

курс «Машинное обучение»

Термины

Александр Дьяконов

06 сентября 2021 года

Ключевые слова

Наука о данных (Data Science)

Статистика (Statistics)

Искусственный интеллект (Artificial Intelligence)

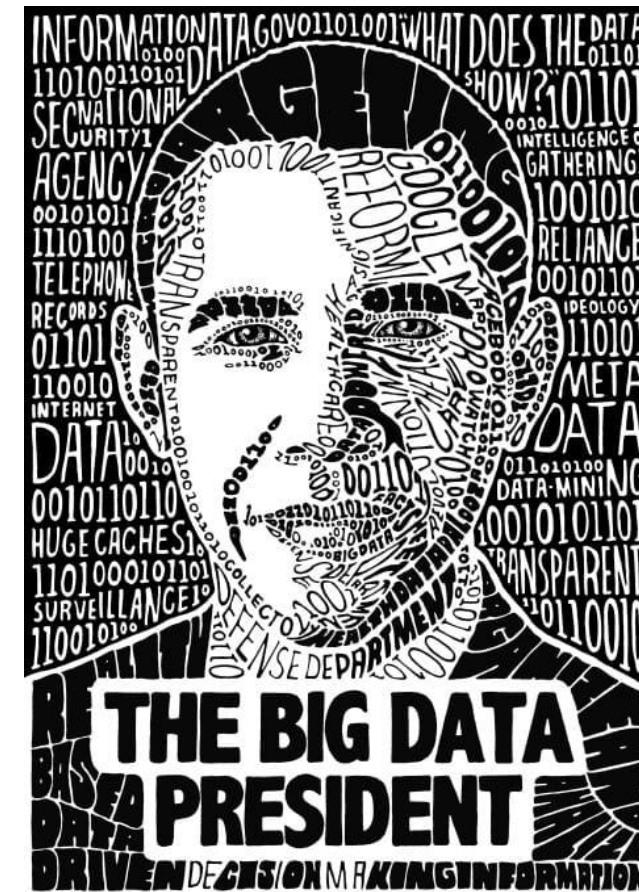
Анализ данных (Data Mining)

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

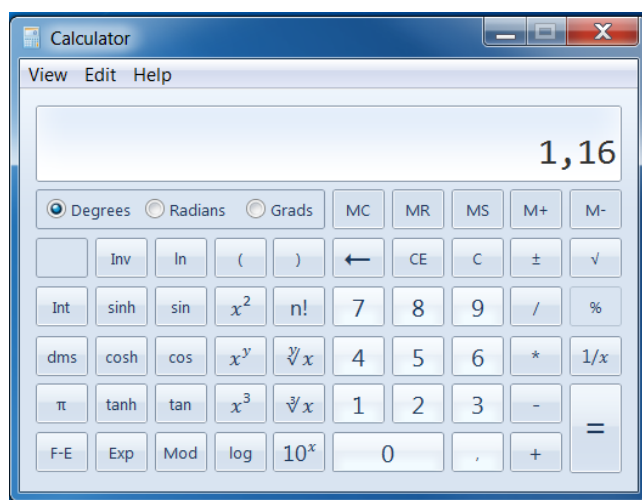
В школе – дают определения

Машинное обучение

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

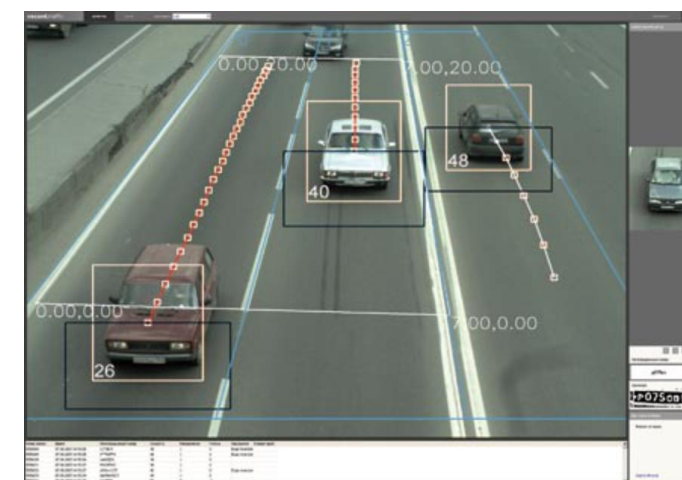
A.L. Samuel Some Studies in Machine Learning Using the Game of Checkers // IBM Journal. July 1959. P. 210–229.

Программирование



Программируем последовательность действий

Обучение



Программируем алгоритм анализа информации

«Машинное обучение» – наука!

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

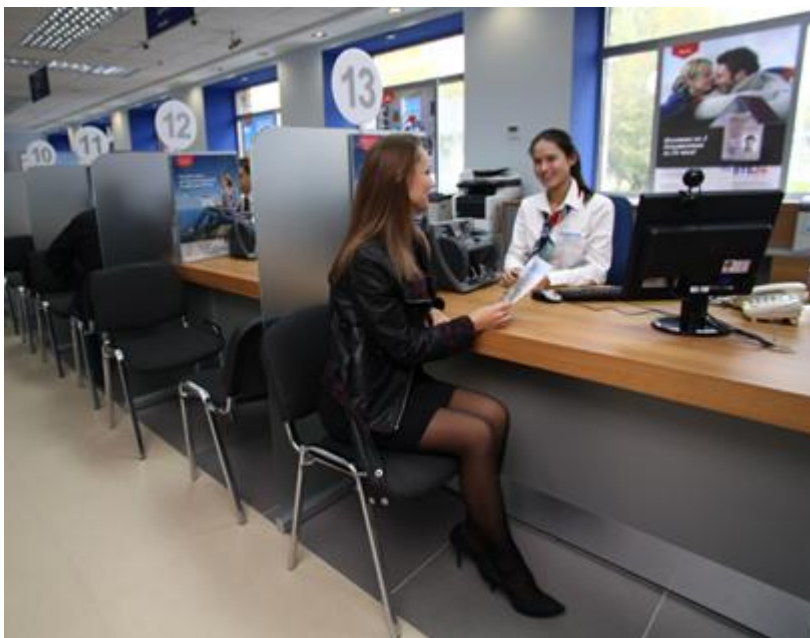
- диагностика болезней, прогнозирование эффективности лекарства
- распознавание образов, символов (Character/ Handwriting Recognition)
- распознавание речи (Speech Recognition)
- распознавание лиц (Face detection)
- классификация спама (Spam filtering)
- идентификация (Person identification / Authentication) 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



<i>Iris setosa</i>		<i>Iris virginica</i>		<i>Iris versicolor</i>
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)

- наука и технология создания интеллектуальных машин
(в том числе, программ)
- свойство интеллектуальных систем выполнять творческие функции, которые традиционно считаются прерогативой человека
 - 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?

AI в слабом смысле

творческие функции человека (например, придумывание сказок)

AI в сильном смысле

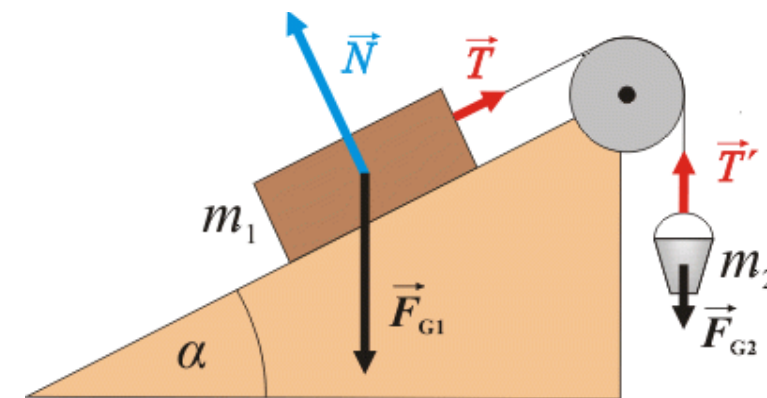
компьютеры могут приобрести способность мыслить и осознавать себя как отдельную личность (в частности, понимать собственные мысли)

Проблема сознания

- самоидентификация
- идентификация других и противопоставление
- борьба за ресурсы

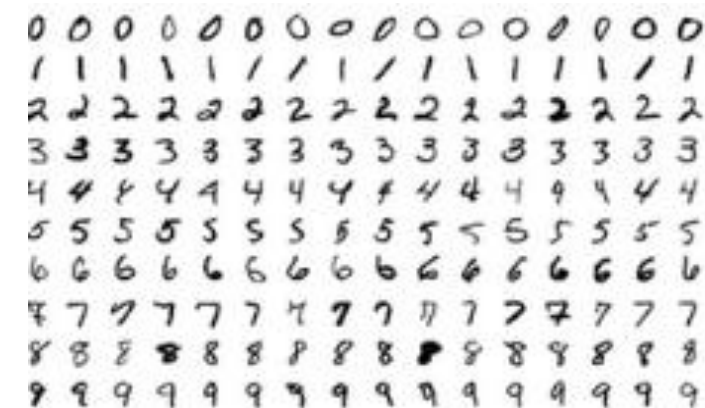
Наш курс = машинное обучение + анализ данных

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»



