Большие данные

д.т.н., профессор
Холод Иван Иванович
iiholod@mail.ru

Введение в курс.

Цель и задачи курса

Цель: формирование представления о развитии средств и методов обработки и анализа Больших данных.

Задачи:

- изучение знаний по базовым методам и алгоритмам обработки и анализа Больших данных и их усовершенствования для выполнения в параллельной и распределенной среде;
- формирование умений и практических навыков разработки алгоритмического и программного обеспечения методов анализа Больших данных;
- освоение навыков применения методов и алгоритмов анализа Больших данных;

Структура курса

1. введение в курс;
2. поколения платформ данных;
3. хранение Больших данных;
4. обработка потоковых данных ;
5. распределенная обработка данных;
6. алгоритмы анализа Больших данных;
7. федеративное обучение;
8. алгоритмы федеративного обучения;

Анализ Больших данных. On-line

- **Модуль 1. Введение в анализ Больших данных** (Тест 20 вопросов):
  - Блок 1. Большие данные
  - Блок 2. Методы анализа данных
  - Блок 3. Поколения платформ Больших данных
  - Блок 4. Анализ распределенных данных
- **Модуль 2. Хранение Больших данных** (Тест 20 вопросов):
  - Блок 1. Реляционные (SQL) БД
  - Блок 2. noSQL БД
  - Блок 3. newSQL БД
  - Блок 4. Распределенные хранилища данных
- **Модуль 3. Распределенный анализ Больших данных** (тест 20 вопросов)
  - Блок 1. Масштабирование вычислений для Больших данных
  - Блок 2. Параллельное обучение
  - Блок 3. Распределенные вычисления
  - Блок 4. Парадигма Map Reduce
  - Блок 5. Системы распределенного анализа данных
- **Модуль 4. Обработка потоковых данных** (Тест 20 вопросов)
  - Блок 1. Проблемы потоковой обработки
  - Блок 2. Требования к потоковой обработке
  - Блок 3. Архитектура
- **Модуль 5. Федеративное обучение** (Тест 20 вопросов)
  - Блок 1. Новая концепция анализа распределенных данных
  - Блок 2. Распределение данных
  - Блок 3. Классы систем федеративного обучения
  - Блок 4.1 Безопасность федеративного обучения. Виды атак.
  - Блок 4.2 Безопасность федеративного обучения. Методы защиты
  - Блок 5. Библиотеки федеративного обучения
  - Блок 6. Применение федеративного обучения

Структура курса

	н1	н2	н3	н4	н5	н6	н7	н8	н9	н10	н11	н12	н13	н14	н15	н16	н17
Модули	0	1	-	2	2	2	3	3	3	3	4	4	5	5	5	5	5
Доклады					1	1		2	2	3		4		5	5	6	6
Практика	к	к	к	1	1	1	к	к	2	2	2	к	к	3	3	3	3

Практические занятия:

•теоретическая часть

- on-line лекции (LETITech)
- доклады на парах
  1. системы хранения Больших данных
  2. системы распределенной обработки Больших данных
  3. алгоритмы распределенных данных
  4. системы обработки потоковых данных
  5. фреймворки Федеративного обучения
  6. алгоритмы федеративного обучения

•практическая часть

- теоретическое проектирование
- практическая реализация.

Практика. Треки

1. **Основной трек**: Анализ Больших данных. Apache Spark в Yandex.Cloud. DataSphere

   Выбор: набора данных, задачи анализа, алгоритма из Apache Spark MLlib

   Последовательный анализ набора данных

   Параллельный анализ набора данных (2, 8, 32 процессора)

2. **Альтернативный трек**: Федеративное обучение

   2.1. Анализ с использование Python FL framework (TFF, FATE, PySyft, Flower, FedML)

- простой набор данных – пр.практика (ЛЭТИ)

   2.2. Анализ с использованием FL4J – пр.практика (ЛЭТИ)

   2.3. Разработка алгоритмов FL для FL4J – пр.практика (ЛЭТИ)

   2.4. Доработка фреймворка FL4J – пр.практика (ЛЭТИ)

   2.5. Доработка GUI (Vue.js) для FL4J – пр.практика (ЛЭТИ)

Основной трек. Анализ данных

Реализация аналитической задачи методом машинного обучения на данных большого объема алгоритмом из модуля Apache Spark MLlib в Yandex.Cloud и Data Shpere

Группы по 2 человека:

-инженер данных;

-ML аналитик;

На выбор:

-набор данных;

- решаемая задача

- используемые ML алгоритм из модуля Apache Spark MLlib;

Порядок выполнения:

1. Теоретическое проектирование: **Доклад №1**
2. Практическая реализация: Доклад №2 по результатам
    1. Последовательное выполнение ML алгоритма
    2. Параллельное выполнение ML алгоритма

Альтернативный трек. Фед. обучение

Реализация аналитической задачи методом федеративного обучения на данных большого объема в Yandex.Cloud с использование ВМ

Группы по 2 человека:

- инженер;
- FL аналитик;

На выбор:

- набор данных (три набора данных);
- фреймворк;
- FL алгоритм;

Порядок выполнения:

1. Теоретическое проектирование: **Доклад №1**
2. Практическая реализация: Доклад №2 по результатам
    1. **выполнение FL алгоритма на одном узле**
    2. **выполнение FL алгоритма не нескольких узлах**

План доклада №1

Подходы к анализу данных

1. Общая информация по данным: источник, кем предоставлены, когда и для каких задач могут использоваться.
2. Описание целевой задачи анализа данных исходя из данных
3. Метаинформация: формат, количество атрибутов и векторов, типы атрибутов, классы и т.п.
4. Ограничения данных: пропущенные значения, аномалии, и т.п..
5. Предлагаемый ML алгоритм (**из модуля Apache Spark MLlib**) для решения целевой задачи
6. Необходимые настройки данных для каждого алгоритма.
7. Ожидаемые модели знаний, построенные алгоритмами.
8. Предлагаемые методы и критерии оценки построенных моделей.

План доклада №2

Результаты анализа данных

1. Процесс анализа: выполненные этапы анализа и итерации;
2. Настройки/преобразования данных (привести фрагменты физических и логических данных).
3. Настройки функций (привести скриншот).
4. Настройки алгоритма (привести скриншоты).
5. Построение моделей алгоритмом: время построения в зависимости от объема данных;
6. Построение модели : время построения в зависимости от числа вычислителей;
7. Построенные модели и их оценки.
8. Выводы.

Структура пояснительной записки

1. Общая информация по данным: источник, кем предоставлены, когда и для каких задач могут использоваться.
2. Описание целевой задачи анализа данных
3. Метаинформация: формат, количество атрибутов и векторов, типы атрибутов, классы и т.п.
4. Ограничения данных: пропущенные значения, аномалии, и т.п..
5. Предлагаемые алгоритмы Data Mining для решения целевой задачи (не менее 3х алгоритмов)
6. Необходимые настройки данных для каждого алгоритма.
7. Ожидаемые модели знаний, построенные алгоритмами.
8. Предлагаемые методы и критерии оценки построенных моделей.
9. Выбранная среда для анализа: разработчик, лицензия, версия, и т.п.
10. Процесс анализа: выполненные этапы анализа и итерации;
11. Настройки/преобразования данных (привести фрагменты физических и логических данных).
12. Настройки функций (привести скриншот).
13. Настройки каждого алгоритма (привести скриншоты).
14. Построение моделей каждым алгоритмом: время построения в зависимости от числа данных;
15. Построенные модели (каждым алгоритмом) и их оценки.
16. Выводы.

Оценка знаний

Оценка по дисциплине формируется из:

1. оценки за теоретическую часть (минимум 10, максимум >40 баллов);
2. оценки за практическую часть (минимум 20, максимум 40 баллов).

Допуск к экзамену если выполнены два условия:

1. выполнена практическая часть ≥ 20;
2. пройдены тесты в on-line курсе ≥ 10 (правильные ответы минимум 50%).

Итоговая оценка вычисляется следующим образом:

1. 5 если набрано баллов ≥ 80 баллов;
2. 4 если 60 < набрано баллов < 80;
3. 3 если 40 < набрано баллов ≤ 60;
4. не аттестован если ≤ 40 баллов

Оценка знаний. Теоретическая часть

Оценка **за теоретическую часть** может быть получена:

1. тесты в on-line курсе – минимум 10, максимум 20 баллов (*1 балл за 5 ответов*)
2. за ответы на вопросы на лекциях;
3. за доклад - максимум 10 баллов;
4. за оппонирование докладов - максимум 5 баллов
5. за экзамен - максимум 10 баллов за вопрос.

Оценка знаний. Практическая часть

Оценка **за практическую часть** может быть получена за доклады и выполнение заданий:

- доклад 1 – максимум 10 баллов
- оклад 2 – максимум 20 баллов:
  - последовательный алгоритм (10 баллов)
  - масштабированный алгоритм (10 баллов)

Презентации присылаются на проверку за 1 неделю до доклада.

График сдачи работы:

Доклад	Срок	Штраф
Теоретическое проектирование	20.03.2023	3 балла до 20.04, 6 после
Последовательное выполнение алгоритма	24.04.2023	3 балла до 22.05, 6 после
Параллельное выполнение алгоритма	19.05.2023	5 баллов после 20.05

10 дополнительных баллов за выполнения ВСЕЙ практики до 01.05

Вопросы?