Учебный план курса

Раздел 1. Введение в машинное обучение

Занятие 1 (лекция). Введение в машинное обучение. Специфика задач обработки данных в физике.

Основные понятия. Типология задач, решаемых методами машинного обучения (МО): задачи регрессии, классификации, оптимизации и кластеризации; временные ряды как особый тип задач. Бинарная, многоклассовая и многометочная классификация. Анализ данных (data mining) и МО. МО и искусственный интеллект. Краткая историческая справка. Обзор больших успехов третьей волны в физике. Особенности данных на примере задач из физики: нелинейность, высокая размерность, мультиколлинеарность, плохая представительность, противоречивость и неполнота данных, дискретность значений данных, наличие шумов. Способы работы с каждой из этих особенностей.

Раздел 2. Подготовка и предобработка данных. Оценка качества моделей

Занятие 2 (практикум). Основные понятия языка Python и приёмы работы с ним.

Основные понятия языка Python. Основные библиотеки подготовки и обработки данных (Pandas, NumPy, SciPy). Основные библиотеки визуализации результатов (Matplotlib, Seaborn). Основные библиотеки машинного обучения (scikit-learn, tensorflow, keras, pytorch, fastai, mxnet). Инструменты разработки - Google Colab.

Занятие 3 (лекция). Подготовка данных. Оценка качества моделей.

Кодирование и нормировка данных. Методы оценки качества данных. Удаление выбросов. Заполнение пропусков в данных. Способы работы с несбалансированными данными. Методы и метрики оценки качества моделей для задач регрессии, классификации и кластеризации. Кросс-валидация.

Занятие 4 (лекция). Отбор и преобразование входных признаков. Оценка значимости входов.

Методы понижения входной размерности данных: отбор и преобразование входных признаков. Типы методов отбора признаков: фильтры, встроенные методы, обёртки. Комплексные методы отбора признаков. Фрактальная размерность данных и алгоритмы её определения.

Занятие 5 (практикум). Практические основы предобработки данных.

Основы предобработки данных. Исследование данных. Работа с табличными данными и изображениями.

Занятие 6 (лекция). Анализ главных компонент и методы на его основе. Кластер-анализ. НС Кохонена и самоорганизующиеся карты Кохонена.

Линейный и нелинейный анализ главных компонент. Метод проекций на латентные структуры. Многомерное разрешение кривых. Проекция t-SNE. Кластер-анализ. НС Кохонена и самоорганизующиеся карты Кохонена.

Раздел 3. Основные методы машинного обучения

Занятие 7 (лекция). Базовые методы машинного обучения.

Линейная регрессия. Логистическая регрессия. Регуляризация L1 и L2, ElasticNet. Регуляризация как встроенный метод отбора признаков. Машины опорных векторов. Kernel Trick. Метод k ближайших соседей. Деревья решений. Алгоритм случайного леса. Градиентный бустинг.

Занятие 8 (лекция). Многослойные персептроны. Алгоритм обратного распространения ошибки.

Формальный нейрон. Многослойный персептрон (МСП) как универсальный аппроксиматор. Алгоритм обратного распространения ошибки и его модификации. Переобучение (переучивание) и методы борьбы с ним. Нейросетевые архитектуры на основе МСП. Автоэнкодеры. Встроенные методы отбора признаков для МСП: ABHC, распад весов и т.д. Выбор оптимальных параметров МСП.

Занятие 9 (практикум). Решение обратных задач (ОЗ) в физике. Решение ОЗ спектроскопии.

Решение обратных задач в физике с помощью ИНС и других методов МО. Подходы от модели, от эксперимента и квазимодельный. Основные приёмы решения ОЗ спектроскопии.

Поиск выбросов, нормировки, сжатие данных (МГК, ПЛС, вейвлет-фильтрация), обучение базовой модели, оценка качества работы. Кластеризация и обучение моделей слабых регрессоров.

Занятие 10 (лекция). Глубокие и свёрточные НС.
Глубокие НС. Предобучение. Перенос обучения. Свёрточные НС.

Занятие 11 (лекция). Некоторые технологии работы с глубокими сетями.
Стратегии обучения нейронных сетей. Функция потерь для задач регрессии и классификации. Регуляризация. Dropout. Механизмы принятия решений нейронной сетью. "Сжатие" моделей.

Занятие 12 (лекция). Рекуррентные НС.
Рекуррентные НС. Сети Джордана-Элмана. Сети LSTM/GRU. Механизм внимания. Трансформеры.

Занятие 13 (лекция). МО и генерация данных
Генеративные состязательные сети. Вариационные автоэнкодеры. Генерация данных. Аугментация данных.

Раздел 4. Решение некоторых типов практических задач обработки данных в физике

Занятие 14 (практикум). Параметры элементарных частиц в камере Вильсона
Задачи сегментации\определения ключевых точек и их параметризация. Генерация фотореалистических изображений из параметризованных схематичных линий (pix2pix GAN?). Обучение на сгенерированном наборе данных.

Занятие 15 (лекция). Анализ временных рядов. Комбинированные алгоритмы.
Анализ временных рядов. Методы на основе скользящего среднего (ARIMA etc.). Погружение временного ряда. Комбинированные алгоритмы. Ансамбли. Бэггинг, бустинг, стекинг. Различные подходы к решению задач многоклассовой и многометочной классификации: one-vs-one, one-vs-all, ECOC и пр.

Занятие 16 (практикум). Решение задачи прогнозирования временного ряда в космической физике.
Стратегии обучения нейронных сетей для решения задач прогнозирования. Подготовка данных (нормировка, заполнение пропусков, устранение дневной периодичности и т.д.), обучение базовой модели, обучение рекуррентной сети, модификация целевой функции, обучение ансамбля модели, сравнение качества.