

Машинное обучение и анализ данных

Вводная лекция

Дьяконов А.Г.

**Московский государственный университет
имени М.В. Ломоносова (Москва, Россия)**



Структура курса

начинает читаться весной 2019 года
для 3 потока 4 курса ВМК

лекции со слайдами
семинаров нет

понедельник, 16:20, П-5

Февраль

пн	вт	ср	чт	пт	сб	вс
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22*	23	24
25	26	27	28	1	2	3

Апрель

пн	вт	ср	чт	пт	сб	вс
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30*	1	2	3	4	5

Март

пн	вт	ср	чт	пт	сб	вс
25	26	27	28	1	2	3
4	5	6	7*	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	31

Май

пн	вт	ср	чт	пт	сб	вс
29	30	1	2	3	4	5
6	7	8*	9	10	11	12
13	14	15	16	17	18	19
20	21	22	23	24	25	26
27	28	29	30	31	1	2

Отчётность:
экзамен (письменный)

В течение семестра задания

- онлайн-тесты
- офлайн-тесты
- решение прикладных задач (в виде соревнования)

Важно: вовремя сдавать все задания

Лектор

Дьяконов Александр Геннадьевич
кафедра ММП ВМК МГУ

канал в телеграмме для общения «MSU_Dyakov's_courses»:

<https://t.me/joinchat/DlgJbg8Lec9KKDvWC8L0cg>

страница курса

<https://github.com/Dyakov/MLDM/>

Другие курсы

- **«Введение в машинное обучение» бакалавриат ВМК (спецкурс)**
 - **«Глубокое обучение» каф. ММП ВМК (лекции + семинары)**
 - **«МОАД» бакалавриат 3 поток 4 курс ВМК (лекции)**
 - **«ПЗАД» Магистратура 1г. ММП ВМК (лекции + семинары)**

Цель курса

Дать основы машинного обучения

- термины
- классические методы
- этапы решения прикладных задач

нет теорем и доказательств, но есть математика

Нужны

- знания Python
- специализированных библиотек
Scikit-Learn, Numpy, Scipy, Pandas, Matplotlib

Первое домашнее задание

- пройти опрос (обязательно – это регистрация на курс)
- начать изучение языка Python и специализированных пакетов
в помощь: <https://github.com/Dyakonov/IML>

Ключевые слова

Наука о данных (Data Science)

Статистика (Statistics)

Искусственный интеллект (Artificial Intelligence)

Анализ данных (Data Mining)

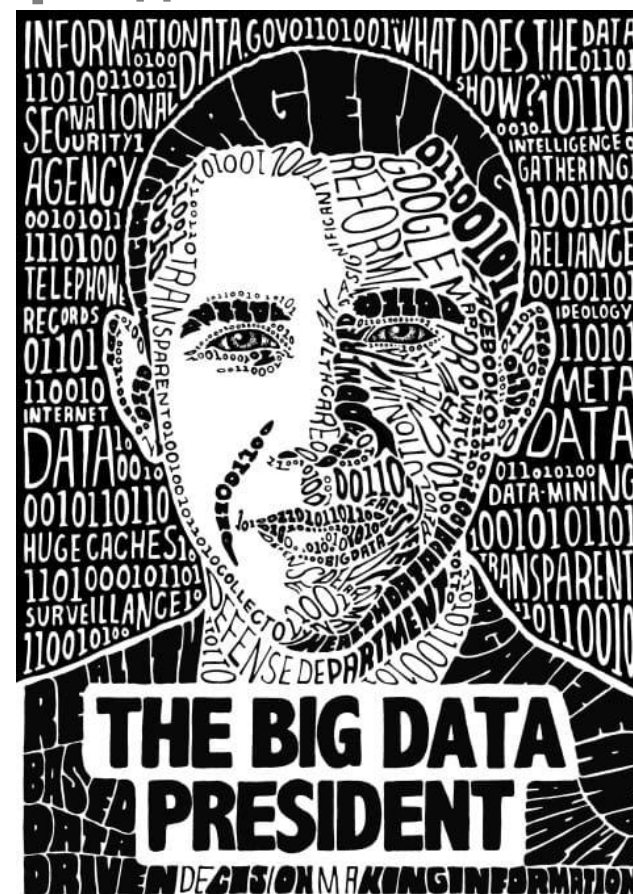
Машинное обучение (Machine learning)

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

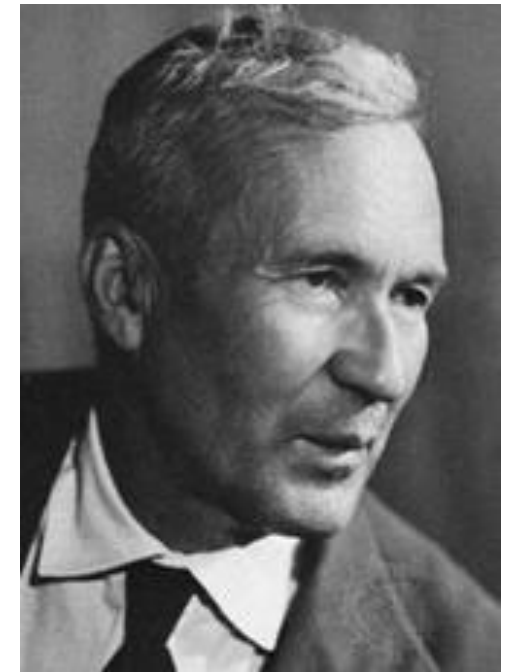
- **валидны**
(соответствуют действительности и есть в новых данных)
- **полезны**
(экономят время, ресурсы, позволяют заработать \$)
- **нетривиальны**
(неочевидны до анализа)
- **понятны / интерпретируемы**
(описываются, могут быть объяснены специалистам)



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)



Обучение — приобретение необходимой функциональности посредством опыта

Обучение на примерах

Учимся ходить

Делаем шаг – получилось / нет

Учим названия животных

Показывают и называют

Обучение по определениям

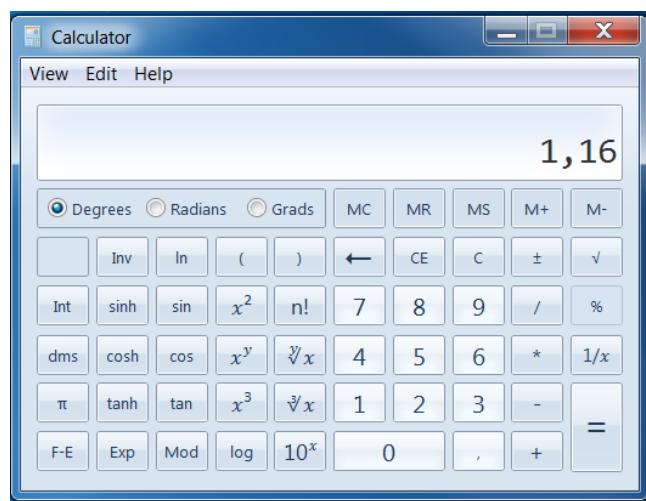
В школе – дают определения

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.

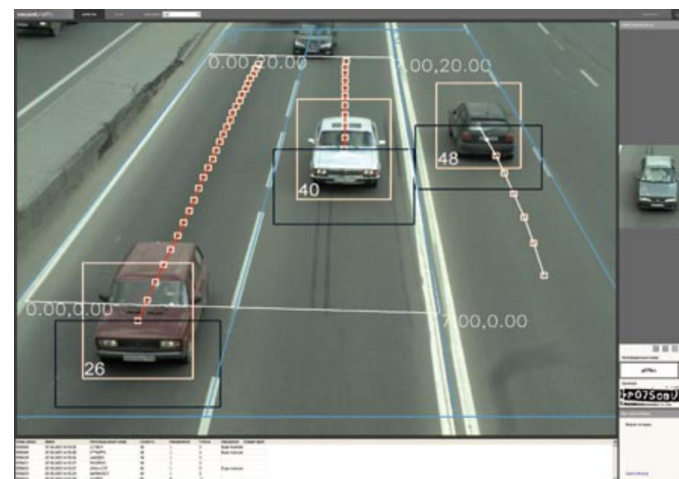
Программирование



**Программируем
последовательность действий**

«Машинное обучение» – наука!

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
- распознавание символов (Character/ Handwriting Recognition)
 - распознавание речи
 - 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)
- анализ форумов, поиск оскорблений, жалоб, автоматическая модерация
 - предсказание поведения клиента / пользователя
- 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



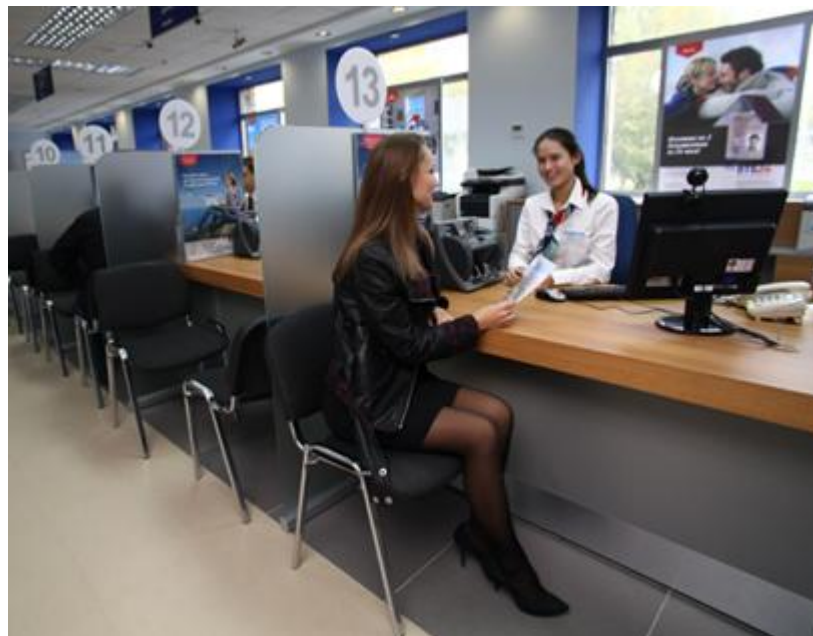
Iris virginica



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 помощью 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

коммерческий и технологический термин

- 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)

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



IBM построила Watson, который выиграл в Jeopardy

сейчас самый популярный термин

Наш курс

Машинное обучение + анализ данных

model based reasoning

можем записать уравнение « $F = M \times A$ »

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)**

Совет по инструментарию



Язык программирования Python

<https://www.python.org/>



Библиотека для матричных вычислений и линейной алгебры

<http://www.numpy.org/>



Библиотека для научных вычислений

<https://www.scipy.org/>



Библиотека для визуализации

<https://matplotlib.org/>

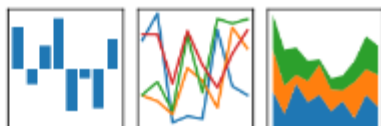


Библиотека для машинного обучения

<http://scikit-learn.org/>

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Библиотека для обработки данных



<https://pandas.pydata.org/>

Совет по инструментарию



**научный дистрибутив
Anaconda Python
от Continuum**

<https://www.anaconda.com/download/>

Python 3.6 version *	Python 2.7 version *
	
64-Bit Graphical Installer (631 MB) ?	64-Bit Graphical Installer (564 MB) ?
32-Bit Graphical Installer (506 MB)	32-Bit Graphical Installer (443 MB)

Совет по инструментарию



Python, R, Julia, Scala, F#

<http://jupyter.org/>



<https://www.jetbrains.com/pycharm/>

эволюция IPython Notebook

**для создания и обмена
«ноутбуками»:**

- код
- полнотекстовые комментарии
- уравнения
- визуализация

**интегрированная среда
разработки для языка
программирования Python**

Совет по инструментарию

Basic Numerical Integration: the Trapezoid Rule

A simple illustration of the trapezoid rule for definite integration:

$$\int_a^b f(x) dx \approx \frac{1}{2} \sum_{k=1}^N (x_k - x_{k-1}) (f(x_k) + f(x_{k-1})).$$

First, we define a simple function and sample it between 0 and 10 at 200 points

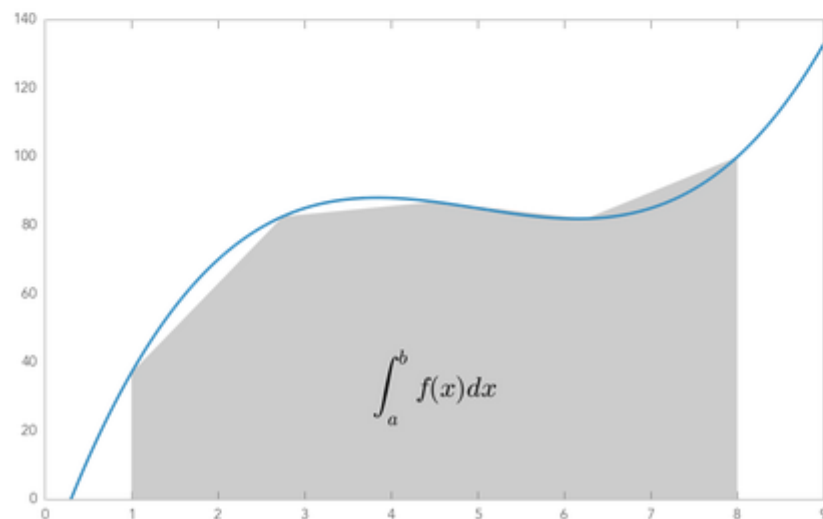
```
In [1]: @matplotlib inline
import numpy as np
import matplotlib.pyplot as plt
```

```
In [2]: def f(x):
        return (x-3)*(x-5)*(x-7)+85

x = np.linspace(0, 10, 200)
y = f(x)
```

Choose a region to integrate over and take only a few points in that region

```
In [4]: plt.plot(x, y, lw=2)
plt.axis([0, 9, 0, 140])
plt.fill_between(xint, 0, yint, facecolor='gray', alpha=0.4)
plt.text(0.5*(a+b), 30, r"$\int_a^b f(x)dx$", horizontalalignment='center', fontsize=20);
```



Пример решения задачи ML

https://github.com/Dyakonov/notebooks/blob/master/dj_benchmark_GMSC_01.ipynb

```
from sklearn.ensemble import RandomForestClassifier

model = RandomForestClassifier(n_estimators=100,
                              max_depth=2,
                              random_state=0) # модель

model.fit(train, y) # обучение

a = model.predict_proba(test)[:,1] # предсказание
```

Материалы по курсу

Лекции К.В. Воронцова

http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%28%D0%BA%D1%83%D1%80%D1%81_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9%2C_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2%29

The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2nd Edition, Springer, 2009.

<https://web.stanford.edu/~hastie/Papers/ESLII.pdf>

Charu C. Aggarwal Data Mining: The Textbook. Springer, 2015.

Книга по современному глубокому обучению

<https://www.deeplearningbook.org>

Соревнования по анализу данных

<https://www.kaggle.com/>

Обзорная книга

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

Онлайн-курсы:

- <https://www.coursera.org/learn/machine-learning>
- <https://www.coursera.org/learn/introduction-machine-learning>
- <https://coursera.org/specializations/machine-learning-data-analysis>