Аннотация рабочей программы дисциплины

Б1.В.ДВ.01.02 «DATA-CENTRIC MACHINE LEARNING»

Курс 3 Семестр 6 Количество з.е. 2
Объем трудоемкости: 2 зачетных единиц (72 ч., из них – 34,2 час. аудиторной нагрузки: лекционных 16 ч., лабораторных работ - 16 ч., 37,8 часов самостоятельной работы, 2 часов КСР, 0,2 часа ИКР.), форма контроля – зачет.

Цель освоения дисциплины: формирование у студентов систематизированных знаний, практических умений и навыков применения современных методов искусственного интеллекта, машинного обучения для решения задач подготовки данных для дальнейшего применения моделях различных предметных областей.

Дисциплина направлена на развитие способности собирать, размечать, предобрабатывать данные и оценивать качество подготовленных данных.

Задачи дисциплины

1. Кроме методов решения типовых задач подготовки данных: обработка пропущенных значений, кодирование категориальных признаков, масштабирование и нормализация числовых данных и методов обработки выбросов (аномалий) в данных, изученных ранее в дисциплине «Многомерный статистический анализ и машинное обучение», усвоить принципы и методы feature engineering (создания и преобразования признаков) для повышения эффективности моделей машинного обучения.
2. Применять на практике методы работы с несбалансированными данными и подходы к разметке данных (labeling).
3. Применять на практике критерии и метрики для оценки качества подготовленных данных и их пригодности для решения конкретной задачи.
4. Приобрести практический навык применения методов предобработки данных: очистка от шума, обработка пропусков, кодирование категориальных переменных, масштабирование признаков. Сформировать умение создавать новые признаки (Feature Engineering) на основе существующих для улучшения предсказательной способности моделей. Освоить методы селекции признаков (Feature Selection) для отбора наиболее информативных переменных и уменьшения размерности данных.
5. Приобрести умение оценивать качество подготовленного набора данных с помощью визуализации и статистических метрик перед передачей его на этап моделирования.

Место дисциплины (модуля) в структуре образовательной программы

Дисциплина «Data-Centric Machine Learning» относится к части, формируемой участниками образовательных отношений Блока "Дисциплины (модули) по выбору" учебного плана (Б1.В.ДВ).

Дисциплина изучается в 6-м семестре. Для успешного освоения необходимы знания, полученные в дисциплинах: «Алгебра и введение в тензорный анализ», «Теория вероятностей и математическая статистка», «Многомерный статистический анализ», и «Машинное обучение», «Программирование».

Преподавание ведется в виде лекций и лабораторных занятий с использованием интерактивных методов. Лабораторные работы направлены на практическое освоение методов и инструментов классификации на реальных данных.

Дисциплина формирует компетенции, необходимые для выполнения выпускной квалификационной работы и профессиональной деятельности в области вычислительных технологий.

Результаты обучения (знания, умения, опыт, компетенции):

BD-1 **Способен осуществлять поиск, сбор, очистку и предварительный анализ данных (П)**

BD-1.1 Обосновывает способы и варианты применения методов предварительного анализа данных в задачах ИИ, включая их математическое (алгоритмическое) преобразование и адаптацию к специфике задачи.

Знает: Методы анализа распределений данных, поиска аномалий и выбросов, статистические тесты для оценки качества данных (нормальность, стационарность), принципы анализа мультимодальных и неструктурированных данных.

Умеет: выбирать методы EDA в зависимости от типа данных и задачи ML, интерпретировать результаты анализа для формулирования гипотез о качестве данных, адаптировать стандартные методы анализа под специфику доменной области.

Владеет библиотеками: pandas-profiling, sweetviz, dataprep, умеет строить интерактивные дашборды для анализа данных. Автоматизирует пайплайны предварительного анализа.

D-1.2 Применяет методы анализа данных для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ.

Знает: Методы формирования и проверки статистических гипотез, принципы A/B тестирования для оценки качества данных, метрики качества данных для ML (class imbalance, label noise, feature quality).

Умеет: Формулировать гипотезы о проблемах в данных на основе EDA, планировать эксперименты по оценке влияния качества данных на модель, валидировать гипотезы о смещениях в данных (bias detection).

Владеет: scipy.stats, statsmodels для статистического тестирования, проводит power analysis для экспериментов с данными, автоматизирует проверку гипотез о качестве данных.

BD-1.3 Применяет методы понижения размерности для первичной интерпретации и визуализации многомерных данных

Знает: алгоритмы линейного и нелинейного снижения размерности (PCA, t-SNE, UMAP), методы отбора признаков на основе важности и корреляций, метрики оценки качества снижения размерности

Умеет: Выбирать метод снижения размерности в зависимости от задачи и типа данных, интерпретировать результаты визуализации для выявления кластеров и аномалий, оценивать информативность признаков после преобразования.

Владеет: sklearn.decomposition, umap-learn, seaborn, умеет создавать интерактивные визуализации с plotly, оптимизирует параметры снижения размерности для интерпретируемости.

BD-1.4 **Знает** и умеет применить методы отбора признаков.

Владеет способностью применять методы отбора признаков данных, значимых для исследования.

Умеет отбирать признаки данных, значимые для исследования,

Владеет sklearn.feature_selection, rfpimp, SHAP, умеет проводить рекурсивное исключение признаков, строит матрицы корреляций и анализирует мультиколлинеарность.

BD-2 **Способен определять требования к наборам данных для решения задач машинного обучения, проводить разметку и анализ наборов данных, оценивать качество данных, обеспечивать непрерывную интеграцию данных**

BD-2.1 Определяет требования к наборам и качеству данных для решения задач машинного обучения.

Знает, как сформировать требования для набора данных. Владеет умениями по формированию требований к наборам и качеству данных для решения задач машинного обучения

BD-2.2 Работает с данными, в том числе собирает данные из разрозненных источников, проверяет данные на корректность.

Знает: методы интеграции данных из разнородных источников, принципы валидации и верификации данных, техники обработки пропущенных значений и аномалий.

Умеет: Проектировать ETL/ELT процессы для сбора данных, разрабатывать правила валидации данных (data contracts), выявлять и устранять проблемы согласованности данных.

Владеет: pandas, pyarrow, dask для обработки больших данных,умеет работать с API, базами данных, облачными хранилищами, автоматизирует проверки качества данных в пайплайнах.

LLM-2 **Способен дообучать, адаптировать и оптимизировать генеративные модели под специфические задачи и условия применения**

LLM-2.1 **Понимает принципы fine-tune**

Знает: Архитектуры и принципы работы современных LLM, методы адаптации моделей: full fine-tuning, LoRA, QLoRA, P-tuning, особенности подготовки данных для дообучения генеративных моделей.

Умеет: Выбирать стратегию дообучения в зависимости от задачи и ресурсов, оценивать риски катастрофического забывания и способы его смягчения, планировать эксперименты по оценке эффективности дообучения.

Понимает технические ограничения разных методов fine-tuning, умеет оценивать вычислительные требования для адаптации моделей, анализирует trade-offs между качеством и стоимостью дообучения.

LLM-2.2 **Создаёт обучающие наборы данных.**

Знает: Требования к данным для fine-tune: релевантность, объем, разнообразие, качество разметки. Форматы данных для разных методов дообучения (SFT, DPO, RLHF), принципы построения диалоговых систем и инструктивных данных, методы аугментации и синтеза данных для NLP, проектировать схемы разметки для задач дообучения LLM.

Умеет: формировать сбалансированные и разнообразные обучающие выборки

Оценивать качество и репрезентативность созданных датасетов. Умеет создавать синтетические данные с помощью LLM

Владеет инструментами для разметки текстовых данных (Label Studio, etc.), **Методами** обеспечения репрезентативности и сбалансированности создаваемого набора данных. **Технологиями** создания синтетических данных для задач, где

реальных данных недостаточно. **Полным циклом** подготовки данных: от сбора сырых данных до формирования готового для обучения объекта (DataLoader, Dataset).

MF-4 **Способен применять статистические методы для анализа данных, валидации моделей машинного обучения и проведения экспериментов в области ИИ**

MF-4.1 Применяет статистические методы анализа и машинного обучения для решения задач анализа данных и проведения экспериментов на данных.

Знает: дескриптивную статистику и методы анализа распределений, статистические тесты для сравнения групп и выявления различий, методы анализа временных рядов и последовательностей.

Умеет: выбирать статистические методы в зависимости от типа данных и гипотез, интерпретировать результаты статистического анализа для принятия решений. формулировать выводы на основе статистических доказательств.

Владеет: scipy.stats, statsmodels, pingouin, умеет проводить анализ мощности (power analysis), визуализирует результаты статистического анализа.

MF-4.2 Способен применять статистические методы для построения предсказательных моделей, включая методы для анализа и прогнозирования временных рядов, а также моделирования нестационарных случайных процессов.

Знает: теоретические основы и предположения линейной и логистической регрессии. Понятие стационарности временного ряда, методы проверки (ADF test) и приведения к стационарному виду (дифференцирование, декомпозиция). Классические модели прогнозирования временных рядов (ARIMA, SARIMA, ETS).**Умеет** формализовывать и применять статистические методы идентификации регрессионных и классификационных моделей, понимает основы базовых вероятностных моделей для временных рядов на основе авторегрессионных зависимостей.. **Владеет** приемами построения модели динамических систем для многомерных временных рядов и полей.

MF-4.3 Способен применять статистические методы для оценки качества моделей ИИ, включая метрики и критерии для регрессии, классификации и кластеризации, а также для проведения статистических тестов для сравнения моделей.

Знает метрики и меры качества моделей регрессии (в т.ч. на временных рядах), классификации, кластеризации.

Умеет оценивать качество моделей МО.

Владеет умением применять статистические методы для оценки качества моделей ИИ, включая метрики и критерии для регрессии, классификации и кластеризации, а также для проведения статистических тестов для сравнения моделей.

ML-2 **Способен применять фундаментальные принципы и методы машинного обучения включая подготовку данных оценку качества моделей и работу с признаками**

ML-2.1 Различает основные типы задач машинного обучения и применяет на практике принципы их решения.

Знает и различает основные типы задач машинного обучения (обучением с учителем, без учителя и с подкреплением). **Умеет** применить типовые подходы к решению базовых задач с использованием готовых инструментов и библиотек (ScikitLearn) (Б)

Умеет обоснованно применять методы решения задач машинного обучения с учётом характеристик данных и бизнес-контекста, настраивает базовые модели и проводит их оценку (П)
Владеет приемами и инструментами проектирования и реализации комплексных решений машинного обучения для нестандартных задач, включая разработку пайплайнов, оптимизацию моделей и интерпретацию результатов (Э)

Содержание и структура дисциплины:

Распределение видов учебной работы и их трудоемкости по разделам дисциплины.
Разделы дисциплины, изучаемые в _6_ семестре *(очная форма)*

№	Наименование разделов (тем)	Количество часов				
		Всего	Аудиторная работа			Внеаудиторная работа
			Л	ПЗ	ЛР	СРС
1.	Введение в Data-Centric Machine Learning. Процесс разметки данных (Data Labeling)	8	2		2	4
2.	Стратегии разметки данных. Согласованность разметки и разрешение конфликтов	8	2		2	4
3.	Поиск и очистка от шума в данных (Data Cleaning).	8	2		2	4
4.	Обогащение данных: Аугментация и работа с несбалансированными данными (Class Imbalance)	8	2		2	4
5.	Глубинный анализ ошибок и срезов данных (Error & Slice Analysis)	8	2		2	4
6.	Синтетические данные и управление данными	9	2		2	5
7.	Data-Centric подходы для NLP и CV	8,8	2		2	4,8
8.	Инфраструктура, мониторинг и этика	11	2		2	7
	ИТОГО по разделам дисциплины	69,8	16		16	37,8
	Контроль самостоятельной работы (КСР)	2				
	Промежуточная аттестация (ИКР)	0,2				
	Подготовка к текущему контролю	-				
	Общая трудоемкость по дисциплине	72				

Примечание: Л – лекции, КСР – контрольные и самостоятельные работы, ЛР – лабораторные занятия, СРС – самостоятельная работа студента

Курсовые проекты или работы.
Не предусмотрены учебным планом

Вид аттестации: ЛР, Комплексная итоговая работа, зачет.

Автор Приходько Т.А. – кандидат технических наук, доцент кафедры вычислительных технологий;