курс «Машинное обучение»

Термины

Александр Дьяконов

Немного о курсе: команда

Дьяконов Александр Геннадьевич

- академический руководитель ЦУ, доктор физ-мат наук, лектор программы AI Masters
- лучший молодой преподаватель вузов в IT (Росрейтинг, 2014)
- работал в компаниях Ozon (5 лет), Algomost (6 лет), Dasha.ai (4 года)
- 1 место в рейтинге Kaggle (2012)

Малашенко Сергей Анатольевич

- Выпускник СарФТИ НИЯУ МИФИ (прикладная математика)
- Lead Data Scientist / Team Lead в Erlyvideo, разработка систем видеоаналитики
- Senior Software Engineer в V5Systems, разработка систем видеоаналитики на встраиваемых системах (детектирование объектов)
- Senior Software Engineer в ЗАО Intel

Васильев Роман Александрович

- Выпускник ВМК МГУ
- Lead Data Scientist в Яндекс.Лавке (e-com). До этого работал Data Scientist'ом в МегаФоне (телеком), Senior Data Scientist'ом в Магните (ритейл)
- Курсы по машинному обучению в ВШЭ

Немного о курсе: программа

Лекции

1. Общие термины: AI, Big Data, ML и т.п.
2. Оценка среднего и вероятности
3. Постановка задачи ML и общие термины
4. Визуализация (часть 1)
5. Визуализация (часть 2)
6. Метрические алгоритмы
7. Контроль качества и выбор моделей
8. Линейные алгоритмы (часть 1: линейная и логистическая регрессии)
9. Линейные алгоритмы (часть 2: SVM, суррогатные функции)
10. Нелинейные алгоритмы
11. Деревья решений
12. Ансамбли (общий взгляд)
13. Предобработка данных
14. Генерация признаков
15. Селекция признаков
16. Кластеризация

Семинары

1. Введение в Python
2. Введение в ООП
3. Библиотека numpy + scipy + pandas. Примеры использования
4. Простые примеры с визуализацией данных (matplotlib + seaborn + plotly)
5. Сложные примеры с визуализацией данных (matplotlib + seaborn + plotly)
6. Scikit-learn. KNN: практическое применение
7. Оценка качества моделей, кросс-валидация, подбор метапараметров
8. Линейные алгоритмы, задача и демонстрация регуляризации
9. SVM и условная оптимизация
10. Ядерные линейные алгоритмы. Примеры, визуализации
11. Деревья решений. Примеры, визуализации
12. Демонстрация примера, где используется (бэггинг, бустинг, стекинг)
13. Демонстрация предобработки данных на некотором датасете
14. Конвейеры обработки данных в Scikit-learn
15. Методы отборы признаков для линейных и деревянных моделей
16. K-means. И разные метрики для выбора моделей

Ключевые слова

Наука о данных (Data Science)

Статистика (Statistics)

Искусственный интеллект (Artificial Intelligence)

Анализ данных (Data Mining)

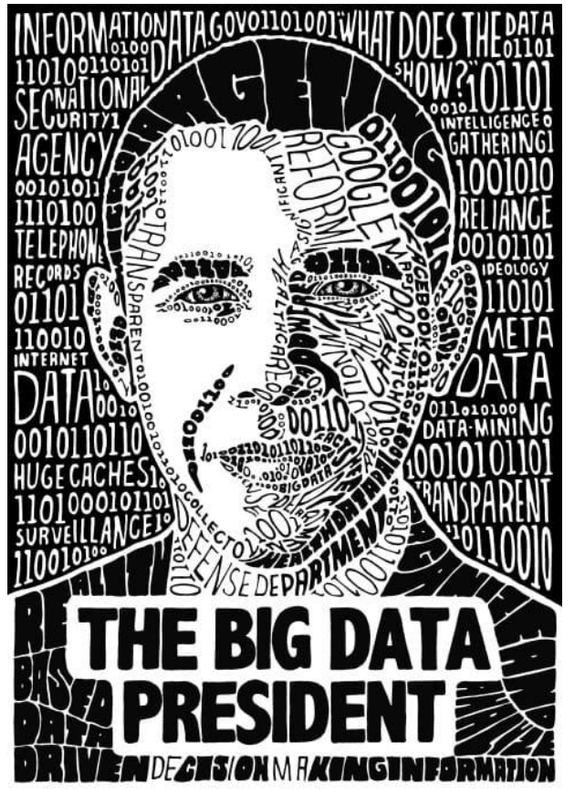
Машинное обучение (Machine learning)

Большие данные (Big Data)

Наука о данных (Data Science)

– направление науки и технологий представления, сбора, обработки, хранения, анализа и использования данных в цифровой форме

всё перечисленное выше – разделы DS



https://www.washingtonpost.com/opinions/obama-the-big-data-president/2013/06/14/1d71fe2e-d391-11e2-b05f-3ea3f0e7bb5a_story.html

Анализ данных (Data Mining)

– нахождение закономерностей и моделей, которые

- **валидны**
  (соответствуют действительности и есть в новых данных)
- **полезны**
  (экономят время, ресурсы, позволяют заработать $)
- **нетривиальны**
  (неочевидны до анализа)
- **понятны / интерпретируемы**
  (описываются, могут быть объяснены специалистам)

в широком смысле – область человеческой деятельности
(не наука! т.к. также искусство, ремесло, спорт)

Математическая статистика

– математическая дисциплина, разрабатывающая математические методы систематизации и использования статистических данных для научных и практических выводов

уже была в обязательных курсах...

Машинное обучение (Machine Learning)

Что такое обучение?

Машинное обучение (Machine Learning)

Обучение — приобретение необходимой функциональности посредством опыта

Обучение на примерах

Учимся ходить
Делаем шаг – получилось / нет

Учим названия животных
Показывают и называют

Обучение по определениям

В школе – дают определения

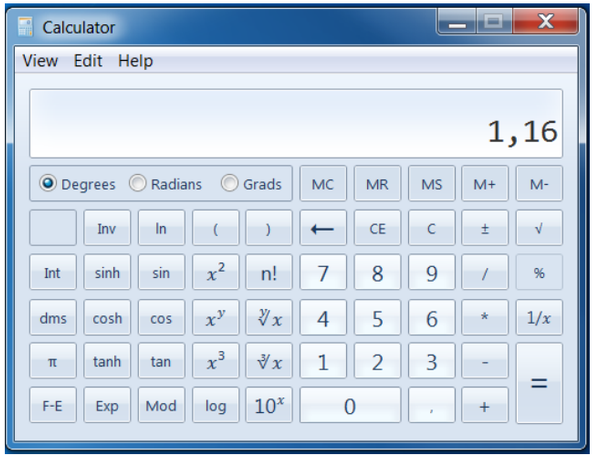
Машинное обучение

Машинное обучение — процесс, в результате которого машина способна показывать поведение, которое в нее не было явно запрограммировано

– наука, которая занимается созданием и исследованием таких процессов

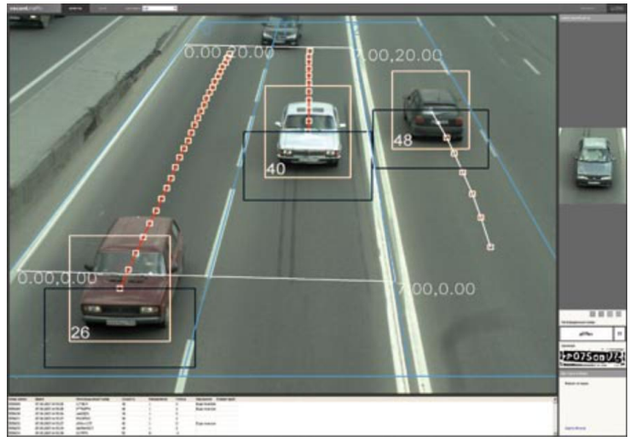
A.L. Samuel Some Studies in Machine Learning Using the Game of Checkers // IBM Journal. July 1959. P. 210–229.

Программирование



Программируем последовательность действий

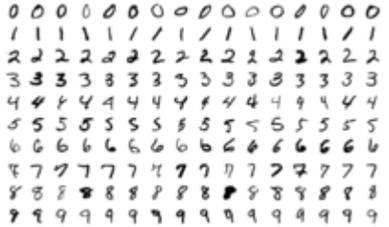
Обучение



Программируем алгоритм анализа информации

Машинное обучение

«Компьютерная программа обучается из опыта E в классе задач T с мерой качества P, если качество измеренное с помощью P в классе задач T увеличивается по мере увеличения опыта E». Том Митчел



Задача: распознавание символов
Мера: процент правильно распознанных
Опыт: база, размеченных вручную, изображений символов

Задача: игра в шашки / шахматы / го
Мера: процент побед
Опыт: игра программы против себя

Задача: рекомендация товаров/услуг/видео
Мера: процент успешных рекомендаций
Опыт: список товаров, просмотренных/ купленных/оцененных пользователями

Примеры задач

- диагностика болезней, прогнозирование эффективности лекарства
- распознавание образов, символов (Character/ Handwriting Recognition)
- распознавание речи (Speech Recognition)
- распознавание лиц (Face detection)
- классификация спама (Spam filtering)
- идентификация (Person identification / Authentication) лица, отпечатков, радужка глаза и т.п.
- тональность текста (sentimental analysis)
- прогноз спроса / выручки (Demand Forecasting)
- скоринг (Credit scoring) – определение кредитоспособности
- определение суммы / пакета страхования
- психотип по профилю соцсети / фотографии
- предсказание оттока (ухода сотрудника / абонента)
- поиск кандидатов на вакансии
- рекомендации товаров
- ранжирование Web-страниц
- ожидание прибыли магазина (учитывая GPS) / рейтинга фильма / доходности сделки
- анализ форумов, поиск оскорблений, жалоб, автоматическая модерация
- предсказание поведения клиента / пользователя (ex: трат клиента)
- поиск похожих объектов, документов, событий (например, юридических дел)
- обнаружение нетипичных пользователей, фрода, инсайдеров
- нахождение зависимостей
- сегментация изображений
- тегирование/аннотирование документов (automatic summarization)

Пример задачи машинного обучения – классификация

Iris setosa

Iris virginica

Iris versicolor

Длина чашелистника	Ширина чашелистника	Длина лепестка	Ширина лепестка	Вид ириса
4.3	3.0	1.1	0.1	setosa
4.4	2.9	1.4	0.2	setosa
4.4	3.0	1.3	0.2	setosa
				…
4.9	2.5	4.5	1.7	virginica
5.6	2.8	4.9	2.0	virginica
				…
5.0	2.0	3.5	1.0	versicolor
5.1	2.5	3.3	1.1	versicolor

Пример задачи машинного обучения – скоринг

Id	статус	г.р.	Пол	офис	На счету	просрочки	возврат
43223	физ	1967	М	54	10000	0	Да
43224	физ	1970	Ж	33	2000	2	Нет
43225	юр	1954	М	54	23500	0	Да

Прогноз поведения пользователя с помощь описания (и кредитной истории)

Большие данные (Big Data)

– технологии сбора, хранения, обработки и анализа данных огромных объёмов и значительного многообразия

Характеристики:

VELOCITY
скорость поступления
VOLUME
объёмы
VARIETY
разнообразие

VERACITY
достоверность

Причины

- удешевление средств хранения
- ускорение средств обработки
- миниатюризация устройств (смартфоны, датчики и т.п.)
- новые форматы / неструктурированность
- новые технологии (GPS)
- интерес бизнеса
- успехи отдельных подходов в ML (например, DL)

коммерческий и технологический термин

Большие данные (Big Data)

Пример:

Google Flu Trends

https://www.google.org/flutrends/about/

- **анализ поисковых запросов**
- **корреляция с известными эпидемиями**
- **прогнозная модель**

Виктор Майер-Шенбергер и Кеннет Кукьер «Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим»

Искусственный интеллект (Artificial Intelligence)

– свойство интеллектуальных систем выполнять «творческие» функции, которые традиционно считаются прерогативой человека

– эти интеллектуальные системы (программы) и машины, в которых они реализованы

– наука и технология создания интеллектуальных этих интеллектуальных систем

«творческие» функции

- логическое мышление (понимание противоречий, умение делать выводы),
- креативные (сочинять истории и музыку, рисовать и т.п.),
- разговорные (понимать речь, отвечать на вопросы, поддерживать диалог и т.п.),
- ориентация (планирование маршрута, узнавание знакомых мест и т.п.),
- координация (ходьба, бег, акробатические упражнения)

...

Искусственный интеллект (Artificial Intelligence)

- умные чат-боты
- автомобили-беспилотники
- умный дом

Пример:
IBM построила Watson, который выиграл в Jeopardy

сейчас самый популярный термин

Искусственный интеллект (Artificial Intelligence)

Проблема «почти реализации»

Как только машина «учится новым способностям» выясняется, что за этим стоят простые вычисления. Можно ли считать это AI?

https://cameralabs.org/9808-20-let-nazad-kompyuter-vpervye-vyigral-shakhmatnuyu-partiyu-u-chempiona-mira

Искусственный интеллект (Artificial Intelligence)

AI в слабом смысле (weak AI / narrow AI)

имитация конкретной творческой деятельности человека

AI в сильном смысле (AGI / strong AI / full AI)

компьютеры могут приобрести способность мыслить и осознавать себя как отдельную личность (в частности, понимать собственные мысли)

также говорят про самообучение, способность самому ставить задачи, понимать свои мысли и т.п.

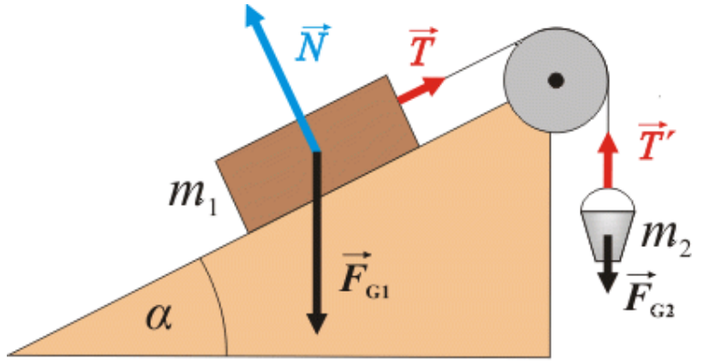
Проблема сознания

- самоидентификация
- идентификация других и противопоставление
- борьба за ресурсы

Наш курс = машинное обучение + анализ данных

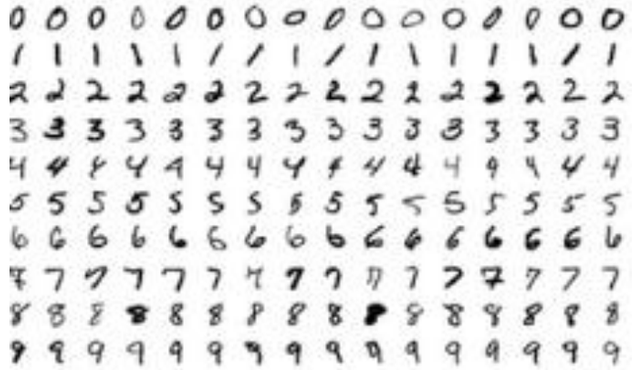
model based reasoning

можем записать уравнение



case based reasoning

~ на основе прецедентов: известна выборка



Зависимость дана

- **неполностью (прецедентно)**
- **потенциально очень сложная (не получится формулы)**
- **часто зависимость не от чисел (пример: тональность текста)**

Литература / ссылки

Виктор Майер-Шенбергер и Кеннет Кукьер
«Большие данные: Революция, которая изменит то, как мы живем, работаем и мыслим»

Том Таулли
«Основы искусственного интеллекта: нетехническое введение»

Педро Домингос
«Верховный алгоритм»