Программа учебной дисциплины

«Современные методы принятия решений:

Алгоритмы обработки больших данных»

Утверждена

Академическим советом ООП

Протокол № 2.3-09/ 2706-01 от «27» июня 2018

Автор	Зимовнов Андрей Вадимович
Число кредитов	4
Контактная работа (час.)	38
Самостоятельная работа (час.)	114
Курс	1
Формат изучения дисциплины	Без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целью освоения дисциплины является ознакомление студентов с основными задачами машинного обучения на больших данных, их особенностями и ограничениями.

В результате освоения дисциплины студент должен:

- знать особенности распараллеливания алгоритмов машинного обучения для применения на больших данных;
- владеть инструментами обработки данных в парадигме MapReduce;
- уметь работать с большими данными в реальных задачах.

Для освоения учебной дисциплины студенты должны владеть знаниями и компетенциями следующих дисциплин:

- Математический анализ
- Линейная алгебра и геометрия
- Теория вероятностей
- Математическая статистика
- Алгоритмы и структуры данных
- Введение в машинное обучение

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

1. Онлайн обучение и линейные модели.

   Онлайн подход к обучению на больших данных на примере линейных моделей. Разбор принципов работы vowpal wabbit. Progressive validation, трюк

с хэшированием. Запуск обучения на кластере. Разбор задачи предсказания кликов для онлайн-рекламы Criteo.

2. Введение в Apache Spark и оптимизация гиперпараметров Обобщение парадигмы Map-Reduce, модель графов вычислений, RDD, DataFrame API, Mllib. Простейшее использование Apache Spark для оптимизации гиперпараметров.
3. Рекомендательные системы.
   Особенности построения рекомендательных систем на больших данных. Content-based, collaborative filtering, ALS, iALS.
4. Бустинг для больших данных (1 лекция. 1 семинар)
   Подходы к распараллеливанию бустинга над решающими деревьями. Обзор реализации xgboost.
5. Введение в TensorFlow
   Вычислительная модель TensorFlow, примеры графов дифференцируемых вычислений для различных задач. Почему GPU дает ускорение. Рассмотрение задачи расчета word2vec представлений для слов.
6. Глубокие нейронные сети для классификации изображений
   Обзор основных архитектур, датасет ImageNet, особенности сверточных сетей. Эффективное использование нескольких GPU, использование нескольких машин с GPU.
7. Глубокие нейронные сети для классификации текстов
   Основы рекурентных нейросетей. Задача определения интента фразы по тексту. Распараллеливание обучения.
8. LSH для нахождения похожих объектов
   Нахождение похожих объектов на примере меры Жаккара. LSH на примере задачи нахождения похожих новостей.
9. Кластеризация больших данных
   Распределенный вариант алгоритма K-Means.

III. ОЦЕНИВАНИЕ

Курс подразумевает следующие способы контроля знаний:

- Домашние работы
- Проверочные на семинарах (письменные)
- Экзамен (в конце курса, письменный).

Результирующая оценка по дисциплине рассчитывается по формуле:

$$O_{\text{итог}} = 0.7\ O_{\text{накопл}} + 0.3\ O_{\text{экз}}$$

Накопленная и итоговая оценки округляются арифметически.

Накопленная оценка рассчитывается по формуле:

$$O_{\text{накопл}} = 0.3\ O_{\text{самост}} + 0.7\ O_{\text{дз}}$$

Оценка за домашние задания рассчитывается как среднее значение оценок за все выданные домашние задания. Оценка за самостоятельную работу рассчитывается как среднее значение оценок за все проверочные работы, проведённые на семинарских занятиях.

Студенту, получившему отличную накопленную оценку, данная оценка может быть выставлена в качестве итоговой на усмотрение семинариста и лектора.

Блокирующие элементы оценки отсутствуют.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Примеры экзаменационных вопросов:

- Онлайн обучение и линейные модели.
- Особенности построения рекомендательных систем на больших данных.
- Подходы к распараллеливанию бустинга над решающими деревьями.
- Вычислительная модель TensorFlow, задача word2vec.
- Глубокие нейронные сети для классификации изображений.
- Глубокие нейронные сети для классификации текстов.
- LSH для нахождения похожих объектов для меры Жаккара.
- Распределенный вариант алгоритма K-Means.

V. РЕСУРСЫ

1. Основная литература

- Jure Leskovec, Anand Rajaraman, Jeff Ullman. Mining of Massive Datasets, Cambridge University Press, 2014. (http://infolab.stanford.edu/~ullman/mmds/book.pdf)
- Ron Bekkerman, Mikhail Bilenko, John Langford. Scaling up Machine Learning: Parallel and Distributed Approaches, Cambridge University Press, 2011.

2. Дополнительная литература

- Ian Goodfellow, Yoshua Bengio, Aaron Courville. Deep Learning (Adaptive Computation and Machine Learning series), The MIT Press, 2016.
- Sandy Ryza, Uri Laserson, Sean Owen, Josh Wills. Advanced Analytics with Spark: Patterns for Learning from Data at Scale, O'Reilly Media, 2015.

3. Программное обеспечение

4. Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Интерпретатор python 3.5+	*Свободно распространяемое ПО*
2.	Пакеты jupyter, numpy, pytorch	*Свободно распространяемое ПО*

5. Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Профессиональные базы данных, информационно-справочные системы	
1.	Arxiv	*URL: https://arxiv.org*
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Pytorch documentation	URL: https://pytorch.org

6. Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

- ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);
- мультимедийный проектор с дистанционным управлением.

Учебные аудитории для семинарских и самостоятельных занятий по дисциплине не требуют специального технического оснащения.