ВВЕДЕНИЕ В НАУКУ О ДАННЫХ И ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

Аннотация. Курс содержит популярное введение в науку о больших данных и основные методы искусственного интеллекта для их обработки. Он призван мотивировать студентов старших курсов для поступления в магистратуру "Наука о Данных и Искусственный Интеллект", но доступен также студентам младших курсов и самой широкой публике. Общая продолжительность курса составляет примерно 40 часов.

Введение. Почему Наука о Данных и Искусственный Интеллект?

Тема 1. Основные элементы анализа данных (8 ч.)

1.1. Подготовка данных
   Формат данных
   Типы переменных
   Выбор переменных
   Конструирование признаков
   Неполные данные
1.2. Выбор алгоритма
   Обучение без учителя
   Обучение с учителем
   Обучение с подкреплением
   Другие факторы
1.3. Настройка параметров
1.4. Оценка результатов
   Метрики классификации
   Метрика регрессии
   Валидация
1.5. Обзор темы

Контрольные вопросы:

1. Какие четыре ключевых шага предполагает исследование в рамках науки о данных?
2. В чем заключается этап подготовки данных?
3. Как осуществляется выбор алгоритмов для моделирования данных?
4. Как проходит настройка алгоритмов для оптимизации моделей?
5. Как оценивается точность моделей?

Тема 2. Кластеризация методом k-средних (2 ч.)

2.1. Поиск кластеров клиентов
2.2. Пример: профили кинозрителей
2.3. Определение кластеров
Сколько кластеров существует?
Что включают кластеры?
2.4. Ограничения
2.5. Обзор темы

Контрольные вопросы:

1. Что такое кластеризация методом k-средних?
2. Как определяется число кластеров k?
3. Какие два шага используются для группировки элементов данных?
4. Когда происходит остановка этих двух шагов алгоритма?
5. Для каких типов кластеров лучше всего работает кластеризация методом k-средних и почему?

Тема 3. Метод главных компонент (2 ч.)

3.1. Изучение пищевой ценности
3.2. Главные компоненты
3.3. Пример: анализ пищевых групп
3.4. Ограничения
3.5. Обзор темы

Контрольные вопросы:

1. В чем заключается метод главных компонент (МГК)?
2. Как используется МГК в анализе данных?
3. Что такое главная компонента?
4. Как используются главные компоненты для анализа и визуализации данных?
5. С какими информационными измерениями МГК работает лучше всего и почему?

Тема 4. Ассоциативные правила (2 ч.)

4.1. Поиск покупательских шаблонов
4.2. Поддержка, достоверность и лифт
4.3. Пример: ведение продуктовых продаж
4.4. Принцип A priori
Поиск товарных наборов с высокой поддержкой
Поиск товарных правил с высокой достоверностью или лифтом
4.5. Ограничения
4.6. Обзор темы

Контрольные вопросы:

1. Что выявляют ассоциативные правила?
2. Каковы три основных способа оценки ассоциации?
3. Что такое "поддержка"?
4. Что такое "достоверность"?
5. Что такое "лифт"?
6. В чем заключается принцип a priori?

Тема 5. Анализ социальных сетей (2 ч.)

5.1. Составление схемы отношений
5.2. Пример: геополитика в торговле оружием
5.3. Лувенский метод
5.4. Алгоритм PageRank
5.5. Ограничения
5.6. Обзор темы

Контрольные вопросы:

1. Как определяется метод для анализа социальных сетей?
2. Что такое Лувенский метод?
3. Когда лучше всего работает Лувенский метод?
4. Как работает алгоритм PageRank?
5. В чем преимущества и недостатки PageRank?

Тема 6. Регрессионный анализ (2 ч.)

6.1. Выведение линии тренда
6.2. Пример: предсказание цен на дома
6.3. Градиентный спуск
6.4. Коэффициенты регрессии
6.5. Коэффициенты корреляции
6.6. Ограничения
6.7. Обзор темы

Контрольные вопросы:

1. Для чего используется регрессионный анализ?
2. Как определяются предиктор и вес предиктора?
3. Что показывает вес предиктора?
4. Как выводится линия тренда?
5. В каких случаях регрессионный анализ работает лучше всего?

Тема 7. Метод k-ближайших соседей и обнаружение аномалий (4 ч.)

7.1. Пищевая экспертиза
7.2. Яблоко от яблони недалеко падает
7.3. Пример: истинные различия в вине
7.4. Обнаружение аномалий
7.5. Ограничения
7.6. Обзор темы

Контрольные вопросы:

1. Что представляет собой метод k-ближайших соседей?
2. Что такое кросс-валидация?
3. Как определяется число k в методе k-ближайших соседей?
4. Когда лучше всего работает метод k-ближайших соседей?
5. Что является верным признаком возможных аномалий?

Тема 8. Метод опорных векторов (2 ч.)

8.1 «Нет» или «о, нет!»?
8.2. Пример: обнаружение сердечно-сосудистых заболеваний
8.3. Построение оптимальной границы
8.4. Ограничения
8.5. Обзор темы

Контрольные вопросы:

1. Что такое опорный вектор?
2. Как работает метод опорных векторов (МОВ)?
3. Что такое функция ядра?
4. По отношению к каким значениям устойчив МОВ?
5. Когда лучше всего работает МОВ?

Тема 9. Дерево решений (2 ч.)

9.1. Прогноз выживания в катастрофе
9.2. Пример: спасение с тонущего «Титаника»
9.3. Создание дерева решений
9.4. Ограничения
9.5. Обзор темы

Контрольные вопросы:

1. Какой прогноз создает дерево решений?
2. В чем заключается процесс рекурсивного деления?
3. Каков критерий остановки рекурсивного деления?
4. Каковы достоинства и недостатки деревьев решений?
5. Какая альтернатива используется для преодоления недостатков этого подхода?

Тема 10. Случайные леса (2 ч.)

10.1. Мудрость толпы
10.2. Пример: предсказание криминальной активности
10.3. Ансамбли
10.4. Бэггинг
10.5. Ограничения
10.6. Обзор темы

Контрольные вопросы:

1. Что такое бэггинг?
2. Что такое ансамблирование?
3. Как случайные леса задействуют бэггинг и ансамблирование?
4. Сравните прогнозы деревьев решений и случайных лесов.
5. Как оцениваются предикторы случайных лесов?

Тема 11. Нейронные сети (6 ч.)

11.1. Создание мозга
11.2. Пример: распознавание рукописных цифр
11.3. Компоненты нейронной сети
11.4. Правила активации
11.5. Ограничения
11.6. Обзор темы

Контрольные вопросы:

1. Опишите общую архитектуру нейронных сетей. Как определяются нейроны, слои нейронов и их количество?
2. Как происходит активация нейронов первого слоя и что формируется в последнем слое нейронной сети?
3. Что такое правило активации?
4. Какой процесс называется методом обратного распространения ошибки?
5. Когда нейронные сети работают лучше всего?

Тема 12. A/B-тестирование и многорукие бандиты (2 ч.)

12.1. Основы A/B-тестирования
12.2. Ограничения A/B-тестирования
12.3. Стратегия снижения эпсилона
12.4. Пример: многорукие бандиты
12.5. Забавный факт: ставка на победителя
12.6. Ограничения стратегии снижения эпсилона
12.7. Обзор темы

Контрольные вопросы:

1. Какой вопрос решает проблема многорукого бандита?
2. Какая стратегия называется A/B-тестированием?
3. Что такое стратегия снижения эпсилона?
4. Какая из двух стратегий работает лучше и когда?

Приложения (4 ч.)

A. Обзор алгоритмов обучения без учителя
B. Обзор алгоритмов обучения с учителем
C. Список параметров настройки
D. Другие метрики оценки
Метрики классификации
Метрики регрессии

Итого: 40 ч.

Рекомендуемая литература

Основная:
Бринк Хенрик, Ричардс Джозеф, Феверолф Марк. Машинное обучение. — СПб.: Питер, 2018. — 336 с.: ил.
Бхаргава А. Грокаем алгоритмы. Иллюстрированное пособие для программистов и любопытствующих. — СПб.: Питер, 2018. — 288 с.: ил.
Винстон Уэйн. Бизнес-моделирование и анализ данных. Решение актуальных задач с помощью Microsoft Excel. 5-е издание. — СПб.: Питер, 2018. — 864 с.: ил.
Клеппман М. Высоконагруженные приложения. Программирование, масштабирование, поддержка. — СПб.: Питер, 2019. — 640 с.: ил.
Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. — СПб.: Питер, 2018. — 480 с.: ил.
Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — СПб.: Питер, 2018. — 576 с.: ил.
Седжвик Р., Уэйн К. Computer Science: основы программирования на Java, ООП, алгоритмы и структуры данных. — СПб.: Питер, 2018. — 1072 с.: ил.
Силен Дэви, Мейсман Арно, Али Мохамед. Основы Data Science и Big Data. Python и наука о данных. — СПб.: Питер, 2018. — 336 с.: ил.
Феррейра Фило Владстон. Теоретический минимум по Computer Science. Все, что нужно программисту и разработчику. — СПб.: Питер, 2019. — 224 с.: ил.
Шолле Франсуа. Глубокое обучение на Python. — СПб.: Питер, 2018. — 400 с.: ил.
Шолле Франсуа. Глубокое обучение на R. — СПб.: Питер, 2018. — 400 с.: ил.

Дополнительная:

Kosinski, M., Matz, S., Gosling, S., Popov, V., & Stillwell, D. (2015). Facebook as a Social Science Research Tool: Opportunities, Challenges, Ethical Considerations and Practical Guidelines. American Psychologist.
Agricultural Research Service, United States Department of Agriculture (2015). USDA Food Composition Databases.
Dataset is included in the following R package: Hahsler, M., Buchta, C., Gruen, B., & Hornik, K. (2016).
Stockholm International Peace Research Institute (2015).
Harrison, D., & Rubinfeld, D. (1993). Boston Housing Data.
Forina, M., et al. (1998). Wine Recognition Data.
Robert Detrano (M.D., Ph.D), from Virginia Medical Center, Long Beach and Cleveland Clinic Foundation (1988). Heart Disease Database (Cleveland).
British Board of Trade Inquiry (1990). Titanic Data.
SF OpenData, City and County of San Francisco (2016). Crime Incidents.
National Oceanic and Atmospheric Administration, National Centers for Environmental Information (2016). Quality Controlled Local Climatological Data (QCLCD).
LeCun, Y., & Cortes, C. (1998). The MNIST Database of Handwritten Digits.

Преподаватель: проф. Сергей Павлович Левашкин, заведующий Научно-исследовательской Лабораторией Искусственного Интеллекта, действительный член Мексиканской Академии Наук, выпускник МГУ им. М.В. Ломоносова, ученый с более чем 20-летним опытом работы в университетах и компаниях России, Северной Америки и Европы levashkin.com