. [БИК ID: RU\bookru\bibl\957129]

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

1. Электронная библиотека Финансового университета (ЭБ) <http://elib.fa.ru/> (<http://library.fa.ru/files/elibfa.pdf>)
2. Электронно-библиотечная система BOOK.RU <http://www.book.ru>
3. Электронно-библиотечная система "Университетская библиотека ОНЛАЙН" <http://biblioclub.ru/>
4. Образовательная платформа "ЮРАЙТ" <https://urait.ru/>
5. Информационно-образовательный портал Финуниверситета: <https://org.fa.ru>
6. Python Documentation <http://python.org/doc/>

7. 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.

При подготовке к семинарским занятиям необходимо активно работать с документацией библиотеки Scikit-learn. Файлы с выполненными заданиями необходимо хранить в личной сетевой папке в компьютерной сети вуза.

При подготовке к контрольной и проектной работе следует использовать нормативные документы Финансового университета, Методические рекомендации по планированию и организации внеаудиторной самостоятельной работы студентов по образовательным программам бакалавриата и магистратуры в Финансовом университете, утвержденные приказом Финуниверситета от 11.05.2021 г. № 1040/о (см. сайт Финансового Университета: на главной странице раздел "Наш университет" → "Единая правовая база Финуниверситета" → "Организация учебного процесса" → "Нормативные документы по самостоятельной работе").

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. Jupyter Notebook (Windows 10)
3. Kaspersky
4. Astra Linux

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. **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)

2. **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)