курс «Прикладные задачи анализа данных»

Вводная лекция

Александр Дьяконов

1 сентября 2020 года

История курса

2013 «ПЗАД» спецкурс (лекции + задания)

2014 «ПЗАД» спецкурс (лекции + задания)

2015 «ПЗАД» спецкурс (лекции + задания)

2015 «АМА» Магистратура 1г. ММП ВМК (лекции + семинары)

2016-19 «ПЗАД» Магистратура 1г. ММП ВМК (лекции + семинары)

2020 «ПЗАД» Магистратура 1г. ММП ВМК (лекции + семинары)

+ «Введение в машинное обучение» (спецкурс)

+ «Машинное обучение и анализ данных» (потоковый курс)

+ «Глубокое обучение / обучение с подкреплением» (часть курса)

Цель

Решать реальные задачи

Обратить внимание на некоторую теорию

«Advanced ML»

Источники

- **Соревновательные платформы**
- **Бизнес**

Задачи

«small data» ~ до 200Гб

Максимально удобно для новичков

Простые форматы данных

Можно решать на компьютере

Лекции

- Разбор кейсов (курс ПЗАД)
- Теория по отдельным темам

Семинары

- Решение задач по теории
- Отчёты по выполнению ДЗ (публичное выступление)
- Обсуждение решений (мозговой штурм)
- Опросы по темам / контрольные работы
- Обучению программированию (R / Matlab / Python)

Зачем

- **Привлечь к решению задач**
- **Необходим практикум (текст, звуки, html и т.п.)**

сейчас, в основном, табличные данные

- **Обмен опытом**
- **Разделение труда / стратегии в соревнованиях**
- **Написание статей**
- **Помощь на начальном этапе (код)**

Почему стоит заниматься «практикой»

- Любая теория – для решения задач
- Задачника по АД нет!
- Есть возможность решать современные задачи
  - BCI
  - Flickr
  - Предсказание результатов тестирования
- Любая теория – для решения задач
- Объективная оценка исследований

Что даст

- **Нет плохих / хороших алгоритмов**
- **Есть плохие / хорошие решения**
- **Реальные задачи – не те, которым учат**

Фундаментальный уровень	**часто нет чистой регрессии, классификации и т.п.** **Пример: прогноз покупок товара,** **где 0 – не ноль покупок, а нет на складе**
Постановочный уровень	**просто другие формулировки,** **другие требования к решению** **Пример: улучшение на доли процентов**
Практический уровень	**другие объёмы** **Пример: 50 объектов задачи 70х годов**

Что требуется от слушателей

- **Компьютер (ОЗУ!)**
- **Python, …**
- **Знания (машинное обучение, статистика и т.д.)**
- **Время!**
- **Внимание (волшебный признак в задача о страховке)**

Что такое соревнование

Компания – Платформа – Решатели

1. Данные
2. Функционал качества
3. Регламент
4. Система обмена опытом

Что такое соревнование

Правила

1. Один ник
2. Нельзя обмениваться кодом/идеями
...

Наше участие

ник: Ivan Grozniy (MMP, MSU, Russia)

Что такое бизнес-задача

Данные

Неполные

Противоречивые

В разных форматах

Много/мало

Качество

Слабая формализация

Несколько функционалов

Бизнес-термины

Пример: проценты

Сроки

Система внедрения (это решение, которое будет работать!)

Система оценки

серия заданий

по каждому от 0 до 10 штрафных баллов, 0 – верно и в срок

задание (практическое)

- качественное решение (например, в LB + преодоление бенчмарка)
- код решения (например, выкладывается в форуме) – для группы!
- отчёт по решению / слайды (+ делается доклад на семинаре) – для группы!

задание по лекции (помечается ДЗ)

- провести исследование / ответить на вопрос
- сообщить о неточностях / ошибках / опечатках
- есть соревновательный фактор (важно быть первым / исправить предыдущего)

в течение 2 лекций выполнить хотя бы одно (иначе – 10 шб)

контрольные работы

репетиция экзамена

бонусы

- до 10 антиштрафных баллов за лучшее решение
- опционально за активность (обсуждение в форуме)

итоговая оценка:

<= 0 ш.б. – отлично автоматом

<= 10 ш.б. – отлично (базовая оценка)

<= 20 ш.б. – хорошо (базовая оценка)

<= 30 ш.б. – удовлетворительно (базовая оценка)

> 30 ш.б. – неудовлетворительно автоматом

Экзамен письменный – с коррекцией оценки

(как – позже уточним)

страница курса

http://goo.gl/PwWBbr

https://github.com/Dyakonov/PZAD

отдельно будет дана ссылка на группу в телеграме

https://classroom.google.com/c/MTU4NDUxMjU5MDk1?cjc=2drok7u

– ссылка в классруме

Максимально релевантный нашему курс

Steven Skiena CSE 519 – «Data Science» https://www3.cs.stonybrook.edu/~skiena/519/