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: uomu5zez8UcAMlk5hCKznxAmdRLabNt6
Дата: 25.03.2025 г.

В. И. Шин, В. О. Миронов

# Построение и оценка моделей машинного обучения

**Рабочая программа дисциплины**

для студентов, обучающихся по направлению подготовки
01.04.02 - Прикладная математика и информатика

Направленность программы магистратуры
Инженер машинного обучения

*Рекомендовано*
*Факультет информационных технологий и анализа больших данных*
*(протокол №56 от 17.06.2025)*

*Одобрено*
*Кафедра искусственного интеллекта*
*(протокол №11 от 27.05.2025)*

**Москва 2025**

# Содержание


| **Наименование разделов РПД** | | **Стр.** |
|---|---|---|
| 1. | Наименование дисциплины | 5 |
| 2. | Перечень планируемых результатов освоения образовательной программы (перечень компетенций) с указанием индикаторов их достижения и планируемых результатов обучения по дисциплине | 5 |
| 3. | Место дисциплины в структуре образовательной программы | 10 |
| 4. | Объем дисциплины(модуля) в зачетных единицах и в академических часах с выделением объема аудиторной (лекции, семинары) и самостоятельной работы обучающихся | 10 |
| 5. | Содержание дисциплины, структурированное по темам (разделам) дисциплины с указанием их объемов (в академических часах) и видов учебных занятий | 11 |
| 5.1. | Содержание дисциплины | 11 |
| 5.2. | Учебно-тематический план | 19 |
| 5.3. | Содержание семинаров, практических занятий | 20 |
| 6. | Перечень учебно-методического обеспечения для | 32 |

| | | |
|---|---|---|
| | самостоятельной работы обучающихся по дисциплине | |
| 6.1. | Перечень вопросов, отводимых на самостоятельное освоение дисциплины, формы внеаудиторной самостоятельной работы | 32 |
| 6.2. | Перечень вопросов, заданий, тем для подготовки к текущему контролю | 37 |
| 7. | Фонд оценочных средств для проведения промежуточной аттестации обучающихся по дисциплине | 41 |
| 8. | Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины | 47 |
| 9. | Перечень ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины | 47 |
| 10. | Методические указания для обучающихся по освоению дисциплины | 48 |
| 11. | Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине, включая перечень необходимого программного обеспечения и информационных | 52 |

| | | |
|---|---|---|
| | справочных систем | |
| 12. | Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине | 52 |

## 1. Наименование дисциплины

«Построение и оценка моделей машинного обучения».

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

| **Код компетенции** | **Наименование компетенции** | **Индикатор достижения компетенции** | **Результаты обучения** |
|---|---|---|---|
| ПК-2 | Способность разрабатывать прикладные решения с использованием методов компьютерного зрения | Использовать основные модели и методы компьютерной обработки изображений и распознавания образов, модели и методы компьютерного зрения | **Знать:** основные алгоритмы и методы, используемые в компьютерной обработке изображений, такие как фильтрация, сегментация, детекция границ, преобразование Фурье и др. **Уметь:** выбирать и применять соответствующие модели и методы компьютерного зрения для решения конкретных задач, таких как распознавание лиц, анализ медицинских изображений или автоматизация процессов в производстве |
| | | Решать с | **Знать:** |

| | | | |
|---|---|---|---|
| | | помощью моделей и методов компьютерного зрения прикладные задачи предметной области | основные алгоритмы распознавания образов, такие как методы машинного обучения (например, SVM, нейронные сети), а также современные подходы, такие как глубокое обучение (например, сверточные нейронные сети)<br>**Уметь:** использовать программные библиотеки и фреймворки (например, OpenCV, TensorFlow, PyTorch) для разработки прикладных решений в области компьютерного зрения |
| ПК-3 | Способность разрабатывать прикладные решения с использованием методов анализа текстов, графов и сетевых структур | Использовать основные модели и методы обработки текстов на естественных языках, графов и семантических структур | **Знать:** основы Natural Language Processing (NLP), включая токенизацию, стемминг, лемматизацию, векторизацию текстов, выделение сущностей и отношений, классификацию и тематическое моделирование<br>**Уметь:** проектировать |

| | | | |
|---|---|---|---|
| | | | архитектуры решений, комбинирующие анализ текстов и графов, реализовывать алгоритмы на языках программирования вроде Python, JavaScript или R, и эффективно интегрировать их в существующие информационные системы |
| | | Решать с помощью прикладных моделей NLP, графовой и сетевой аналитики прикладных задач предметной области | **Знать:** основы методов NLP ( токенизация, лемматизация, удаление стоп-слов, очистка данных);<br><br>векторизацию текстов (TF-IDF, Word2Vec, BERT и другие эмбеддинги);<br><br>машинное обучение и глубокое обучение для задач классификации, сентимент-анализа, перевода и др;<br><br>основные концепции графовой аналитики и сетевого анализа: графы и способы их представления (списки |

| | | | |
|---|---|---|---|
| | | | смежности, матрица смежности);<br><br>центральность узлов (Degree Centrality, Betweenness Centrality, Eigenvector Centrality);<br><br>выделение сообществ и кластеров (Louvain Method, Girvan-Newman Algorithm);<br><br>библиотеки и инструменты для реализации: NLP: SpaCy, NLTK, Gensim, Transformers (HuggingFace);<br><br>графовую аналитику: NetworkX, igraph, Gephi, Neo4j;<br><br>применение этих технологий в рамках своей специальности (экономика, медицина, маркетинг и т.п.).<br>**Уметь:**<br>разрабатывать и настраивать NLP-модели для решения прикладных задач (например, классификация документов, |

| | | | |
|---|---|---|---|
| | | | анализ тональности отзывов, поиск похожих текстов);<br><br>строить и анализировать графы для представления сложных данных и связей (социальные сети, транзакционные данные, библиометрические показатели);<br><br>использовать готовые библиотеки и API для быстрого прототипирования и тестирования решений;<br><br>формулировать задачи прикладного характера, адекватно выбирать инструментарий и создавать эффективные решения для бизнес-задач или научных проектов;<br><br>визуализировать результаты графового анализа и представлять их визуально привлекательно и понятно; |

| | | | |
|---|---|---|---|
| | | | интегрировать методы NLP и графового анализа в IT-инфраструктуру компаний или лабораторий для автоматизации процессов принятия решений; |

## 3. Место дисциплины в структуре образовательной программы

Дисциплина «Построение и оценка моделей машинного обучения» относится к «Модулю направленности программы магистратуры»

## 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)** | **Модуль 2 (в часах)** |
|---|---|---|
| **Общая трудоемкость дисциплины** | **4/144** | **144** |
| ***Контактная работа - Аудиторные занятия*** | **40** | **40** |
| *Лекции* | 10 | 10 |
| *Семинары, практические занятия* | 30 | 30 |
| ***Самостоятельная работа*** | 104 | 104 |
| Вид текущего контроля | Контрольная работа; | Контрольная работа |
| Вид промежуточной аттестации | Экзамен; | 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

Введение в машинное обучение. Искусственный интеллект, машинное обучение и глубокое обучение. История анализа данных. Основные понятия: обучение с учителем, обучение без учителя, обучение с подкреплением, модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль, обучающая и тестовая выборка. Кросс-валидация. Основные типы данных: таблицы, тексты, изображения, звук. Современные достижения.

Постановки задач в машинном обучении: классификация, регрессия, ранжирование, кластеризация, прогнозирование, сокращение размерности, обработка естественного языка (NLP), компьютерное зрение (CV), генеративные модели и рекомендательные системы, работа с временными рядами и анализ временных сигналов, этические и правовые аспекты машинного обучения, практическое внедрение моделей в производство (ML Ops), гиперпараметры и оптимизация гиперпараметров, основы параллельных вычислений и распределённых систем

Примеры прикладных задач.

• Классификация: Автоматическая диагностика заболеваний легких на рентгеновских изображениях пациентов. Здесь задача состоит в классификации каждого изображения либо как здорового, либо как пораженного заболеванием (например, пневмонией). Это пример задачи компьютерного зрения, решаемой методами глубокого обучения.

• Регрессия: Прогнозирование стоимости недвижимости исходя из характеристик жилья (площади, района, количества комнат и прочих факторов). Задача заключается в построении модели, способной точно оценить стоимость дома или квартиры на рынке. Типичный случай применения методов регрессии.

• Ранжирование: Создание эффективной системы рекомендаций фильмов на платформе потокового вещания. Модель должна правильно сортировать фильмы, предлагая пользователям наиболее релевантные именно для них варианты. Эта задача решается с помощью подходов рекомендательных систем.

• Кластеризация: Анализ покупательского поведения клиентов магазина для формирования маркетинговых стратегий. Модели группируют покупателей по схожести покупок, выявляя различные сегменты аудитории (например, любители сладостей, приверженцы здоровой пищи и др.). Этот подход относится к задаче безучёбного обучения и используется для выявления скрытых закономерностей в данных.

• Прогнозирование временных рядов: Предсказание загруженности дорог города в течение суток. Используются исторические данные о траффике, погоде и событиях для построения точной модели прогнозирования уровня пробок в городе. Такой вид задач встречается в логистике, городском управлении и транспортной аналитике, и относится к обработке временных рядов.

• Обработка естественного языка (NLP): Автоматический анализ отзывов клиентов на товары или услуги для определения общей удовлетворённости потребителей и выделения ключевых проблем. Используется классификация настроений (sentiment analysis), позволяющая бизнесу быстро реагировать на обратную связь и улучшать качество продуктов.

• Генеративные модели: Генерация реалистичных фотографий несуществующих людей для нужд дизайна интерфейсов, рекламных кампаний или тестирования систем безопасности. Такие модели создаются с применением технологий Generative Adversarial Networks (GANs) и широко используются в креативных индустриях.

• Компьютерное зрение (CV): Система автоматического распознавания автомобильных номеров для контроля дорожного движения и автоматической оплаты парковки. Применяются методы сверточных нейронных сетей (Convolutional Neural Networks, CNN), позволяющие эффективно обрабатывать визуальные образы и извлекать необходимую информацию.

• Рекомендательные системы: Персонализированные рекомендации книг на платформах электронной коммерции, основанные на истории покупок и предпочтениях конкретного пользователя. Такие системы позволяют увеличить продажи и повысить уровень лояльности клиентов.

• ML Ops (DevOps для машинного обучения): Построение конвейера автоматизированного обновления и развёртывания моделей машинного обучения для прогнозирования спроса на продукцию в крупной розничной сети. Ключевое преимущество — снижение человеческого вмешательства и ускоренное обновление моделей при изменении рыночных условий.

• Коллаборативная фильтрация: Платформа совместного прослушивания музыки рекомендует плейлисты друзьям пользователя на основе сходства предпочтений. Использование метода коллаборативной фильтрации позволяет создать удобную систему рекомендаций на основе вкусов сообщества пользователей.

## Тема 2. Подготовка и предварительный анализ данных

Табличные данные. Бинарные данные. Номинальные данные. Номинальные упорядоченные данные. Количественные данные. Трансформация данных. Борьба с пропусками. Источники данных. Импорт табличных данных в Pandas. Экспорт табличных данных в Pandas.

Описательная статистика. Корреляционный анализ. Фильтрация данных. Визуализация данных. Гистограммы. Линейные графики. Диаграммы рассеяния. Секторные диаграммы. Столбчатые диаграммы. Ящики с усами. Plotly. Matplotlib. Признаковое пространство. Целевая переменная.

Преобразование категориальных данных: One-Hot Encoding, Label Encoding, Target Encoding и бинаризация (Binarization).

Масштабирование и нормализация данных: MinMax Scaler, StandardScaler и RobustScaler.

Фичь-инженерия: Создание новых полезных признаков из исходных данных.

Балансировка классов: Oversampling (SMOTE), undersampling и комбинированные подходы.

Методы борьбы с выбросами: Z-score, IQR-метода, Boxplot-анализа и продвинутые методы, такие как Isolation Forest.

Подготовка временных рядов: Разложение временного ряда на компоненты (тренд, сезонность, остатки), агрегирование и сглаживание данных.

Дублирование и аномалии: Умение находить дублирующиеся записи и справляться с ними.Оценка корреляции между признаками: Методы оценки взаимосвязи признаков, помимо простой корреляционной матрицы Пирсона, такие как Спирмена и Кендалла.

**Тема 3. Модели регрессии**

Теорема Гаусса-Маркова. Метод наименьших квадратов. Модель линейной регрессии. Парная регрессия. Множественная регрессия. Параметрическая регрессия. Оценка качества модели: множественный и парный коэффициент корреляции, индекс детерминации. Оценка значимости регрессии. Средне-квадратическая ошибка RMSE. Средняя абсолютная относительная ошибка MAPE. Нелинейные модели регрессии. Линеаризирующие преобразования функций. Непараметрическая регрессия. Локально взвешенный метод наименьших квадратов. Робастная непараметрическая регрессия. Алгоритм LOWESS. Логарифмическая и степенная регрессия. Регуляризация. Полиномиальная регрессия. Деревья решений и ансамбли деревьев для регрессии. Машины опорных векторов (SVM) для регрессии. Байесовская регрессия. Метод Ridge для многомерной регрессии. Robust regression (Робастная регрессия). Метрики качества модели. Интерпретация результатов регрессии. Модели временной регрессии (Time Series Regression Models).

**Тема 4. Логистическая регрессия**

Предсказание вероятности принадлежности к классу. Логистическая функция или сигмоида. Переход от линейной регрессии к логистической. Понятие наблюдаемой вероятности и шансов. Функция потерь логистической регрессии. Градиентный спуск. Пакетный градиентный спуск. Стохастический градиентный спуск. Мини-пакетный градиентный спуск. Label encoding. One-hot encoding. Логистическая регрессия в задаче классификации. Многоклассовая логистическая регрессия (Multinomial Logistic Regression). Проблема разреженности данных (Sparse data problem). Стратегии инициализации весов. Эксперимент с гиперпараметрами (Hyperparameter tuning). Переполнение экспоненты и численная устойчивость. Решение проблемы Overfitting и Underfitting. Оптимизационные алгоритмы второго порядка. Приложения логистической регрессии. Выбор порогового значения (Threshold selection). Логистическая регрессия с регуляризацией (Regularized logistic regression).

**Тема 5. Байесовский подход**

Наивный Байес. Байесовский классификатор. Параметрическое восстановление плотности распределения вероятностей. Непараметрическое восстановление плотности распределения вероятностей. Метод окна Парзена с постоянной и переменной шириной окна. Ядерное сглаживание. Оценка Надарая-Ватсона с постоянной и переменной шириной окна. Выбор функции ядра. Проблема выбросов. Задача отсева выбросов. Мультиколлинеарность признаков. Методы устранения мульт-коллинеарности. Регуляризация корреляционной матрицы. Типичные случаи применения Naive Bayes. Графовые модели Байеса (Bayesian Graphical Models). Параметрические vs. Непараметрические методы. Сравнение методов Parzen Window и Kernel Density Estimation (KDE). Интуитивное объяснение эффекта оконного сглаживания. Методы выбора функции ядра. Диагностические тесты на выявление мультиколлинеарности. Регулярные матричные операции. Отсев выбросов (Outlier detection techniques). Алгоритмы очистки данных (Data Cleaning Algorithms). Продвинутые регулярные схемы (Advanced regularization schemes).

**Тема 6. Метрические алгоритмы**

Обобщённый метрический классификатор. Метод ближайших соседей kNN и его обобщения. Понятие выступа объекта. Алгоритм отбора эталонов. Проблема размерности. Выбор метрики. Обобщённый метрический регрессор. Метод потенциальных функций и его связь с линейной моделью классификации. Матрица ошибок, ошибка первого рода (ложно отрицательная), ошибка второго рода (ложно положительная). Критерии качества классификации: чувствительность и специфичность, ROC-кривая и AUC-кривая, доля верных распознаваний (Accuracy), точность (Precision) и полнота (Recall), F-мера, AUC-PR. Весовые метрики и адаптивные веса. Эффект локального соседства и лоскутная структура данных. Обучение метрики (Metric Learning). Автоматический отбор наилучшего значения k (k-optimization). Снижение размерности и визуализация данных. Отбор признаков (Feature Selection). Проблема класса миноритариев (Imbalanced classes). Совместимость с глубокими сетевыми архитектурами. Классификация с ограничениями (Constrained Classification). Расширение kNN на пространственно-временные данные.

## Тема 7. Методы сокращения размерности

Методы отбора признаков. Метод главных компонент. Собственное значение. Собственный вектор. Варимакс вращение. Факторные нагрузки. Многомерное шкалирование. Сингулярное разложение матриц. Сокращение размерности для визуализации. Методы стохастического выбора признаков (Stochastic Feature Selection). Отбор признаков на основе энтропии (Entropy-Based Feature Selection). Редукция ранга матрицы (Matrix Rank Reduction). Самоорганизующиеся карты (Self-Organizing Maps, SOM). Autoencoder и Variational Autoencoder (VAE). T-distributed Stochastic Neighbor Embedding (t-SNE). Uniform Manifold Approximation and Projection (UMAP). Метод ядра (Kernel Trick). Маргинальное управление признаковым пространством (Margin Control in Feature Space).

## Тема 8. Линейные алгоритмы классификации

Классификация линейной функцией. Линейная модель нейрона МакКаллока-Питтса. Градиентный метод численной минимизации. Персептрон Розенблатта. Дельта-правило ADALINE. Правило Хэбба.

Понятие выступа для линейного классификатора. Непрерывные функции потерь: кусочно-линейная (SVM), кусочно-линейная (Hebb's rule), логарифмическая, квадратичная, сигмоидная, экспоненциальная, пороговая функция потерь. Проблема переобучения. Валидационная выборка. Разные штрафы за ошибки. Критерий минимума среднего риска. Support Vector Machines (SVM). Softmax и Multinomial Logistic Regression. Logistic Loss Function. Binary Cross Entropy. ReLU Activation Function. Пороговая логика (Threshold Logic Unit). Cross Validation Techniques. Ensemble Methods for Linear Classifiers. Bias-Variance Tradeoff. Model Interpretability. Early Stopping. Curse of Dimensionality.

### Тема 9. Модели классификации Pandas и Scikit-Learn

Библиотека Scikit-Learn. Линейные классификаторы. Метод опорных векторов. Оптимальная разделяющая гиперплоскость. Понятие зазора между классами. Случаи линейной разделимости и отсутствия линейной разделимости. Классификаторы на базе деревьев решений. Энтропия. Глубина дерева. Алгоритм CART. Бинарная классификация. Мультиклассовая классификация. Мультиклассовая логистическая регрессия. Многослойный персептрон. Переход к ансамблевым методам (Ensemble Methods). Наборы данных (Datasets). Преимущества и ограничения моделей классификации. Параметры и гиперпараметры моделей. Оценка качества моделей. Особенности работы с Pandas. Visualization with Matplotlib & Seaborn. Pipeline и Workflows. Deep Dive into Evaluation Strategies.

### Тема 10. Композиции алгоритмов

Понятие композиции (ансамбля) алгоритмов. Теоретические предпосылки композиционных методов. Одиночные модели и ансамбли моделей. Простое голосование. Взвешенное голосование. Аппроксимация функции потерь: сигмоидная, логарифмическая, кусочно-линейная, экспоненциальная, квадратичная. Последовательные методы: бустинг-алгоритмы AdaBoost, AnyBoost, категориальный бустинг (CatBoost), градиентный бустинг (модели xgBoost, LightGBM). Стохастические методы построения композиций (параллельные методы комитетов): бутстреп, бэггинг, метод случайных подпространств. Случайный лес (Random Forest). Оптимизация моделей. Прунинг. Stacking (Метаклассификация).

Плюсы и минусы Ensembling. Staking vs Blending. Evaluation of Ensembles. Handling Imbalanced Classes. Sequential Boosting Variants. Deep Bagging and Deep Forests. Optimization in Ensemble Learning. Error Propagation in Ensembles. Interpretable Ensembles.

## Тема 11. Задача кластеризации

Постановка задачи кластеризации. Примеры прикладных задач. Типы кластерных структур. Оптимизационные постановки задач кластеризации и частичного обучения. Алгоритмы кластеризации: распространение сходства (Affinity Propagation), агломеративная (иерархическая) кластеризация (Agglomerative Clustering), сбалансированное итеративное сокращение и кластеризация с помощью иерархий (BIRCH), основанная на плотности пространственная кластеризация для приложений с шумами (DBSCAN), дивизивный алгоритм K-средних (K-means), мини-пакетный K-средних (Mini-Batch K-Means), сдвиг среднего значения (Mean Shift), упорядочение точек для обнаружения кластерной структуры (OPTICS), спектральная кластеризация (Spectral Clustering), модель гауссовой смеси (Mixture of Gaussians). EM-алгоритм. Построение дендрограммы. Оценка числа кластеров: метод локтя, коэффициент силуэта, критерий Акаике (AIC), Байесовский информационный критерий (BIC). Оценка качества кластеризации: индексы ARI и AMI, гомогенность, полнота, V-мера.

## Тема 12. Прогнозирование временных рядов

Понятие временного ряда. Тренд и сезонность. Авторегрессия (AR). Скользящее среднее (MA). Авторегрессионное скользящее среднее (ARMA). Авторегрессионное интегрированное скользящее среднее (ARIMA). Сезонная авторегрессия с интегрированным скользящим средним (SARIMA). Сезонное авторегрессионное интегрированное скользящее среднее с экзогенными регрессорами (SARIMAX). Векторная авторегрессия (VAR). Векторная авторегрессия скользящего среднего (VARMA). Скользящее среднее векторной авторегрессии с экзогенными регрессорами (VARMAX). Простое экспоненциальное сглаживание (SES). Экспоненциальное сглаживание Хольта-Винтерса (HWES). Подбор параметров моделей. Прогнозирование и оценка качества. Спектральный анализ временных рядов. Wavelet Transform (вейвлет-анализ).

Гетероскедастичность и архимедийская волатильность (ARCH/GARCH models). Хаффман-кодирование и сжатие временных рядов. Forecast Combination (Комбинация прогнозов). Anomaly Detection (Обнаружение аномалий). Predictive Maintenance (Прогнозирующее обслуживание). Deep Learning for Time Series Prediction. Seasonal Adjustment (Корректировка сезонности). Time Series Clustering (Кластеризация временных рядов). Explainable AI (XAI) in Time Series Forecasting.

### 5.2. Учебно-тематический план

Очная форма обучения

| № п/п | Наименование тем(разделов) дисциплины | Трудоемкость в часах | | | | |
|---|---|---|---|---|---|---|
| | | Всего | Контактная работа - Аудиторная работа | | | Самостоятельная работа |
| | | | Общая, в т.ч.: | Лекции | Семинары, практические занятия | |
| 1 | Введение в машинное обучение | 12 | 3 | 1 | 2 | 9 |
| 2 | Подготовка и предварительный анализ данных | 12 | 3 | 1 | 2 | 9 |
| 3 | Модели регрессии | 12 | 3 | 1 | 2 | 9 |
| 4 | Логистическая регрессия | 12 | 3 | 1 | 2 | 9 |
| 5 | Байесовский подход | 12 | 3 | 1 | 2 | 9 |
| 6 | Метрические алгорит | 11 | 2 | 0 | 2 | 9 |

| | | | | | | |
|---|---|---|---|---|---|---|
| | мы | | | | | |
| 7 | Методы сокращения размерности | 13 | 4 | 1 | 3 | 9 |
| 8 | Линейные алгоритмы классификации | 12 | 3 | 0 | 3 | 9 |
| 9 | Модели классификации Pandas и Scikit-Learn | 13 | 4 | 1 | 3 | 9 |
| 10 | Композиции алгоритмов | 13 | 4 | 1 | 3 | 9 |
| 11 | Задача кластеризации | 11 | 4 | 1 | 3 | 7 |
| 12 | Прогнозирование временных рядов | 11 | 4 | 1 | 3 | 7 |
| | Итого | 144 | 40 | 10 | 30 | 104 |

### 5.3. Содержание семинаров, практических занятий

| **Наименование тем (разделов) дисциплины** | **Перечень вопросов для обсуждения на семинарах, практических занятиях** | **Формы проведения занятий** |
|---|---|---|
| Введение в машинное обучение | История анализа данных. Основные понятия: обучение с учителем, обучение без учителя, обучение с подкреплением, модель алгоритмов, | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль, обучающая и тестовая выборка. Основные типы данных: таблицы, тексты, изображения, звук. Современные достижения. Задачи машинного обучения. Важнейшие термины и понятия. Современное состояние дел и перспективы развития. Популярные библиотеки и платформы. Применение машинного обучения в индустрии. Этические аспекты и регулирование. Будущие вызовы и направления исследований. | |
| Подготовка и предварительный анализ данных | Табличные данные. Бинарные данные. Номинальные данные. Номинальные упорядоченные данные. Количественные данные. Гистограммы. Линейные графики. Диаграммы рассеяния. Секторные диаграммы. Столбчатые диаграммы. Ящики с усами. Обработка отсутствующих данных. Номинальные и количественные | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | признаки. Проверка наличия выбросов. Анализ корреляции признаков. Форматы представления данных. Типы переменных и измерение центральной тенденции. Нормализация и стандартизация данных. Методы визуализации данных. Превращение категорийных данных в числовой формат. | |
| Модели регрессии | Теорема Гаусса-Маркова. Метод наименьших квадратов. Модель линейной регрессии. Нелинейные модели регрессии. Линеаризирующие преобразования функций. Непараметрическая регрессия. Оценка качества регрессионных моделей: среднеквадратическая ошибка (MSE), средняя абсолютная ошибка (MAE), коэффициент детерминации (R2). Регуляризация в регрессии: L1 (лассо), L2 (ридж), ElasticNet. Проверка гомоскедастичности остатков, тесты на гетероскедастичность (например, тест Голдфельда-Квандта). Взаимодействие признаков и полиномиальные | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | регрессии. Методы подбора лучшей модели: AIC/BIC, перекрестная проверка (кросс-валидация). Деревья решений и случайные леса в задачах регрессии. Байесовская регрессия и сравнение с традиционной (частотной) регрессией. Нейронные сети для регрессионных задач. Гребневая регрессия (ridge regression) и её применение в проблемах мультиколлинеарности. Квазирегрессия (quasi-regrssion) и её роль в борьбе с ненормальностью остатков. | |
| Логистическая регрессия | Переход от линейной регрессии к логистической. Понятие наблюдаемой вероятности и шансов. Функция потерь логистической регрессии. Градиентный спуск. Пакетный градиентный спуск. Стохастический градиентный спуск. Мини-пакетный градиентный спуск. Значение логистической функции (сигмоиды) и её интерпретация. Преимущества и ограничения логистической регрессии. Метод максимальной | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | правдоподобия и его связь с логистической регрессией. Регуляризация в логистической регрессии (L1, L2, Elastic Net). Кривая ROC и показатель AUC как мера качества модели. Интерпретация коэффициентов логистической регрессии. Шаги для подбора гиперпараметров логистической регрессии. Внедрение логистической регрессии с использованием Python-библиотеки Scikit-learn. Ранжирование и ранжирующая логистическая регрессия (RankNet). Проблемы мультиколлинеарности и способы их разрешения в логистической регрессии. | |
| Байесовский подход | Оценка Надарая-Ватсона с постоянной и переменной шириной окна. Выбор функции ядра. Проблема выбросов. Задача отсева выбросов. Мультиколлинеарность признаков. Методы устранения мультколлинеарности. Регуляризация корреляционной матрицы. Принцип Байесовской оценки и её отличие от традиционного | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | подхода. Основы Байесовского вывода и формула Байеса. Байесовская линейная регрессия и её преимущества. Инструменты для работы с неопределённостью и апостериорные распределения. Варианты выбора a priori распределения (например, нормальное, Лапласа, гамма-распределение). Решение проблемы переобучения с помощью Байесовского подхода. Гибридные методы (байес + регуляризация) и их реализация. Преимущества и ограничения Байесовского подхода в машинном обучении. Сравнительный анализ частотного и Байесовского подходов. Примеры применения Байесовского подхода в реальных задачах (например, медицинские диагнозы, прогнозирование спроса). | |
| Метрические алгоритмы | Обобщённый метрический регрессор. Метод потенциальных функций и его связь с линейной моделью классификации. AUC-PR кривая. Основные виды метрик | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | расстояния (Евклидово расстояние, Манхэттенское расстояние, Махаланобисовское расстояние). Классификация на основе ближайшего соседа (kNN) и методы выбора оптимального числа соседей (k). Недостатки простого kNN и способы их устранения (взвешенный kNN, использование различных метрик расстояния). Улучшенные версии kNN: Locally Weighted Learning (LWL), Radius-Based Nearest Neighbour (RBNN). Обоснование использования метрических методов в задачах регрессии и классификации. Способы выбора оптимальной метрики и функции ядра для обобщённого метрического регрессора. Интерпретация PR-кривой и отличие от ROC-кривой. Связь между AUC-PR и качеством классификатора при неравномерном классе (дисбаланс классов). | |
| Методы сокращения размерности | Варимакс вращение. Факторные нагрузки. Многомерное шкалирование. Сингулярное разложение матриц. Метод главных | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | компонент (PCA). Линейный дискриминантный анализ (LDA). Нелинейные методы снижения размерности. Автокодировщики (Autoencoders). Регуляризация при снижении размерности. Выбор числа главных компонент (выбор размерности). Интерпретация факторных нагрузок. Недостатки и ограничения методов сокращения размерности. Методы селекции признаков (Feature Selection). Способы нормализации данных перед снижением размерности. Аппаратное обеспечение и требования к памяти при использовании PCA и аналогичных методов. Использование PCA совместно с другими алгоритмами машинного обучения. Другие типы вращения, кроме варимакс. Альтернативы PCA для задач классификации. Проблемы переноса полученных знаний на другие задачи и классы данных. Принципы подбора оптимального числа главных компонент (номер главной | |

| | | |
|---|---|---|
| | компоненты). Оценка результата снижения размерности. Вероятностные методы PCA и Sparse PCA. | |
| Линейные алгоритмы классификации | Классификация линейной функцией. Линейная модель нейрона МакКаллока-Питтса. Градиентный метод численной минимизации. Проблема переобучения. Валидационная выборка. Разные штрафы за ошибки. Логистическая регрессия и её связь с линейной классификацией. Регуляризация в линейных классификаторах (L1, L2). Стохастический градиентный спуск (SGD). Многоклассовая классификация (One-Vs-Rest, One-Vs-One). Support Vector Machines (SVM) и их связь с линейными классификаторами. Cross-validation (кросс-валидация). Ridge classification (регуляризация гребнем). Выбор функционала потерь (Loss function). | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |
| Модели классификации Pandas и Scikit-Learn | Понятие зазора между классами. Случаи линейной разделимости и отсутствия линейной разделимости. Мультиклассовая логистическая регрессия. Многослойные | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | персептрон. Методы регулярных выражений (regularization) в логистической регрессии и нейронных сетях. Градиентный спуск и его модификации. Критерии качества классификации: точность (accuracy), precision, recall, f1-score, ROC-AUC. Проблема дисбаланса классов (class imbalance). Гиперпараметры моделей и их настройка. Работа с категориальными признаками. | |
| Композиции алгоритмов | Теоретические предпосылки композиционных методов. Одиночные модели и ансамбли моделей. Простое голосование. Взвешенное голосование. Аппроксимация функции потерь: сигмоидная, логарифмическая, кусочно-линейная, экспоненциальная, квадратичная. Метод случайных подпространств. Оптимизация моделей. Прунинг. Последовательные методы композиции (бустинг). Параллельные методы композиции (бэггинг). Случайный лес (Random Forest). Stacking (метакомпозиция). Dropouts и | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | регуляризация в композициях. Практические аспекты ансамблей. | |
| Задача кластеризации | Типы кластерных структур. Оптимизационные постановки задач кластеризации и частичного обучения. Распространение сходства (Affinity Propagation), мини-пакетный K-средних (Mini-Batch K-Means), сдвиг среднего значения (Mean Shift), упорядочение точек для обнаружения кластерной структуры (OPTICS), спектральная кластеризация (Spectral Clustering). EM-алгоритм. Построение дендрограммы. Критерий Акаике (AIC), Байесовский информационный критерий (BIC). Оценка качества кластеризации: индексы ARI и AMI, гомогенность, полнота, V-мера. DBSCAN (Density-Based Spatial Clustering of Applications with Noise). Birch (Balanced Iterative Reducing and Clustering using Hierarchies). Агломеративная кластеризация (Hierarchical clustering). K-medoids (Partition Around Medoids). Silhouette score (силуэтовая оценка). Gap-statistics (критерий Gap). Spectral embedding (спектральное вложение). Graph-based | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | clustering (кластеризация на графах). EM-algorithm for Gaussian mixture model (алгоритм ожидания-увеличения для смеси гауссовых распределений). Визуализация результатов кластеризации. | |
| Прогнозирование временных рядов | Сезонная авторегрессия с интегрированным скользящим средним (SARIMA). Сезонное авторегрессионное интегрированное скользящее среднее с экзогенными регрессорами (SARIMAX). Скользящее среднее векторной авторегрессии с экзогенными регрессорами (VARMAX). Экспоненциальное сглаживание Хольта-Винтерса (HWES). Подбор параметров моделей. Авторегрессия (AR) и скользящие средние (MA). Простое экспоненциальное сглаживание (Simple Exponential Smoothing, SES). Анализ стационарности временных рядов. Критерии подбора моделей (Akaike Information Criterion, AIC / Bayesian Information Criterion, BIC). Оценка качества прогнозов. Обработка пропущенных значений и аномалий | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | в данных. Использование внешних факторов (экзогенных регрессоров). Проблема мультиколлинеарности и автокорреляции. Разделение данных на тренировочную и тестовую выборки. | |

## 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 | История анализа данных. Основные понятия: обучение с учителем, обучение без учителя, обучение с подкреплением, модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль, обучающая и тестовая выборка. Основные типы данных: таблицы, тексты, изображения, звук. Современные достижения. | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |
| Подготовка и | Табличные данные. | Обзор литературы и |

| | | |
|---|---|---|
| предварительный анализ данных | Бинарные данные. Номинальные данные. Номинальные упорядоченные данные. Количественные данные. Гистограммы. Линейные графики. Диаграммы рассеяния. Секторные диаграммы. Столбчатые диаграммы. Ящики с усами. Промежуточные итоги и сводные таблицы. Фильтрация и отбор данных. Кодирование категориальных признаков. Работа с пропущенными значениями. Стандартизация и нормализация данных. Обнаружение и обработка выбросов. Распознавание и борьба с мультиколлинеарностью. Анализ структуры данных и формирование гипотез. Рабочие инструменты и библиотеки для анализа данных. | веб-источников, выполнение заданий тестирования для самоподготовки. |
| Модели регрессии | Теорема Гаусса-Маркова. Метод наименьших квадратов. Модель линейной регрессии. Нелинейные модели регрессии. Линеаризирующие преобразования функций. Непараметрическая регрессия. Регуляризация в | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | регрессии. Выбор и оценка качества модели. Доверительные интервалы и p-значения коэффициентов. Проблема мультиколлинеарности. Работа с гетероскедастичностью. Применение регрессии в задачах прогнозирования. Частота ложноположительных и ложноотрицательных прогнозов. | |
| Логистическая регрессия | Переход от линейной регрессии к логистической. Понятие наблюдаемой вероятности и шансов. Функция потерь логистической регрессии. Градиентный спуск. Пакетный градиентный спуск. Стохастический градиентный спуск. Мини-пакетный градиентный спуск. Форма уравнения логистической регрессии и интерпретация коэффициентов. Регуляризация в логистической регрессии (L1, L2). Методы выбора оптимального шага обучения (Learning Rate). Проблема мультиколлинеарности в логистической регрессии. | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | Практическое применение логистической регрессии. Сравнение логистической регрессии с другими методами классификации. | |
| Байесовский подход | Оценка Надарая-Ватсона с постоянной и переменной шириной окна. Выбор функции ядра. Проблема выбросов. Задача отсева выбросов. Мультиколлинеарность признаков. Методы устранения мультколлинеарности. Регуляризация корреляционной матрицы. | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |
| Метрические алгоритмы | Обобщённый метрический регрессор. Метод потенциальных функций и его связь с линейной моделью классификации. AUC-PR кривая. | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |
| Методы сокращения размерности | Варимакс вращение. Факторные нагрузки. Многомерное шкалирование. Сингулярное разложение матриц. | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |
| Линейные алгоритмы классификации | Классификация линейной функцией. Линейная модель нейрона МакКаллока-Питтса. Градиентный метод численной минимизации. Проблема переобучения. Валидационная выборка. Разные штрафы за ошибки. | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| Модели классификации Pandas и Scikit-Learn | Понятие зазора между классами. Случаи линейной разделимости и отсутствия линейной разделимости. Мультиклассовая логистическая регрессия. Многослойные персептрон. | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |
| Композиции алгоритмов | Теоретические предпосылки композиционных методов. Одиночные модели и ансамбли моделей. Простое голосование. Взвешенное голосование. Аппроксимация функции потерь: сигмоидная, логарифмическая, кусочно-линейная, экспоненциальная, квадратичная. Метод случайных подпространств. Оптимизация моделей. Прунинг. | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |
| Задача кластеризации | Типы кластерных структур. Оптимизационные постановки задач кластеризации и частичного обучения. Распространение сходства (Affinity Propagation), мини-пакетный K-средних (Mini-Batch K-Means), сдвиг среднего значения (Mean Shift), упорядочение точек для обнаружения кластерной структуры (OPTICS), спектральная кластеризация (Spectral Clustering). EM- | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

| | | |
|---|---|---|
| | алгоритм. Построение дендрограммы. Критерий Акаике (AIC), Байесовский информационный критерий (BIC). Оценка качества кластеризации: индексы ARI и AMI, гомогенность, полнота, V-мера. | |
| Прогнозирование временных рядов | Сезонная авторегрессия с интегрированным скользящим средним (SARIMA). Сезонное авторегрессионное интегрированное скользящее среднее с экзогенными регрессорами (SARIMAX). Скользящее среднее векторной авторегрессии с экзогенными регрессорами (VARMAX). Экспоненциальное сглаживание Хольта-Винтерса (HWES). Подбор параметров моделей. | Обзор литературы и веб-источников, выполнение заданий тестирования для самоподготовки. |

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

### Примерные задания контрольной работы

### Задание 1.

Для информации о проверке качества деталей по множеству признаков с обнаружением дефектных деталей, которая хранится в файле «data_defect.csv»:

• примените не менее 3 способов предобработки данных;

• постройте модель классификации на базе метода опорных векторов, логистической регрессии и градиентного бустинга xgBoost с использованием кросс-валидации и всех наборов предобработанных данных;

• выполните сравнительную оценку качества по метрике Accuracy;

• выполните сравнительную оценку качества по метрике F-мера;

• предложите способы повышения метрик качества;

• сделайте вывод по проведенному анализу.

**Задание 2.**

Предлагается массив данных с информацией о статистике выступления футбольных команд, который хранится в файле «data_football.csv». Для данного массива:

• определите корреляционную связь между признаками;

• выполните стандартизацию данных;

• сократите размерность до 2 главных компонент;

• на основе 2 главных компонент реализуйте алгоритм кластеризации K-средних, предварительно определив по коэффициенту силуэта оптимальное число кластеров;

• визуализируйте результаты наилучшей кластеризации;

• на тех же данных примените алгоритм DBSCAN с расстоянием по умолчанию;

• сравните результат с результатами K-средних;

• сделайте вывод по проведенному анализу.

**Задание 3.**

Банкротство компаний.

Имеется таблица данных о характеристиках российских предприятий, хранящихся в файле bankruptcy_data.csv. Требуется:

- Выполнить очистку данных, заполнив пропуски и устранив выбросы.

- Предобработать данные, выполнив масштабирование и кодирование категориальных признаков.

- Построить модели классификации на основе метода k-ближайших соседей, случайного леса и градиентного бустинга с использованием всех вариантов предобработки.

- Оценить качество моделей по метрикам Accuracy и F1-мере с помощью кросс-валидации.
- Сделать выводы о качестве моделей и предложить методы для повышения метрик.

**Задание 4.**

Клинические данные
Имеется набор клинических данных пациентов с онкологическими заболеваниями, хранящийся в файле clinical_data.csv. Необходимо:

- Найти корреляцию между признаками и удалить высоко-коррелируемые признаки.
- Применить метод главных компонент (PCA) для сведения размерности до трех главных компонент.
- Используя полученный новый набор признаков, реализовать алгоритм кластеризации с помощью метода K-средних и DBSCAN.
- Сравнить качество кластеризации на основе индекса Силуэта и визуализировать результаты.
- Сделать выводы о результатах кластеризации и предложить возможные улучшения.

**Задание 5.**

Покупатели интернет-магазина.
Имеется массив данных о поведении покупателей интернет-магазина, содержащий около 10 тысяч записей и хранимый в файле shop_customers.csv. Поставлена следующая задача:

- Применить три способа предобработки данных (нормализация, масштабирование, заполнение пропусков).
- Разделить данные на обучающий и тестовый наборы.
- Реализовать классификацию на основе логистической регрессии, случайного леса и градиентного бустинга.
- Оценить качество моделей по метрикам точности (Accuracy) и F1-мере с использованием кросс-валидации.

- Вывести заключение по качеству моделей и предложить способы улучшения метрик.

**Задание 6.**

В файле wine_quality.csv содержатся данные о химическом составе вин различного сорта. Ваша задача:

- Определить корреляцию между характеристиками вин и показателем качества.
- Провести стандартизацию данных и сведение размерности с помощью PCA до двух главных компонент.
- Выполнить кластеризацию с помощью алгоритма K-средних и DBSCAN, выбрав оптимальный параметр числа кластеров по индексу Силуэта.
- Визуализировать результаты кластеризации и сравнить два метода.
- Сделать вывод о результатах и возможностях улучшения качества кластеризации.

*Критерии балльной оценки различных форм текущего контроля успеваемости содержатся в соответствующих методических рекомендациях Кафедры искусственного интеллекта.*

**Дополнительная информация:**

**Примеры тестовых заданий**

*1. Для какого типа обучения главная цель научить модель делать выводы и находить закономерности.*

А) обучение без учителя

Б) обучение с подкреплением

В) обучение с учителем

*2. Что из перечисленного не является методом борьбы с пропусками в данных?*

А) Исключение объектов

Б) Интерполяция значений

В) Нормализация данных

*3. С чем, из ниже перечисленного сравнивают линейный классификатор?*

А) с аксоном

Б) с нейроном

В) с правилом Хебба

## 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

#### ПК-2 Способность разрабатывать прикладные решения с использованием методов компьютерного зрения

**1) Использовать основные модели и методы компьютерной обработки изображений и распознавания образов, модели и методы компьютерного зрения**

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

**Знать:** основные алгоритмы и методы, используемые в компьютерной обработке изображений, такие как фильтрация, сегментация, детекция границ, преобразование Фурье и др.

**Уметь:** выбирать и применять соответствующие модели и методы компьютерного зрения для решения конкретных задач, таких как распознавание лиц, анализ медицинских изображений или автоматизация процессов в производстве

**Типовые контрольные задания**

Разработать систему, которая будет распознавать лица людей в видео потоке с веб-камеры. Система должна использовать основные модели и методы компьютерной обработки изображений и распознавания образов, а также методы компьютерного зрения.

**2) Решать с помощью моделей и методов компьютерного зрения прикладные задачи предметной области**

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

**Знать:** основные алгоритмы распознавания образов, такие как методы машинного обучения (например, SVM, нейронные сети), а также современные подходы, такие как глубокое обучение (например, сверточные нейронные сети)

**Уметь:** использовать программные библиотеки и фреймворки (например, OpenCV, TensorFlow, PyTorch) для разработки прикладных решений в области компьютерного зрения

**Типовые контрольные задания**

Требуется создать веб-приложение для анализа аватарок социальной сети на платформе VKontakte.

Собрать данные о группе пользователей и реализовать интерфейс, позволяющий визуализировать социальную сеть, проводить анализ уровней вовлечённости пользователей, находить лидеров мнений и сегменты аудитории.

**ПК-3 Способность разрабатывать прикладные решения с использованием методов анализа текстов, графов и сетевых структур**

**1) Использовать основные модели и методы обработки текстов на естественных языках, графов и семантических структур**

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

**Знать:** основы Natural Language Processing (NLP), включая токенизацию, стемминг, лемматизацию, векторизацию текстов, выделение сущностей и отношений, классификацию и тематическое моделирование

**Уметь:** проектировать архитектуры решений, комбинирующие анализ текстов и графов, реализовывать алгоритмы на языках программирования вроде Python, JavaScript или R, и эффективно интегрировать их в существующие информационные системы

**Типовые контрольные задания**

1. Предоставлена база отзывов клиентов компании. Используя методы NLP, проведите автоматический анализ настроений отзывов, выделив положительные и отрицательные мнения. Разработайте инструмент визуализации частотного словаря негативных и позитивных высказываний.

2. Представлен набор сообщений из чата технической поддержки. Разделите потоки обращений на темы запросов и выработайте автоматизированную схему первичной сортировки клиентских заявок. Проверьте результативность модели на тестовых данных.

**2) Решать с помощью прикладных моделей NLP, графовой и сетевой аналитики прикладных задач предметной области**

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

**Знать:** основы методов NLP ( токенизация, лемматизация, удаление стоп-слов, очистка данных); векторизацию текстов (TF-IDF, Word2Vec, BERT и другие эмбеддинги); машинное обучение и глубокое обучение для задач классификации, сентимент-анализа, перевода и др; основные концепции графовой аналитики и сетевого анализа: графы и способы их представления (списки смежности, матрица смежности); центральность узлов (Degree Centrality, Betweenness Centrality, Eigenvector Centrality); выделение сообществ и кластеров (Louvain Method, Girvan-Newman Algorithm); библиотеки и инструменты для реализации: NLP: SpaCy, NLTK, Gensim, Transformers (HuggingFace); графовую аналитику: NetworkX, igraph, Gephi, Neo4j; применение этих технологий в рамках своей специальности (экономика, медицина, маркетинг и т.п.).

**Уметь:** разрабатывать и настраивать NLP-модели для решения прикладных задач (например, классификация документов, анализ тональности отзывов, поиск похожих текстов); строить и анализировать графы для представления сложных данных и связей (социальные сети, транзакционные данные, библиометрические показатели); использовать готовые библиотеки и API для быстрого прототипирования и тестирования решений; формулировать задачи прикладного характера, адекватно

выбирать инструментарий и создавать эффективные решения для бизнес-задач или научных проектов; визуализировать результаты графового анализа и представлять их визуально привлекательно и понятно; интегрировать методы NLP и графового анализа в IT-инфраструктуру компаний или лабораторий для автоматизации процессов принятия решений;

**Типовые контрольные задания**

1. Даны связи сотрудников в крупной корпорации. Выполните анализ социальной сети предприятия, определив самые влиятельные узлы и сообщества. Постройте график, отображающий уровни взаимодействия сотрудников, а также степень централизации управления в компании.

2. Имеются данные посещаемости сайта компании. Реализуйте графовую модель посещения страниц пользователями, рассчитав коэффициенты переходов между страницами и выделив узкие места навигации ресурса. Представьте отчёт с рекомендациями по улучшению UX.

**Примеры практико-ориентированных заданий**

**Задание 1.**

Имеется тестовый набор данных. Оцените оптимальное число кластеров в данных. Реализуйте алгоритм K-средних с соответствующим числом кластеров. Выполните оценку качества полученного решения.

**Задание 2.**

Имеется тестовый набор данных. Обучите модель линейной и полиномиальной регрессии для прогнозирования целевого признака, предварительно нормализовав данные. Выполните оценку эффективности модели на валидационном датасете.

**Примерные вопросы для подготовки к Экзамену**

1. Исторические вехи развития искусственного интеллекта. Методы и задачи в машинном обучении. Основные виды в МО.

2. Кибернетика и первые нейронные сети. Перцептрон Розенблатта.

3. Обработка естественного языка. Задачи, модели и методы, достижения.

4. Компьютерное зрение. Задачи, модели и методы, достижения.

5. Обучение с подкреплением. Задачи, модели и методы, достижения.

6. Сильный искусственный интеллект. Отличие от классического искусственного интеллекта.

7. Задачи классификации и регрессии в машинном обучении.

8. Задачи ранжирования и кластеризации в машинном обучении.

9. Функции потерь и функции активации. Метод обратного распространения ошибки.

10. Паралич сети, недообучение, переобучение. Методы регуляризации.

11. Проблема данных и этапы процесса машинного обучения. Категории данных. Примеры.

12. Функционал качества в задачах классификации, регрессии.

13. Функционал качества в задачах кластеризации.

14. Методы предобработки данных и борьбы с пропусками.

15. Метод градиентного спуска. Другие методы оптимизации.

16. Теорема Байеса и Байесовский классификатор.

17. Глубокое обучение. Задачи, модели и методы, достижения. Гиперпараметры моделей машинного обучения.

18. Методы контроля обучения.

19. Сверточные нейронные сети и рекуррентные нейронные сети.

20. Библиотеки pandas, numpy и matplotlib для анализа и визуализации данных.

21. Библиотека scikit-learn для построения моделей машинного обучения.

22. Метод K-ближайших соседей, модель дерева решений.

23. Модель линейной регрессии, модель логистической регрессии.

24. Метод опорных векторов. Ядерные методы в машинном обучении.

25. Модель гауссовых смесей. Метод K-средних.

26. Агломеративная кластеризация. Кластеризация DBSCAN.

27. Локально взвешенное сглаживание.

28. Ансамблевые модели. Бустинг и бэггинг.

29. Прогнозирование временных рядов.

**Пример экзаменационного билета**

**Задание 1. (20 баллов). Теоретический вопрос.**

Задача классификации. Модели решения на основе машинного обучения. Оценка качества решения задачи

**Задание 2 (40 баллов). Практическое задание.**

Исследовать модель логистической регрессии для датасета «bank.csv» в наборе данных Bank, предварительно подготовив данные для извлечения информации из категориальных переменных с помощью методов Label Encoding и One-Hot Encoding. Получить метрики precision, recall и accuracy для данных с преобразованием переменной education, education и job, и не менее 4х категориальных переменных. Построить зависимость доли верных распознаваний на тесте от количества закодированных переменных в обучающей выборке размером 85%. Построить модель классификации на тех же данных с использованием кросс-валидации и многослойного перспетрона. Сделать выводы.

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

**Основная литература:**

1. Анализ данных и машинное обучение на платформе MS SQL Server : Учебное пособие / Ю.Н. Кондрашов Электрон. дан. Москва : Русайнс , 2024 303 с. Режим доступа: book.ru Internet access https://book.ru/book/953643 ISBN 978-5-466-06473-5

2. Сопов , Евгений Александрович Многокритериальные нейроэволюционные системы в задачах машинного обучения и человеко-машинного взаимодействия : Монография / Сибирский государственный университет науки и технологий им. академика М.Ф. Решетнева Красноярск : Сибирский федеральный университет , 2019 160 с. ВО - Магистратура https://znanium.com/catalog/document?id=380296 ISBN 978-5-7638-3969-2

**Дополнительная литература:**

1. Математика на языке R [ Электронный ресурс ] / Зададаев С. А. 3-е изд., доп. и перераб. Москва : Прометей , 2024 346 с. Книга из коллекции Прометей - Математика https://e.lanbook.com/book/445976 ISBN 978-5-00172-606-7

2. Анализ данных. Компьютерный практикум : Учебное пособие / В.Н. Калинина, В.И. Соловьев Электрон. дан. Москва : КноРус , 2022 166 с. Режим доступа: book.ru Internet access https://book.ru/book/942681 ISBN 978-5-406-09229-3

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

1. Личный кабинет обучающегося https://org.fa.ru

2. Электронная библиотека Финансового университета (ЭБ) http://elib.fa.ru/ (http://library.fa.ru/files/elibfa.pdf)

3. Электронно-библиотечная система BOOK.RU http://www.book.ru

4. Электронно-библиотечная система "Университетская библиотека ОНЛАЙН" http://biblioclub.ru/

5. Образовательная платформа "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" издательства "Альпина Паблишер" http://lib.alpinadigital.ru/en/library

9. •Научная электронная библиотека eLibrary.ru http://elibrary.ru

10. •Национальная электронная библиотека http://нэб.рф/

11. СПАРК https://spark-interfax.ru/

12. Academic Reference http://ar.cnki.net/ACADREF

13. Пакет баз данных компании EBSCO Publishing, крупнейшего агрегатора научных ресурсов ведущих издательств мира http://search.ebscohost.com

14. •Электронные продукты издательства Elsevier http://www.sciencedirect.com

15. Emerald: Management eJournal Portfolio https://www.emerald.com/insight/

16. Scopus https://www.scopus.com

17. •Электронные коллекции книг и журналов издательства Springer: http://link.springer.com/

## 10. Методические указания для обучающихся по освоению дисциплины

Студентам необходимо ознакомиться:

- с содержанием рабочей программы дисциплины (далее – РПД), с целями и задачами дисциплины, ее связями с другими дисциплинами образовательной программы, методическими

разработками по данной дисциплине, имеющимися на образовательном портале и сайте департамента, с графиком консультаций преподавателей.

Рекомендации по подготовке к лекционным занятиям

(теоретический курс)

Изучение дисциплины требует систематического и последовательного накопления знаний, следовательно, пропуски отдельных тем не позволяют глубоко освоить предмет. Именно поэтому контроль над систематической работой студентов всегда находится в центре внимания департамента.

Студентам рекомендуется:

- перед каждой лекцией просматривать рабочую программу дисциплины, что позволит сэкономить время на записывание темы лекции, ее основных вопросов, рекомендуемой литературы;

- на отдельные лекции приносить соответствующий материал на бумажных или электронных носителях, представленный лектором на портале. Данный материал будет охарактеризован, прокомментирован, дополнен непосредственно не лекции;

- перед очередной лекцией необходимо просмотреть по конспекту материал предыдущей лекции. При затруднениях в восприятии материала следует обратиться к основным литературным источникам, если разобраться в материале опять не удалось, то обратитесь к лектору (по графику его консультаций) или к преподавателю на практических занятиях. Не оставляйте «белых пятен» в освоении материала.

Рекомендации по подготовке к практическим (семинарским) занятиям

Студентам следует:

- приносить с собой рекомендованную преподавателем литературу к конкретному занятию;

- до очередного практического занятия по рекомендованным литературным источникам проработать теоретический материал, соответствующей темы занятия;

- при подготовке к практическим занятиям следует обязательно использовать не только лекции, но и другую учебную литературу;

- в начале занятий задать преподавателю вопросы по материалу, вызвавшему затруднения в его понимании и освоении, при решении задач, заданных для самостоятельного решения;

- в ходе семинара давать конкретные, четкие ответы по существу вопросов;

- на занятии доводить каждую задачу до окончательного решения, демонстрировать понимание проведенных расчетов (анализов, ситуаций), в случае затруднений обращаться к преподавателю.

Методические рекомендации по выполнению различных форм самостоятельных домашних заданий

Самостоятельная работа студентов включает в себя выполнение различного рода заданий, которые ориентированы на более глубокое усвоение материала изучаемой дисциплины. К выполнению заданий для самостоятельной работы предъявляются следующие требования: задания должны выполняться самостоятельно и представляться в установленный срок, а также соответствовать установленным требованиям по оформлению.

Студентам следует:

- руководствоваться графиком самостоятельной работы, определенным РПД;

- выполнять все плановые задания, выдаваемые преподавателем для самостоятельного выполнения, и разбирать на семинарах и консультациях неясные вопросы;

- использовать при подготовке нормативные документы Финансового университета;

- при подготовке к экзамену параллельно прорабатывать соответствующие теоретические и практические разделы дисциплины, фиксируя неясные моменты для их обсуждения на плановой консультации.

Методические рекомендации по работе с литературой

Любая форма самостоятельной работы студента (подготовка к семинарскому занятию, выполнение практического задания для самостоятельной работы, начинается с изучения соответствующей литературы как в библиотеке, так и дома.

К каждой теме учебной дисциплины подобрана основная и дополнительная литература.

При работе с литературой рекомендуется делать записи. Записи в той или иной форме не только способствуют пониманию и усвоению изучаемого материала, но и помогают вырабатывать навыки явного изложения в письменной форме тех или иных теоретических вопросов.

Методические указания по проведению практических занятий

По структуре практические занятия следует разделить на учебные и контрольные.

Учебные практические занятия структурно состоят из следующих компонент:

• проверка наличия выполненного задания самостоятельной работы каждого студента;

• выборочная проверка корректности выполнения домашнего задания;

• разбор типичных ошибок, возникших в самостоятельной работе;

• рассмотрение теоретических вопросов, связанных с текущим практическим занятием;

• разбор методов выполнения практических заданий и решения задач;

• корректировка заданий для самостоятельной работы студентов.

Контрольные практические занятия структурно состоят из следующих компонент:

• проведение аудиторных самостоятельных работ;

• подведение итогов и разбор типичных ошибок, возникших при выполнении самостоятельных работ.

Студенты должны обратить внимание на перечень основных контрольных мероприятий, которые проводятся в соответствии с рабочей программой дисциплины.

Конкретные сроки проведения этих мероприятий своевременно доводятся до сведения студентов.

## 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. R 4.1.2 RStudio
2. Пакет офисных программ
3. Kaspersky

**- Современные профессиональные базы данных и информационные справочные системы:**

1. Информационно-правовая система «Консультант Плюс»
2. Информационно-правовая система «Гарант»
3. Электронная энциклопедия: http://ru.wikipedia.org/wiki/Wiki

**- 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)