Машинное обучение и работа с большими данными

#2. Задачи в ML. Pandas.

ПЛАН

#2. Задачи в ML. Pandas.

01

ВВЕДЕНИЕ
7 МИН

Основные понятия в ML.

Обучение с- и без- учителя. Типы задач и их комбинации.

ОБЗОР ЗАДАЧ
20 МИН

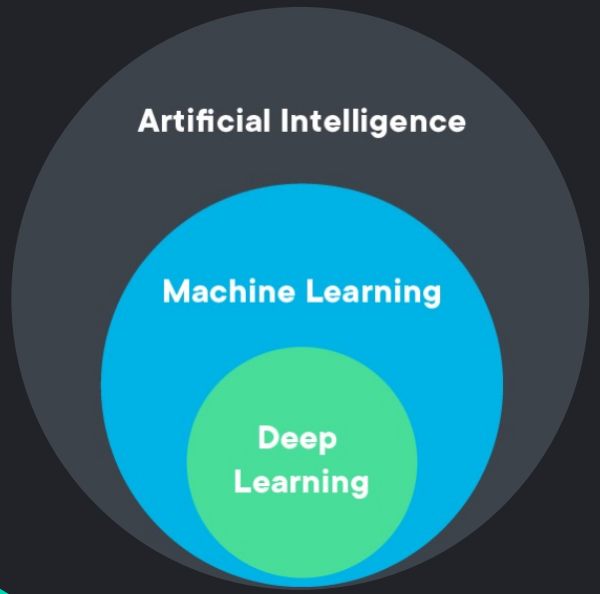
02

03

ПРАКТИКА
10 МИН

Знакомство с библиотеками питона pandas, numpy для анализа данных.

01 ВВЕДЕНИЕ
7 МИН



AI or ML?

Только ML

AI («ИИ» – искусственный интеллект) используют популисты и гуманитарии

ML, DL

Отличия

ML – общее и более широкое понятие
DL – нейронные сети
+ новомодные эвристики

01 ВВЕДЕНИЕ
7 МИН

смотря
сколько details

English terms

Принятая и общая терминология на англ. Иногда буду пытаться русифицировать.

ML начался в 2010-х?

Нет

Основы заложены еще в 50-х годах 20го века.

Rosenblatt, Frank. "The perceptron: a probabilistic model for Information storage and organization in the brain." *Psychological review* 65.6 (1958): 386.

KNN 1960-х
Backpropagation 1970-х
Classic ML 1990-х

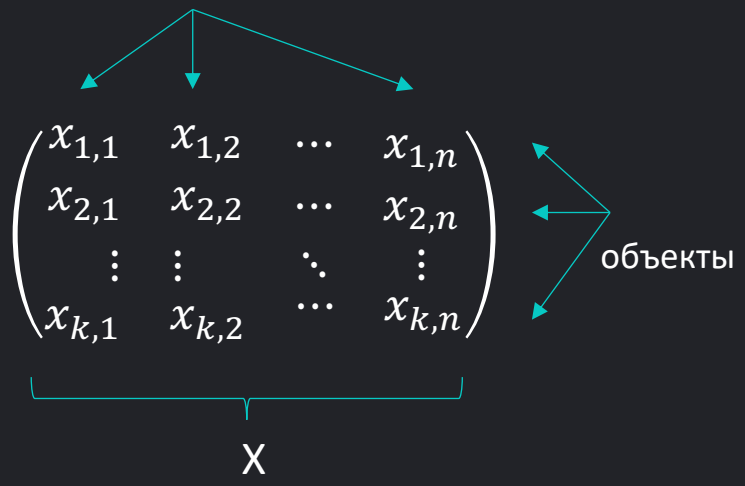
Всплеск 2010-х

Deep Learning

Начал работать лучше, чем человек.

02 ВВЕДЕНИЕ
20 МИН

признаки (features) [характеристики объекта]



X – матрица объектов и признаков.

Основная задача [в терминах ML]:
Каждому объекту $\mathbf{x_i}$ поставить в соответствие значение $\mathbf{y_i}$

$$f(X) = Y$$

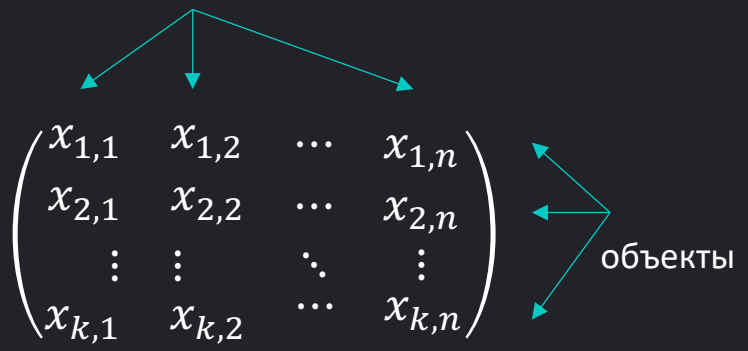
Y – новое знание
f – решающая функция

F_1	F_2	F_3	F_N	Y_1	Y_2	Y_3
8.93	M	…	1	1	5	9.15
9.35	F	…	0	0	3	8.72
…	…	…	…	…	…	…
6.12	M	…	1	1	11	6.52
				0 или 1	0..11	$-\infty;\ +\infty$

X

02 ВВЕДЕНИЕ
20 МИН

признаки (features) [характеристики объекта]



Клиент 0	
Возраст	29
Пол (М – 2, Ж – 1)	2
Доход	120
Размер семьи	5

x_0

29
2
120
5

x_0	29	2	120	5		0.1		6.8	0
x_1	31	1	95	2	×	0.2	=	7.05	1
…	…					0.05		…	…
x_k	44	2	60	0		-0.5		7.8	1

X — матрица объект-признак

W — вектор весов

$X \times W = \bar{Y}$

$\bar{Y}$ — скор

$\bar{Y} > 7$ — ответ классификатора

29*0.1 + 2*0.2 + 120*0.05 – 5*0.5

31*0.1 + 1*0.2 + 95*0.05 – 2*0.5

…

44*0.1 + 2*0.2 + 60*0.05 – 0*0.5

02 ВВЕДЕНИЕ
20 МИН

ML задачи:

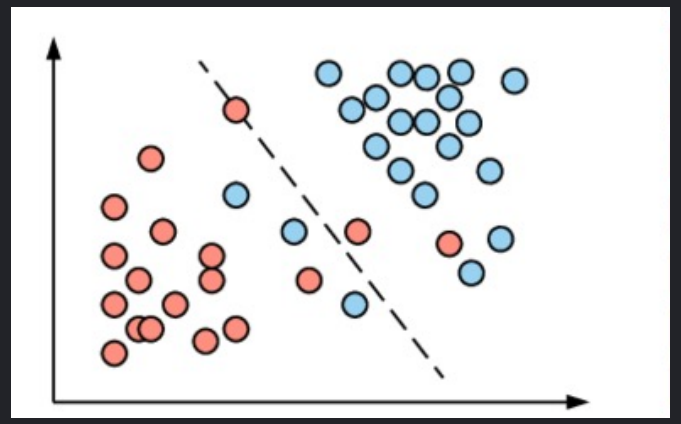
- Supervised
  - classification
  - regression
- Unsupervised
  - clustering

- semi-supervised
- semi-unsupervised
- self-supervised
- ranking

- Supervised

Учимся на размеченных данных

F_1	F_2	F_3	F_N	Y

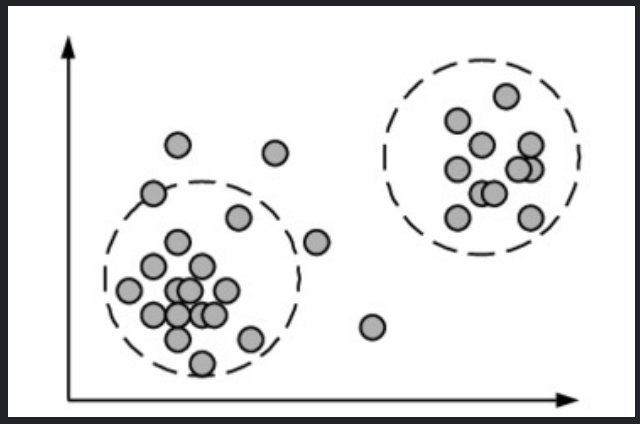


- Unsupervised

Учимся без разметки

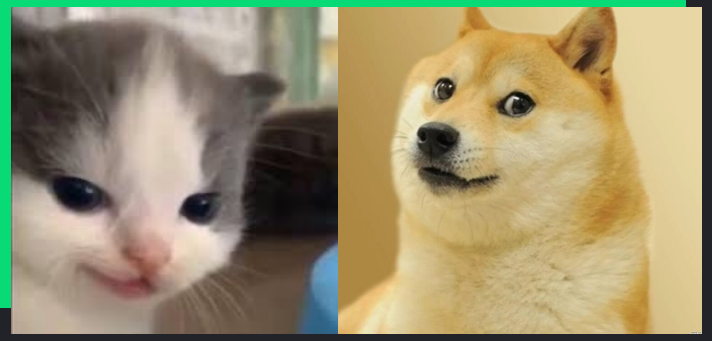
F_1	F_2	F_3	F_N

нет Y*
(правильных ответов)

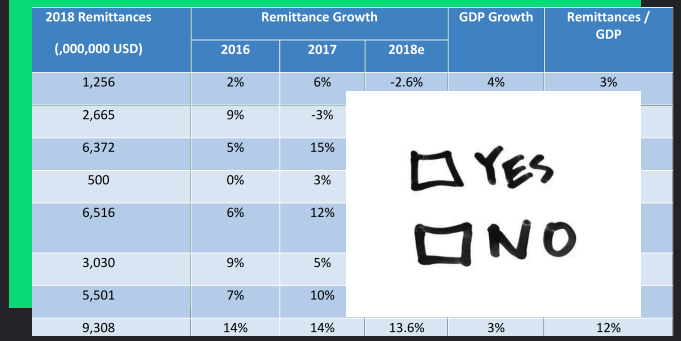


* имеется ввиду, что Y нет с точки зрения «решающего механизма».
Supervised оптимизатор **видит Y** и обучается относительно него.
Unsupervised оптимизатор **не видит Y** и обучается относительно своих критериев.

С учителем | Классификация



Классификация изображений



Кредитный скоринг

Антифрод

Спам-фильтрация

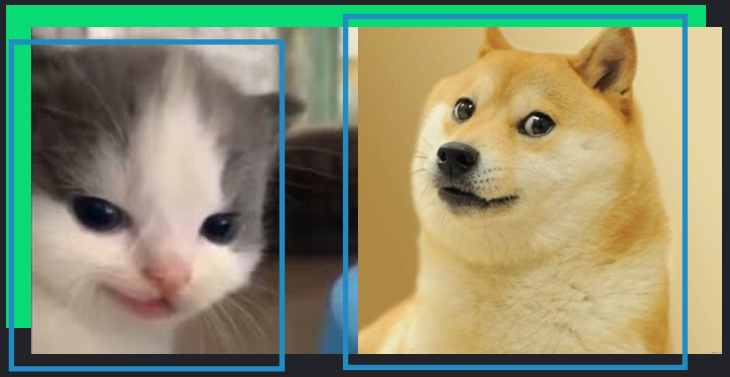
Задача: поставить каждому объекту метку.
Метка – ***дискретная***, принимает одно значение из множества.

$Y \in \{0, 1, 2, \ldots\}$ или $Y \in \{\text{котик, не котик}\}$
Y – класс объекта

Типы задач классификации:

- **binary** [двуклассовая] — **каждый объект – имеет 1 класс из 2 возможных** {0, 1}
- **multiclass** [многоклассовая] — **каждый объект – имеет класс из 2+ возможных** {0, 1, …}
- **multilabel** [??] *многометочная* — **каждый объект – может принадлежать к 1 и более классов**

С учителем | Регрессия



Детекция на изображениях

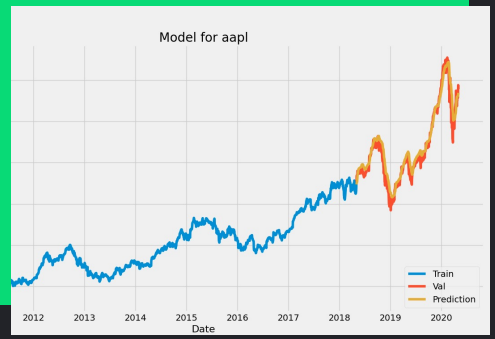
***Отличие:** предсказываем параметры рамки.

Remittance Growth			GDP Growth	Remittances / GDP
2016	2017	2018e		
2%	6%	-2.6%	4%	3%
9%	-3%	15.8%	1%	0%
5%	15%	15.9%	3%	2%
0%	3%	-5.9%	3%	1%
6%	12%			
9%	5%			
7%	10%	9.1%	2%	22%
14%	14%	13.6%	3%	12%

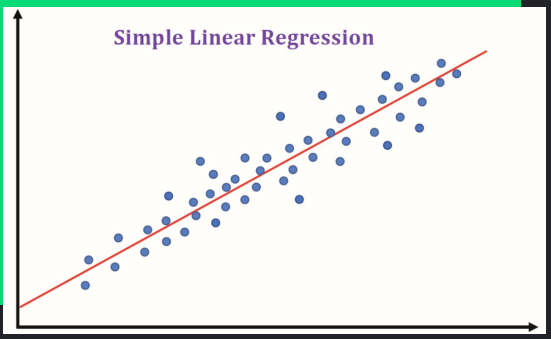
limit = 150 000

Кредитный скоринг

***Отличие:** предсказываем размер кредитного лимита.



Прогнозирование цен акций



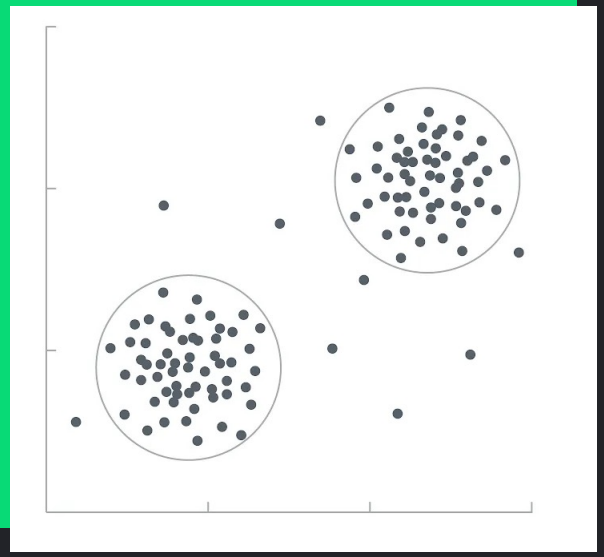
Восстановление зависимости

Задача: поставить каждому объекту число.
Число – ***непрерывное***, ограничено диапазоном.

$Y \in (-\infty; +\infty)$
Y – число, определенное в каком-то диапазоне.
Может принимать **любое** значение.

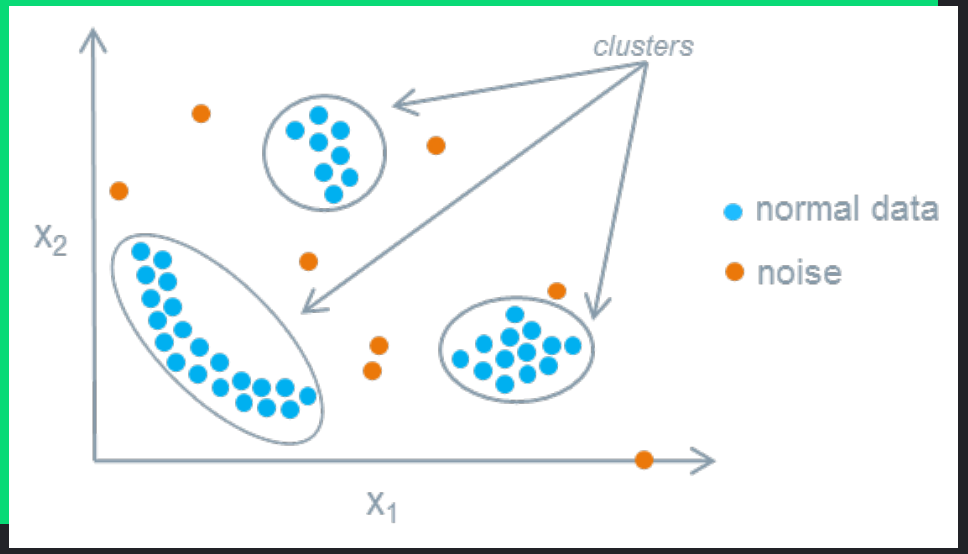
02 ВВЕДЕНИЕ
20 МИН

Без учителя

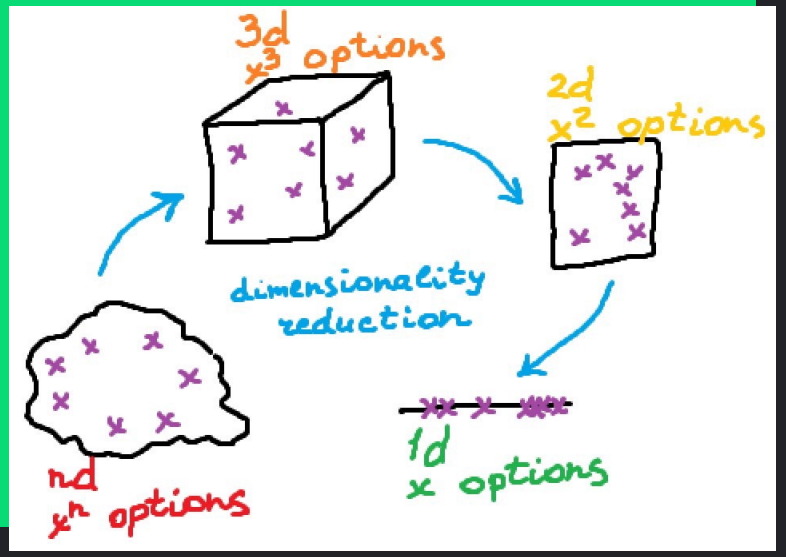


Кластеризация (без учителя)

* бывает кластеризация с учителем



Поиск аномалий



Понижение размерности*

* может быть техникой, а может полноценным методом решения

Задача: построить классификатор / регрессор.
Отличие от обучения с учителем: «решающий механизм» не получает информации из-вне о том, правильно ли он решает или нет.

02 ВВЕДЕНИЕ
20 МИН

Сложный задачи

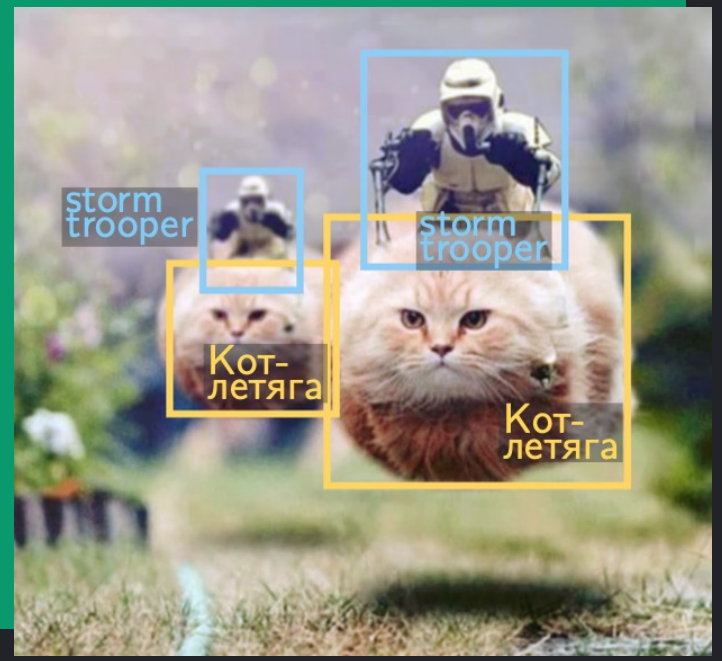
СЕГМЕНТАЦИЯ

- Автопилот
- Составление карт по спутникам
- Маски в Snapchat/IG
- Медицина
- Редактирование фото

ЗАДАЧА ГЛУБИНЫ

- Автопилот
- 3-D фотография
- VR/AR
- Экономия на LIDAR

ДЕТЕКЦИЯ



- Распознавание лиц
- Поиск объектов
- Автопилот
- Маски в Snapchat/IG
- Медицина


— Всегда было
ИИ
— Это просто перемножение матриц?
@ithumor

03 ПРАКТИКА
15 МИН

Практика 1

Интро.

Откройте учебный Notebook из архива с материалами к занятию.