Аннотация рабочей программы дисциплины

Б1.В.ДВ.01.01 «Подготовка данных машинного обучения»

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 «Подготовка данных машинного обучения» является 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 «Подготовка данных машинного обучения» 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 заполнения пропусков в данных и удаления выбросов в табличных данных (случайные величины)
Имеет навыки (**умеет**) очистки зашумленных временных рядов и изображений. Обнаруживает и устраняет выбросы в данных временных рядов. **Владеет** подходами к заполнению пропусков в данных временных рядов и изображений.

BD-1.2 Применяет методы анализа данных для проверки разведочных гипотез и подготовки данных к применению современных методов ИИ
Знает основные методы понижения размерности
Умеет применить основные методы понижения размерности и подбирает оптимальную размерность в зависимости от необходимой доли объяснённой дисперсии.
Владеет методологией применения существующих библиотек, реализующих методы понижения размерности.

BD-1.3 Применяет методы понижения размерности для первичной интерпретации и визуализации многомерных данных
Знает и умеет применить основы методов отбора признаков и выбирает оптимальное подмножество признаков.
Владеет методологией применения существующих библиотек, реализующих методы отбора признаков.

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 методами finetuning

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 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 Знает приемы и инструменты для сбора данных из разрозненных источников. Умеет работать с данными, в том числе собирает данные из разрозненных источников, проверяет данные на корректность. Владеет языками и инструментами для сбора данных и оценки их корректности.

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

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

Знает: основные подходы к тонкой настройке: полная настройка всех параметров, поэтапная разморозка слоев, методы эффективной тонкой настройки (P-Tuning, LoRA, QLoRA, Adapter). Гиперпараметры, критически важные для fine-tune: learning rate, scheduler, batch size, и их отличия от обучения с нуля.

Умеет: Отличать дообучение от первичного обучения, знает базовые процедуры **fine-tune,** анализировать задачу и выбирать наиболее подходящий метод fine-tune (полная настройка vs. эффективные методы).

Владеет: **Навыком** осознанного выбора стратегии fine-tune под ограничения (вычислительные ресурсы, объем данных, требования к качеству). Применяет fine-tune к предобученным моделям на новых датасетах.

• **Методами** анализа и интерпретации процесса дообучения (использование логов, графиков потерь).

• **Критическим мышлением** для оценки целесообразности применения fine-tune в конкретном сценарии versus использования prompt engineering или RAG.

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

Знает: Требования к данным для fine-tune: релевантность, объем, разнообразие, качество разметки.Форматы данных для популярных фреймворков (Hugging Face, TensorFlow, PyTorch) и структур задач (текст-текст, текст-изображение, инстркукции и т.д.). Методы аугментации данных (data augmentation), специфичные для генеративных моделей (e.g., back-translation для текста, модификация промптов). Принципы разбиения данных на обучающую, валидационную и тестовую выборки.

Умеет: Выбирать методы с учетом требований к latency и ресурсам. собирать данные из различных источников: API, веб-скрапинг, открытые датасеты, синтетическая генерация. Очищать и предобрабатывать сырые данные: удаление шума, дубликатов, нормализация текста, приведение к единому формату. Размечать данные в соответствии с поставленной задачей (e.g., составлять пары "инструкция-ответ", аннотировать изображения). Применять методы аугментации данных для увеличения размера и разнообразия обучающего набора.

Владеет: Навыками работы с библиотеками и инструментами для обработки данных (Pandas, NumPy, Hugging Face Datasets). **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 отличия статистического обучения от не статистического, **владеет** классификацией методов статистического машинного обучения. **Умеет** применять и выбирать методы статистического машинного обучения, учитывая особенности данных и задачи, а также объясняет различия между подходами.

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 подготовку данных для МО. Планирование датасета под задачу.	8	2		2	4
2.	Первичный сбор данных и их оценка. Предобработка данных: очистка, обработка пропусков, выбросов.	8	2		2	4
3.	Разметка данных и аугментация.	8	2		2	4
4.	Предобработка изображений	8	2		2	4
5.	Предобработка аудиоданных	8	2		2	4
6.	Предобработка текстовых данных	9	2		2	5
7.	Предобработка данных для временных рядов	8,8	2		2	4,8
8.	Статистический анализ данных (EDA, визуализация, анализ распределений, корреляции, анализ признаков, feature importance).Стратегии по улучшению метрик через данные (feature engineering, отбор признаков, балансировка). Дизайн эксперимента по улучшению данных	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;