ПРОЕКТ

ДОПОЛНИТЕЛЬНАЯ ОБЩЕОБРАЗОВАТЕЛЬНАЯ ОБЩЕРАЗВИВАЮЩАЯ ПРОГРАММА

«Большие данные»

Направленность: техническая

Возраст обучающихся: 15–18 лет

Содержание

Пояснительная записка

Введение

Элементы обучения основам обработки и анализа данных вводятся с первого полугодия 10 класса с постепенным усложнением содержания соответственно возрасту обучающегося и заканчиваются во втором полугодии 11-го класса.

программа составлена на основе:

- Федерального закона № 273-ФЗ от 29 декабря 2012 года «Об образовании в Российской Федерации»
- Федерального государственного образовательного стандарта среднего общего образования (Приказ Минобрнауки России от 17.05.2012 № 413 (ред. от 11.12.2020) «Об утверждении федерального государственного образовательного стандарта среднего общего образования» (Зарегистрировано в Минюсте России 07.06.2012 № 24480)
- Профессионального стандарта «06.001 Разработка программного обеспечения» Министерства юстиции Российской Федерации 18 декабря 2013 года, регистрационный № 30635)
- Профессионального стандарта «08.022 Статистическая деятельность» (Профессиональный стандарт Статистик, утверждённый приказом Министерства труда и социальной защиты РФ от 8 сентября 2015 г. № 605н)

Цели и задачи

Целями курса являются формирование у обучающегося аналитического мышления и, соответственно, знаний и умений, необходимых для успешного развития в отраслях, связанных со сложной аналитикой данных. Для достижения поставленных целей образование по данному направлению должно обеспечить решение следующих задач:

1) овладение реальными и практическими знаниями методов статистического анализа данных;
2) формирование навыков построения математических моделей (от нейронных сетей до кластеризации, от факторного до корреляционного анализа);
3) формирование навыков работы с большими массивами данных;
4) осознание практической важности нахождения уникальной закономерности в данных.

Состав участников образовательного процесса

Программа основного общего образования рассчитана на реализацию в 10– 11 классах общеобразовательных учреждений и учреждений с углубленным изучением отдельных предметов и нацелена на возрастную категорию учащихся 15–18 лет.

Общая характеристика учебного курса

Программа «Большие данные» предназначена для практического освоения учащимися работы с технологиями информационного поиска и обработки больших данных, работы с инструментами анализа данных, основ математической статистики и теории вероятностей, основ математического моделирования. Программа рассчитана на 2 года (10–11 класс), при этом обучение можно разделить на следующие модули:

- Модуль «Введение в вероятностное моделирование» содержит основы исчисления вероятностей, вероятностного анализа данных и начальные сведения о вероятностных моделях, использующихся для решения задач машинного обучения. В рамках модуля излагаются примеры применения изучаемых моделей, методов и алгоритмов, а также типовые алгоритмы решения задач реального мира с использованием вероятностных методов и моделей.
- Модуль «Анализ и визуализация данных на Python» предполагает изучение основных методов, подходов и инструментов для анализа и визуализации данных с использованием возможностей Python и его основных библиотек.
- Модуль «Параллельная обработка и управление большими данными» предполагает изучение теории баз данных, а также современных инструментов и технологий для решения задач, связанных с параллельной обработкой и анализом больших данных.
- Модуль «Машинное обучение» предполагает изучение основных методов и моделей машинного обучения, а также их реализацию на Python. В рамках модуля даются алгоритмы решения типовых задач машинного обучения с примерами вариантов их применения в реальных задачах.

Содержание курса направлено на формирование универсальных учебных действий, обеспечивающих развитие познавательных и коммуникативных качеств личности.

Обучающиеся включаются в проектную и исследовательскую деятельность, основу которой составляют такие учебные действия, как умение видеть закономерности в данных, строить на этих закономерностях модели, а также проводить валидацию и последующую доработку модели.

Требования к результатам освоения учебного курса

Деятельность образовательного учреждения в обучении по направлению «Большие данные» должна быть направлена на достижение обучающимися следующих **личностных результатов**:

- готовность и способность к самостоятельной, творческой и ответственной деятельности; навыки сотрудничества со сверстниками, детьми младшего возраста, взрослыми в образовательной, общественно полезной, учебно-исследовательской, проектной и других видах деятельности;
- готовность и способность к образованию, в том числе самообразованию, на протяжении всей жизни;
- сознательное отношение к непрерывному образованию как условию успешной профессиональной и общественной деятельности;

− осознанный выбор будущей профессии и возможностей реализации собственных жизненных планов; отношение к профессиональной деятельности как возможности участия в решении личных, общественных, государственных, общенациональных проблем.

Метапредметными результатами освоения программы по направлению «Большие данные» являются:

− умение самостоятельно определять цели и составлять планы в различных сферах деятельности, осознавая приоритетные и второстепенные задачи; самостоятельно осуществлять, контролировать и корректировать учебную, внеурочную и внешкольную деятельность с учётом предварительного планирования; использовать различные ресурсы для достижения целей; выбирать успешные стратегии в трудных ситуациях;

− умение продуктивно общаться и взаимодействовать с коллегами по совместной деятельности, учитывать позиции другого (совместное целеполагание и планирование общих способов работы на основе прогнозирования, контроль и коррекция хода и результатов совместной деятельности), эффективно разрешать конфликты;

− владение навыками исследовательской и проектной деятельности (определение целей и задач, планирование проведения исследования, формулирование гипотез и плана их проверки; осуществление наблюдений и экспериментов, использование количественных и качественных методов обработки и анализа полученных данных; построение доказательств в отношении выдвинутых гипотез и формулирование выводов; представление результатов исследования в заданном формате, составление текста отчёта и презентации с использованием информационных и коммуникационных технологий).

Предметными результатами освоения программы по направлению «Большие данные» являются:

− владение базовыми элементами теории вероятностей, методов математической статистики и методов машинного обучения;

− умение находить закономерности в данных, разрабатывать математические модели и модели машинного обучения на этих данных;

− умение выполнять численный анализ данных и визуализировать полученные результаты на языке Python;

− владение практическим опытом решения задач с применением методов математической статистики и машинного обучения.

Планируемые результаты изучения учебного курса

Выпускник научится:

1) владеть основами математической статистики и теории вероятностей;
2) находить закономерности в данных, разрабатывать математические модели и модели машинного обучения на этих данных;
3) визуализировать полученные результаты моделирования.

Учебно-тематический план

№ п/п	Модуль	Наименование раздела	Количество часов
1 полугодие 10 класса			
1.	Введение в вероятностное моделирование	Вводное занятие. Что такое математическая модель?	2
2.		Интуитивные понятия теории вероятностей	2
3.		Исчисление вероятностей и элементы комбинаторики. Текущий контроль	2
4.		Условная и полная вероятность	2
5.		Понятие случайной величины	2
6.		Обработка результатов наблюдений. Понятие статистической оценки. Текущий контроль	2
7.		Числовые оценки выборочных характеристик	2
8.		Вероятностные модели случайной величины	2
9.		Оценка параметров распределения случайной величины. Текущий контроль	2
10		Интервальные оценки и проверка статистических гипотез	2
11.		Базовые понятия из линейной алгебры	2
12.		Элементы многомерного статистического анализа и моделирования. Базовые элементы корреляционного анализа и регрессионного анализа. Текущий контроль	2
13.		Понятие классификации и кластеризации. Как связаны эти две задачи? Чем классификация отличается от регрессии?	2
14.		Понятие градиента. Текущий контроль	2
15.		Реализация итогового проекта	2
16.		Презентация результатов итогового проекта	2
2 полугодие 10 класса			
17.	Анализ и визуализация данных на Python	Анализ данных. Примеры и задачи	2
18.		Одномерный анализ данных. График функции. Гистограммы. Распределения	2
19.		Векторы и матрицы. Текущий контроль	2
20.		Введение в Python. Базовые операции	2
21.		Библиотека numpy. Примеры	2

22.		Библиотека pandas. Примеры. Текущий контроль	2
23.		Библиотека matplotlib. Примеры	2
24.		Понятие корреляции. Примеры на pandas и numpy	2
25.		Обучение с учителем. Примеры. Текущий контроль	2
26		Обучение без учителя. Примеры	2
27.		Кластеризация данных на Python	2
28.		Линейная контроль регрессия на Python. Текущий контроль	2
29.		Логистическая регрессия на Python	2
30.		Работа с изображениями в Python. Текущий контроль	2
31.		Реализация итогового проекта	2
32.		Презентация результатов итогового проекта	2
1 полугодие 11 класса			
33.	Параллельная обработка и управление большими данными	Понятие обработки данных. Виды обработки данных. Виды баз данных	2
34.		Типы данных, таблицы и отношения между ними. Реляционная модель данных	2
35.		Введение в SQL. Примеры в PostgreSQL. Текущий контроль	2
36.		Понятие индекса. Виды индексов	2
37.		Проектирование баз данных. Цели проектирования. Нормализация данных. Проектирование базы данных в PostgreSQL	2
38.		Текстовые и бинарные форматы хранения данных json, csv, parquet. Текущий контроль	2
39.		Обработка данных в памяти. Продвинутый pandas. Знакомство с ataframe’ами. Примеры	2
40.		Колоночные базы данных (NoSQL для больших данных): HBase, ClickHouse	2
41.		Основные понятия распределенной обработки данных. Текущий контроль	2
42.		Знакомство с Apache Spark (PySpark)	2
43.		Парадигма MapReduce. Сравнение с Hadoop	2
44.		Параллельная и распределенная обработка больших данных средствами PySpark	2

45.		Разработка итогового проекта. Постановка задачи организации хранения и обработки данных. Текущий контроль	2
46.		Проектирование хранилища и процесса обработки данных	2
47.		Реализация итогового проекта	2
48.		Презентация результатов итогового проекта	2
2 полугодие 11 класса			
49.	Введение в машинное обучение	Введение в машинное обучение.	2
50.		Типология и метрики качества алгоритмов машинного обучения	2
51.		Метрические алгоритмы классификации. Текущий контроль	2
52.		Логические алгоритмы классификации	2
53.		Введение в ансамблевые методы	2
54.		Модели смесей распределений. Текущий контроль	2
55.		Методы кластеризации и детектирования аномалий	2
56.		Методы снижения размерности многомерных данных	2
57.		Обучение с подкреплением. Текущий контроль	2
58.		Введение в нейронные сети	2
59.		Многослойный перцептрон	2
60.		Свёрточные нейронные сети. Текущий контроль	2
61.		Рекуррентные нейронные сети	2
62.		Глубокое обучение без учителя. Текущий контроль	2
63.		Постановка задачи для итогового проекта. Разработка итогового проекта	2
64		Презентация результатов итогового проекта	2
Итого:			**128**

Примечание. Разделы, относящиеся к одному модулю, могут быть реализованы в различных полугодиях. В том числе, возможно параллельное изучение материала нескольких модулей, если это обосновано логикой освоения материала.

Содержание курса

1-е полугодие 10 класса

ВВОДНОЕ ЗАНЯТИЕ. ЧТО ТАКОЕ МАТЕМАТИЧЕСКАЯ МОДЕЛЬ? Основные принципы и подходы к моделированию. Классификация математических моделей. Бытовое понятие о вероятности.

ИНТУИТИВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ. Элементарные события. Случайные события. Алгебра событий. Примеры. Упражнения.

ИСЧИСЛЕНИЕ ВЕРОЯТНОСТЕЙ И ЭЛЕМЕНТЫ КОМБИНАТОРИКИ. Сложение вероятностей совместных и несовместных событий. Перестановки, выборки и сочетания. Примеры. Упражнения. Текущий контроль.

УСЛОВНАЯ И ПОЛНАЯ ВЕРОЯТНОСТЬ. Понятие условной вероятности. Формула Байеса. Теорема о полной вероятности. Упражнения.

ПОНЯТИЕ СЛУЧАЙНОЙ ВЕЛИЧИНЫ. Дискретная случайная величина. Схемы повторения испытаний. Формула Пуассона. Законы распределения дискретной случайной величины. Непрерывная случайная величина. Функция распределения. Плотность распределения. Равномерный закон распределения. Нормальный закон распределения. Упражнения.

ОБРАБОТКА РЕЗУЛЬТАТОВ НАБЛЮДЕНИЙ. ПОНЯТИЕ СТАТИСТИЧЕСКОЙ ОЦЕНКИ. Что такое статистические оценки и чем занимается математическая статистика? Эмпирическая функция распределения. Принципы построения гистограмм. Ядерная оценка плотности распределения. Упражнения. Текущий контроль.

ЧИСЛОВЫЕ ОЦЕНКИ ВЫБОРОЧНЫХ ХАРАКТЕРИСТИК. Выборочное среднее. Выборочная дисперсия. Выборочное среднее квадратическое отклонение. Упражнения.

ВЕРОЯТНОСТНЫЕ МОДЕЛИ СЛУЧАЙНОЙ ВЕЛИЧИНЫ. Выбор функции распределения как вероятностной модели случайной величины. Вероятностная модель как смесь распределений. Смесь распределений Гаусса. Примеры. Упражнения.

ОЦЕНКА ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ СЛУЧАЙНОЙ ВЕЛИЧИНЫ. Точечные оценки параметров распределения. Метод максимального правдоподобия. Упражнения. Текущий контроль.

ИНТЕРВАЛЬНЫЕ ОЦЕНКИ И ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ. Понятие доверительного интервала как модели для ошибки оцененных параметров. Понятие статистической гипотезы и критериев для ее проверки. Непараметрические и параметрические критерии. Упражнения.

БАЗОВЫЕ ПОНЯТИЯ ИЗ ЛИНЕЙНОЙ АЛГЕБРЫ. Понятие вектора и матрицы. Операции над матрицами. Матричные произведения. Специальные виды матриц. Обратная матрица. Понятие системы линейных алгебраических уравнений (СЛАУ). Методы решения СЛАУ (обзорно). Метод Гаусса. Упражнения.

ЭЛЕМЕНТЫ МНОГОМЕРНОГО СТАТИСТИЧЕСКОГО АНАЛИЗА И МОДЕЛИРОВАНИЯ. БАЗОВЫЕ ЭЛЕМЕНТЫ КОРРЕЛЯЦИОННОГО АНАЛИЗА И РЕГРЕССИОННОГО АНАЛИЗА. Выборочные коэффициенты корреляции. Корреляционная матрица. Уравнение прямой и задача регрессии. Множественная регрессия. Оценка качества регрессионной модели. Понятие вероятностного интервала. Примеры. Упражнения. Текущий контроль.

ПОНЯТИЕ КЛАССИФИКАЦИИ И КЛАСТЕРИЗАЦИИ. КАК СВЯЗАНЫ ЭТИ ДВЕ ЗАДАЧИ? ЧЕМ КЛАССИФИКАЦИЯ ОТЛИЧАЕТСЯ ОТ РЕГРЕССИИ? Расстояние между объектами. Типы кластеров. Методы кластеризации (обзорно). Метод ближайшего соседа. Метод k-means. Кластеризация с помощью вероятностных моделей: разделение смеси Гауссовых распределений (дискриминантный анализ). Методы классификации. Логистическая регрессия. Упражнения.

ПОНЯТИЕ ГРАДИЕНТА. Использование градиента в задачах оптимизации и машинного обучения. Метод градиентного спуска. Стохастический градиентный спуск. Примеры. Упражнения. Текущий контроль.

РЕАЛИЗАЦИЯ ИТОГОВОГО ПРОЕКТА. Постановка задачи по построению одномерной вероятностной модели. Рекомендации по выполнению.

ПРЕЗЕНТАЦИЯ РЕЗУЛЬТАТОВ ИТОГОВОГО ПРОЕКТА.

2-е полугодие 10 класса

АНАЛИЗ ДАННЫХ. ПРИМЕРЫ И ЗАДАЧИ. Какие бывают данные. Понятия числовых, категориальных данных. Способы представления информации. Основные задачи анализа данных: классификация, регрессия, кластеризация (повторение).

ОДНОМЕРНЫЙ АНАЛИЗ ДАННЫХ. ГРАФИК ФУНКЦИИ. ГИСТОГРАММЫ. РАСПРЕДЕЛЕНИЯ. Понятие функции и аргумента. Зависимость и независимость. Построение графика функции по табличным значениям. Понятие гистограммы как способа представления табличных данных, примеры (повторение). Понятие распределения (повторение) и способы визуализации различных распределений.

ВЕКТОРЫ И МАТРИЦЫ. Понятие вектора и понятие матрицы и их физический смысл. Размерность матриц. Связь матриц и таблиц данных. Линейная зависимость и линейная независимость векторов. Понятия коллинеарности и компланарности. Основные операции над матрицами и векторами: сложение, скалярное произведение, умножение матриц и их физический смысл (повторение и углубление). Транспонирование матриц. Обратные матрицы. Системы линейных алгебраических уравнений и способы их решения (повторение и углубление). Текущий контроль.

ВВЕДЕНИЕ В PYTHON. БАЗОВЫЕ ОПЕРАЦИИ. Базовые типы данных в Python**:** численные, строковые, логические переменные. Циклы. Функции. Структуры данных в Python: списки, множества и словари – примеры создания и основные операции с ними. Понятие list comprehension. Пример реализации функции одной переменной. Импорт модулей и функций.

БИБЛИОТЕКА NUMPY. ПРИМЕРЫ. Основные конструкции библиотеки numpy как библиотеки для высокопроизводительных вычислений. Векторизация вычислений. Создание массивов, одномерные и многомерные массивы. Вычисление основных статистических показателей матрицы с помощью numpy: минимум, максимум, среднее, argmax и др. Примеры. Текущий контроль.

БИБЛИОТЕКА PANDAS. ПРИМЕРЫ. Основные конструкции библиотеки pandas. Чтение файлов и запись в файл. Понятие pandas.DataFrame и pandas.Series. Выгрузка данных по условию. Создание таблиц. Агрегация и слияние имеющихся данных. Выполнение сложных запросов к датасету.

БИБЛИОТЕКА MATPLOTLIB. ПРИМЕРЫ. Библиотека matplotlib и визуализация данных. Построение графика функции и создание своего стиля для графика. Линейные и логарифмические шкалы, выбор масштаба представления данных. Гистограммы в matplotlib. Примеры построения нескольких независимых графиков в одном окне: метод subplots. Сохранение графиков в виде изображения. Текущий контроль.

ПОНЯТИЕ КОРРЕЛЯЦИИ. ПРИМЕРЫ НА PANDAS И NUMPY. Понятие корреляции (повторение). Ложные корреляции. Виды зависимостей данных друг от друга. Понятие кросс-корреляции, автокорреляции и свёртки. Понятие ранговых списков. Корреляция Пирсона и корреляция Спирмена. Вычисление попарных корреляций и корреляционных таблиц средствами numpy и pandas. Heatmap и графическое представление таблиц данных.

ОБУЧЕНИЕ С УЧИТЕЛЕМ. ПРИМЕРЫ. Задача обучения с учителем. Обучение по прецедентам. Объекты и целевые переменные. Понятие функции ошибок. Тренировочная и тестовая выборка. Задачи классификации и регрессии – сходства и различия. Данные для обучения в виде таблиц значений. Текущий контроль.

ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ. ПРИМЕРЫ. Задачи обучения без учителя и data mining. Обзор: кластеризация, корреляционный анализ, понижение размерности. Алгоритмы кластеризации (повторение).

КЛАСТЕРИЗАЦИЯ ДАННЫХ НА PYTHON. Библиотека scikit-learn и её использование для кластеризации данных в Python. Изменение параметров методов кластеризации и проверка качества кластеризации. Метрики кластеризации и их реализация в Python. Реализация алгоритмов KNN, SVM и Kmeans в библиотеке scikit-learn. Примеры и визуализация. Текущий контроль.

ЛИНЕЙНАЯ РЕГРЕССИЯ НА PYTHON. Понятие линейной регрессии (повторение). Понятие весовых коэффициентов и настройка параметров модели. Отбор признаков и работа с данными. Скалирование и центрирование данных. Недообучение и переобучение. Понятие регуляризации. Регуляризация модели линейной регрессии – подходы Lasso и Ridge и их отличия. Случай нелинейной зависимости, полиномиальная регрессия. Примеры и упражнения на Python.

ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ НА PYTHON. Логистическая регрессия как модель бинарной классификации. Целевая переменная и виды функции ошибок для задач классификации. Понятие функции активации и её виды: линейная, сигмоида, гиперболический тангенс, ReLU. Примеры и упражнения на Python. Текущий контроль.

РАБОТА С ИЗОБРАЖЕНИЯМИ В PYTHON. Изображение как матрица. Понятие RGB изображения и примеры других цветовых пространств. Понятие яркости и контраста. Основы обработки изображений: фильтрация, бинаризация, выделение границ, размытие. Загрузка изображений в Python и использование библиотеки matplotlib для работы с изображениями. Примеры и упражнения по обработке изображений в Python.

ПОДГОТОВКА ИТОГОВОГО ПРОЕКТА. Реализация итогового проекта по построению модели бинарной классификации в Python с помощью библиотек numpy, pandas, scikit-learn и matplotlib.

ПРЕЗЕНТАЦИЯ РЕЗУЛЬТАТОВ ИТОГОВОГО ПРОЕКТА

1-е полугодие 11 класса

ПОНЯТИЕ ОБРАБОТКИ ДАННЫХ. ВИДЫ ОБРАБОТКИ ДАННЫХ. ВИДЫ БАЗ ДАННЫХ. Обработка цифровой, символьной, текстовой и табличной информации. Реляционные и NoSQl базы данных, их отличия, области применения, примеры использования.

ТИПЫ ДАННЫХ, ТАБЛИЦЫ И ОТНОШЕНИЯ МЕЖДУ НИМИ.

РЕЛЯЦИОННАЯ МОДЕЛЬ ДАННЫХ. Строковые, целочисленные, дробные, дата, время. Понятие таблицы, ключа. Нормальные формы. Ключи, первичные и внешние ключи.

ВВЕДЕНИЕ В SQL. ПРИМЕРЫ В POSTGRESQL. Создание таблиц, вставка, выборка, удаление, изменение данных. Создание ключей на колонки. Выборка данных из таблиц, фильтрация, сортировки, группировки, слияние, подзапросы. Текущий контроль.

ПОНЯТИЕ ИНДЕКСА. ВИДЫ ИНДЕКСОВ. Индексы в базах данных: назначение, влияние на производительность, принципы создания индексов. Индексы: по порядку сортировки, источнику данных, воздействию на источник данных, структуре, количественному составу, характеристике содержимого, механизму обновления, покрытию индексируемого содержимого.

ПРОЕКТИРОВАНИЕ БАЗ ДАННЫХ. ЦЕЛИ ПРОЕКТИРОВАНИЯ. НОРМАЛИЗАЦИЯ ДАННЫХ. ПРОЕКТИРОВАНИЕ БАЗЫ ДАННЫХ В POSTGRESQL. Основные задачи и этапы проектирования баз данных. Концептуальное, логическое, физическое проектирование.
ТЕКСТОВЫЕ И БИНАРНЫЕ ФОРМАТЫ ХРАНЕНИЯ ДАННЫХ JSON, CSV, PARQUET. ОБРАБОТКА ДАННЫХ В ПАМЯТИ. ПРОДВИНУТЫЙ PANDAS. ЗНАКОМСТВО С DATAFRAME'АМИ. ПРИМЕРЫ. Чтение данных, обработка и запись в различные форматы.

КОЛОНОЧНЫЕ БАЗЫ ДАННЫХ (NOSQL ДЛЯ БОЛЬШИХ ДАННЫХ): HBASE, CLICKHOUSE. Понятие колонки, способ представления, отличия от строкового представления. Достоинства и недостатки КБД.

ОСНОВНЫЕ ПОНЯТИЯ РАСПРЕДЕЛЕННОЙ ОБРАБОТКИ ДАННЫХ. АРХИТЕКТУРА «КЛИЕНТ-СЕРВЕР», «МАСТЕР-ВОРКЕР». Достоинства и недостатки распределенной обработки данных. Способы распределения данных: централизованный, децентрализованный, смешанный.

ЗНАКОМСТВО С APACHE SPARK (PYSPARK). Компоненты экосистемы Apache Spark, Особенности Apache Spark, RDD и особенности использования, трансформации и действия.

ПАРАДИГМА MAPREDUCE. СРАВНЕНИЕ С HADOOP. Понятие shuffle, виды реализации shuffle в spark.

ПАРАЛЛЕЛЬНАЯ И РАСПРЕДЕЛЕННАЯ ОБРАБОТКА БОЛЬШИХ ДАННЫХ СРЕДСТВАМИ PYSPARK. Знакомство со Spark-shell. Написание программ в Apache Spark, чтение и запись данных. Понятие DataFrame. Использование DataFrame вместо RDD, простые запросы, фильтрация и агрегация. Продвинутые операции: join, broadcast, udf, udaf.

РАЗРАБОТКА ИТОГОВОГО ПРОЕКТА. ПОСТАНОВКА ЗАДАЧИ ОРГАНИЗАЦИИ ХРАНЕНИЯ И ОБРАБОТКИ ДАННЫХ.
Проектирование хранилища и процесса обработки данных
Программная реализация проекта
Презентация результатов итогового проекта

2-е полугодие 11 класса

ВВЕДЕНИЕ В МАШИННОЕ ОБУЧЕНИЕ. Классификация моделей искусственного интеллекта по Расселу и Норвигу. Имитация когнитивных функций человека современными моделями машинного обучения. Определения машинного обучения. Опыт, задача, качество решения. Способы задания входных данных для алгоритма машинного обучения. Обобщающая способность модели. Дилемма смещения-разброса, понятие недообучения и переобучения.

ТИПОЛОГИЯ И МЕТРИКИ КАЧЕСТВА АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ. Обучение с учителем (карта методов). Разметка данных. Функции потерь. Тренировочная, тестовая (контрольная), валидационная (проверочная) выборка. Кросс-валидация. Метрики качества бинарной классификации. ROC-AUC и Precision-Recall кривые. Метрики качества для несбалансированных выборок. Обучение без учителя. Метрики качества для оценки результатов кластеризации. Модулярность. Коэффициент силуэта. Semi-supervised обучение. Обучение с подкреплением.

МЕТРИЧЕСКИЕ МЕТОДЫ КЛАССИФИКАЦИИ. Метрики расстояния (Манхэттенская, евклидово расстояние, косинусное расстояние). Метод ближайших соседей. Подбор числа соседей. Метод опорных векторов для случая линейно разделимой выборки.

ЛОГИЧЕСКИЕ КЛАССИФИКАТОРЫ. Решающие правила. Конструирование решающих правил. Решающие деревья. Метрики информативности. Подрезка решающих деревьев.

ВВЕДЕНИЕ В АНСАМБЛЕВЫЕ МЕТОДЫ. Бэггинг. Случайный лес. Бустинг. Алгоритм AdaBoost.

МОДЕЛИ СМЕСЕЙ РАСПРЕДЕЛЕНИЙ. Обучение смеси гауссианов. EM-алгоритм. Алгоритмы тематического моделирования. Вероятностный латентно-семантический анализ.

МЕТОДЫ КЛАСТЕРИЗАЦИИ И ДЕТЕКТИРОВАНИЯ АНОМАЛИЙ. Алгоритм к-средних (англ. k-means). Применение EM-алгоритма для алгоритма к-средних. Иерархическая кластеризация. Интерпретация дендрограмм. Кластеризация на графах. Методы детектирования аномалий.

МЕТОДЫ СНИЖЕНИЯ РАЗМЕРНОСТИ МНОГОМЕРНЫХ ДАННЫХ.

Многомерное шкалирование. Метод главных компонент. Методы обучения представлений для текстовых данных. Методы обучения представлений для графовых данных.

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ. Агент и среда. Система подкрепления. Способы обучения с подкреплением. Задача о многоруком бандите. Q-обучение.

ВВЕДЕНИЕ В НЕЙРОННЫЕ СЕТИ. Типология нейронных сетей. Однослойные модели нейронных сетей. Правило Хебба. Карты Кохонена.

МНОГОСЛОЙНЫЙ ПЕРЦЕПТРОН. Алгоритм обратного распространения ошибки. Способы борьбы с переобучением для нейронных сетей.

СВЁРТОЧНЫЕ НЕЙРОННЫЕ СЕТИ. Принцип построения иерархических признаков. Архитектура сверточной нейронной сети. Слои свертки и субдискретизации. Реализация операций пулинга. Современные архитектуры сверточных нейросетей: ImageNet, VGG16.

РЕКУРРЕНТНЫЕ НЕЙРОННЫЕ СЕТИ. Архитектура сети RNN. Архитектура сети LSTM. Архитектура сети GRU. Примеры использования рекуррентных сетей в области машинного перевода и прогнозирования временных рядов.

ГЛУБОКОЕ ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ. Ограниченная машина Больцмана. Автоэнкодеры. Сети глубокого доверия. Генеративно-состязательная сеть.

ПОСТАНОВКА ЗАДАЧИ ДЛЯ ИТОГОВОГО ПРОЕКТА. РАЗРАБОТКА ИТОГОВОГО ПРОЕКТА.

ПРЕЗЕНТАЦИЯ РЕЗУЛЬТАТОВ ИТОГОВОГО ПРОЕКТА