Машинное обучение

Лекция 1 Введение

Логистика Курса

Материалы: https://github.com/WeaselCMC/ml_dpo_2021

Телеграм: По всем вопросам https://t.me/joinchat/u5NGLf6CUoEyNjQy

Логистика Курса: преподаватели

Евгений Егоров



Tg: @iesh00a

Website: https://evgenii-egorov.github.io

Teaching: Skoltech, ВШЭ, ШАД

Индустриальные проекты: Сбербанк, Huawei,

Газпром, Азбука Вкуса

Научные публикации: NeurIPS, ICML, AABI

Руслан Костоев



Tg: @mrstrangelove

Website: Coming soon

Teaching: Skoltech, ВШЭ, ШАД

Индустриальные проекты: Газпром,

Google, Huawei, Северсталь

Логистика Курса: темы

- Введение и основные задачи
- Линейная регрессия
- Градиентные методы обучения
- Классификация (метрики качества классификации, логистическая регрессия и SVM, Многоклассовая классификация)
- Решающие деревья
- Бэггинг и случайные леса, Градиентный бустинг
- Отбор признаков и снижение размерности
- Кластеризация
- Поиск аномалий
- Рекомендательные системы
- Ранжирование
- Заключение

Логистика Курса: оценки

- Домашние задания (HW)
 - 5-6 самостоятельных заданий
 - У всех домашек одинаковый вес
 - Для зачета нужно получить в среднем 6/10

Логистика Курса: работа в классе

- 1ч 20м лекция
 - Можно перебивать
 - Задавать вопросы
- 10м перерыв
- 1ч 30м практическая работа
 - Задание в ноутбуках
 - o Breakout room по 2-3 человека

Логистика Курса

Вопросы?

Что такое машинное обучение ?

"Field of study that gives computers the ability to learn without being explicitly programmed" - Arthur Samuel



https://en.wikipedia.org/wiki/Arthur_Samuel

Задача: Перевести часы в минуты



Задача: Перевести часы в минуты

х - часы

$$f(x) = 60x$$



Задача: Как долго будет падать тело?

- Я кинула объект с высоты х
- Через какое время он упадет на Землю?

Задача: Как долго будет падать тело?

- Я кинула объект с высоты х
- Через какое время он упадет на Землю?
- Если я Галилео Галилей:
 Надо провести эксперименты и замерять время



https://en.wikipedia.org/wiki/Galileo%27 s_Leaning_Tower_of_Pisa_experiment

Задача: Как долго будет падать тело?

- Я кинула объект с высоты х
- Через какое время он упадет на Землю?
- Если я Галилео Галилей:
 Надо провести эксперименты и замерять время
- Яв 21 веке:

$$f(x) = \sqrt{\frac{2x}{g}}$$

Задача: Кто изображен на фото?



Задача: Анализ эмоциональной окраски текста

Какой из отзывов положительный?

- "Не могу сказать, что мне понравился суп. Но он лучше чем ничего"
- "Мне не не понравился этот банан"

Задача: Анализ эмоциональной окраски текста



৭ ≡

Computer Science > Computation and Language

arXiv:1704.05579 (cs)

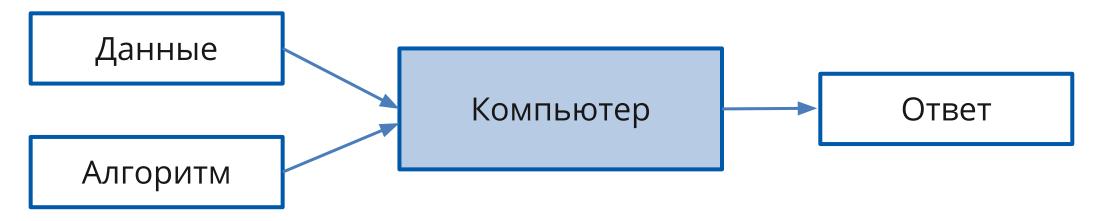
[Submitted on 19 Apr 2017 (v1), last revised 22 Mar 2018 (this version, v4)]

A Large Self-Annotated Corpus for Sarcasm

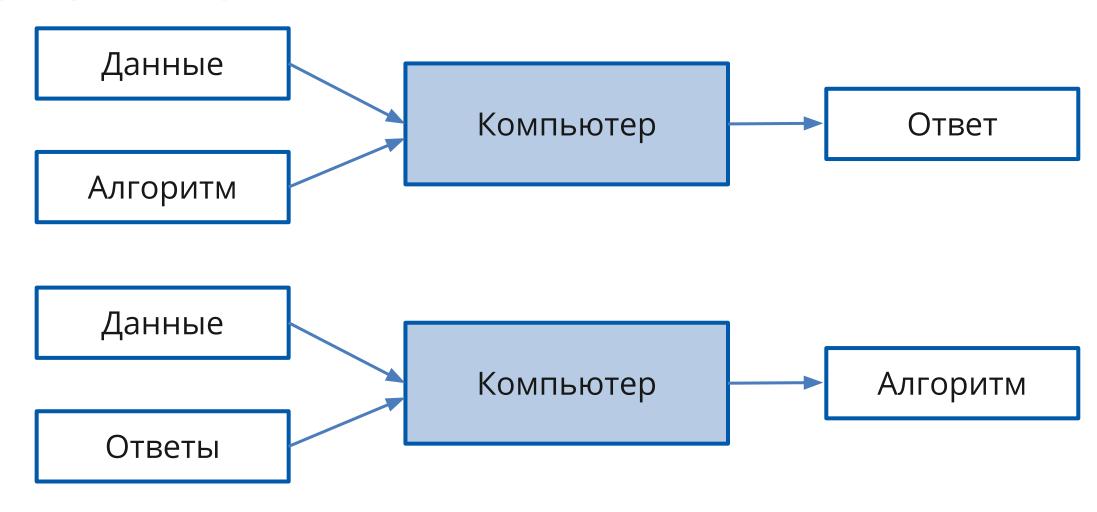
Mikhail Khodak, Nikunj Saunshi, Kiran Vodrahalli

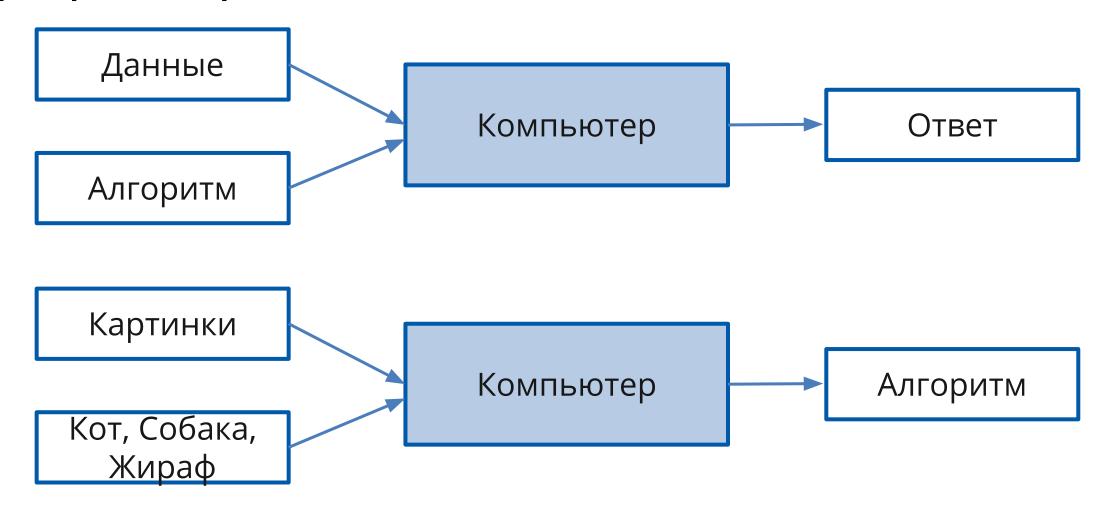
Download PDF

We introduce the Self-Annotated Reddit Corpus (SARC), a large corpus for sarcasm research and for training and evaluating systems for sarcasm detection. The corpus has 1.3 million sarcastic statements – 10 times more than any previous dataset — and many times more instances of non-sarcastic statements, allowing for learning in both balanced and unbalanced label regimes. Each statement is furthermore self-annotated — sarcasm is labeled by the author, not an independent annotator — and provided with user, topic, and conversation context. We evaluate the corpus for accuracy, construct benchmarks for sarcasm detection, and evaluate baseline methods.





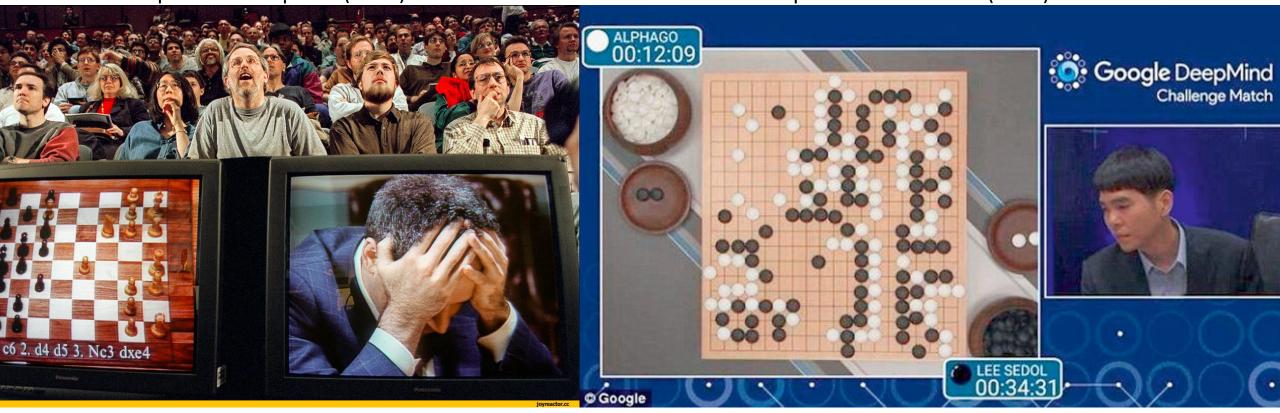




Задача: Шахматы против Го

Deep Blue vs Kasparov (1985)

AlphaGo vs Lee Sedol (2016)

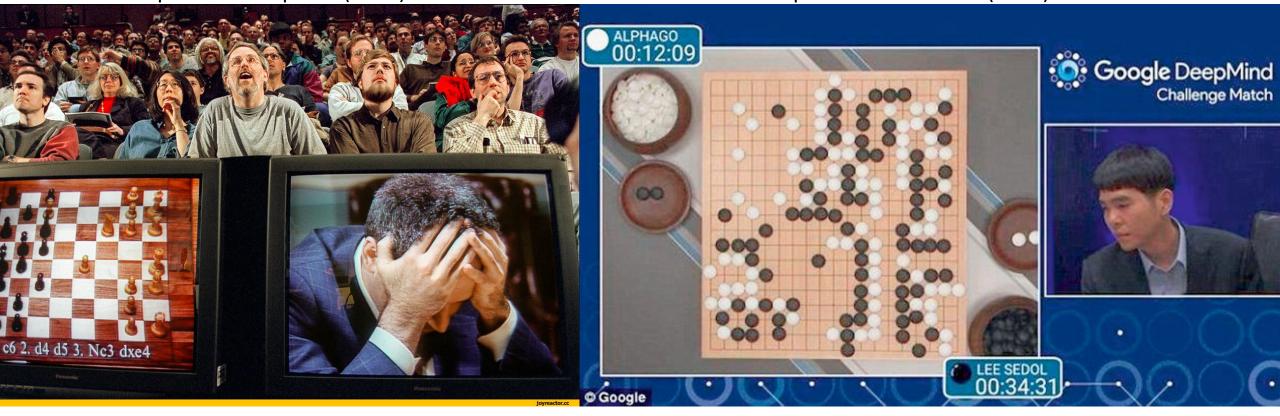


https://deepmind.com/research/case-studies/alphago-the-story-so-far

Почему сейчас?

Deep Blue vs Kasparov (1985)

AlphaGo vs Lee Sedol (2016)



https://deepmind.com/research/case-studies/alphago-the-story-so-far

Почему сейчас?

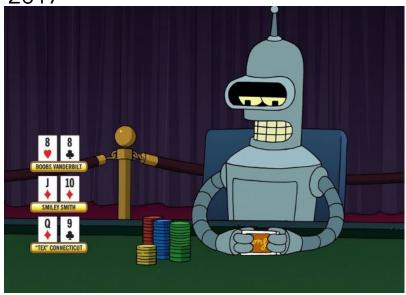
How Long Until Computers Have the Same Power As the Human Brain? Lake Michigan's volume (in fluid ounces) is about the same as our brain's capacity (in calculations per second). Computing power doubles every 18 months. At that rate, you see

very little progress for a long time—and suddenly you're finished.

1940 calcs/second Mother Jones

Еще примеры

2017



https://www.wired.com/2017/01/mystery-ai-just-crushed-best-human-players-poker/

Еще примеры

2017

8 8 BOBS VANDERBILT

J 10

SMILEY SMITH

Q 9

TEX' CONNECTICUIT

https://www.wired.com/2017/01/mystery-ai-just-crushed-best-human-players-poker/

2018



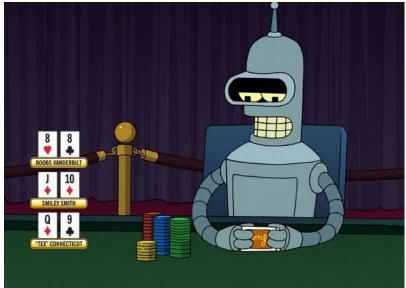
#AI bots just beat humans at the video game Dota 2. That's a big deal, because their victory required teamwork and collaboration – a huge milestone in advancing artificial intelligence.

via Twitter

https://openai.com/projects/five/

Еще примеры

2017



https://www.wired.com/2017/01/mystery-ai-just-crushed-best-human-players-poker/

Bill Gates
@BillGates

#AI bots just beat humans at the video game Dota 2. The

their victory required teamwork and collaboration – a huadvancing artificial intelligence.

via Twitter

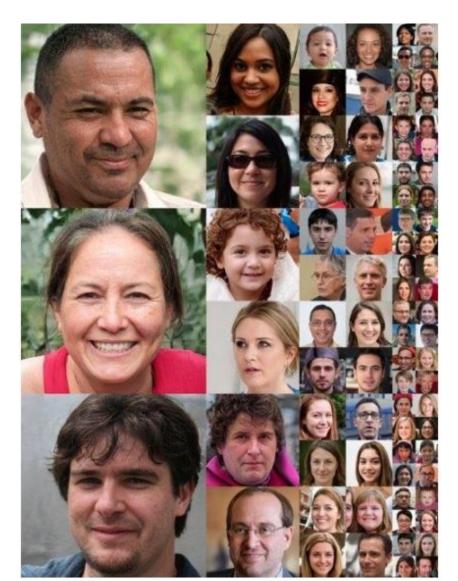
https://openai.com/projects/five/



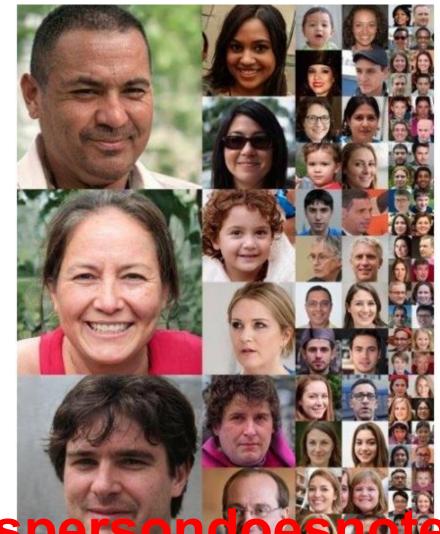
2019

https://deepmind.com/blog/article/alphast ar-mastering-real-time-strategy-game-st arcraft-ii

Не только игры Что общего у этих людей?

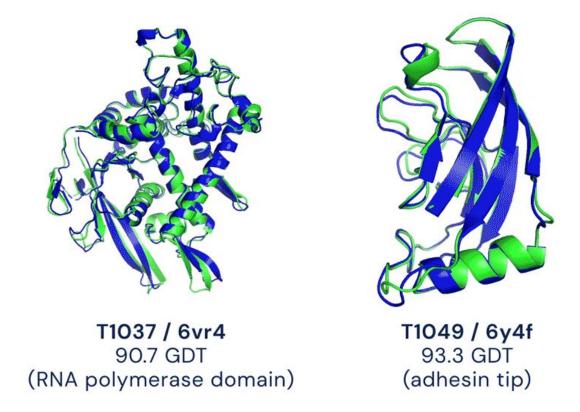


Не только игры Что общего у этих людей?



https://thispersondoesnotexist.com/

Не только игры Фолдинг белка

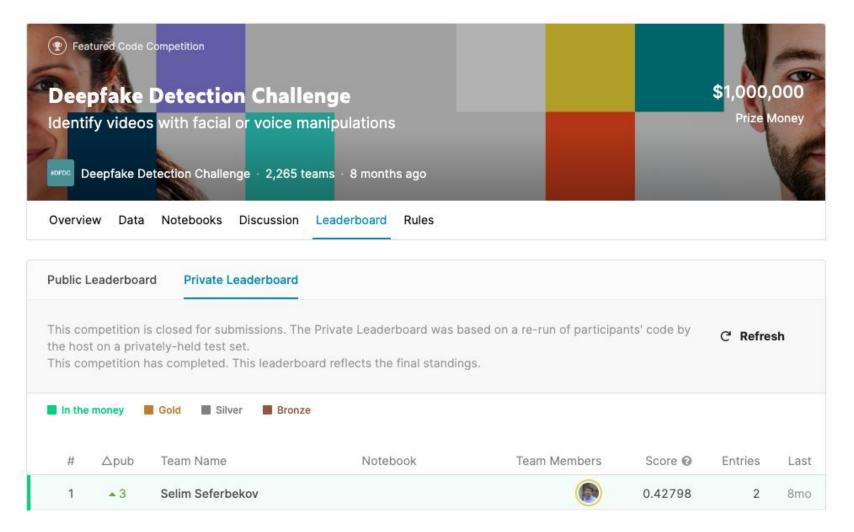


Experimental resultComputational prediction

Не только игры Deep Fakes



Не только игры Deep Fakes Detection



Искусственный Интеллект?

• Слабый (week)







• Сильный (general)



Искусственный Интеллект?

• Слабый (week)







◆ Сильный (general)



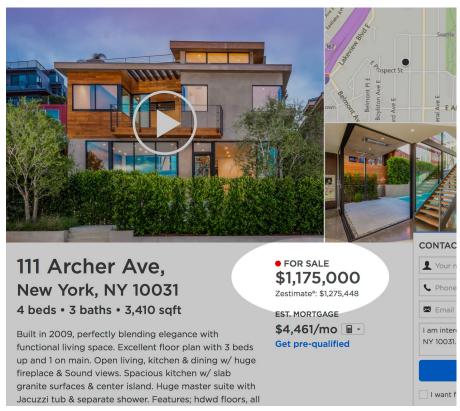
Основные определения

Задача: Сколько стоит дом?

Соревнования на Kaggle.com:

- Zillow's Home Value Prediction (\$25'000)
- Sberbank Russian Housing Market (\$12'000)

Хотим научиться предсказывать стоимость дома



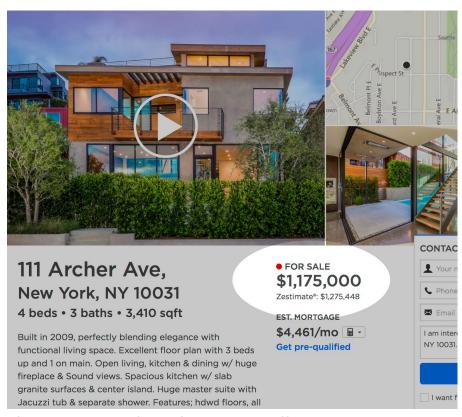
https://www.kaggle.com/c/zillow-prize-1/overview/competition-overview

Задача: Сколько стоит дом?

Соревнования на Kaggle.com:

- Zillow's Home Value Prediction (\$25'000)
- Sberbank Russian Housing Market (\$12'000)

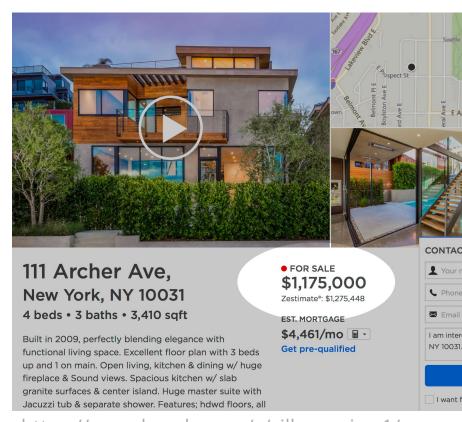
Хотим научиться предсказывать стоимость дома... посмотреть на 100500 других домов



https://www.kaggle.com/c/zillow-prize-1/overview/competition-overview

Обозначения

- *x* объект (sample)
 - Описание дома
- Ж пространство всех возможных объектов
 - Все возможные описание домов
- y ответ, целевая переменная (target)
 - Стоимость одного дома
- У пространство всех возможных ответов
 - Все положительные вещественные числа



https://www.kaggle.com/c/zillow-prize-1/overview/competition-overview

Обозначения: признаки

Что такое объект? Компьютеру нужны числа

Признаки (features) - числовые характеристики объектов

$$x = (x_1, x_2, ..., x_d)$$

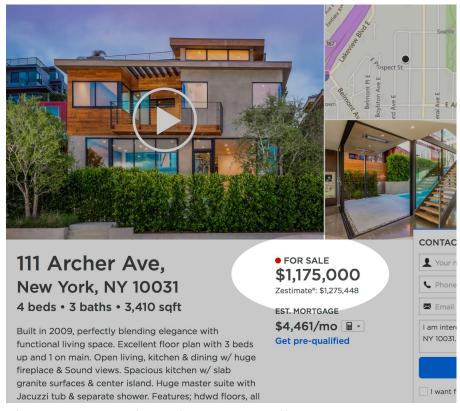
Обозначения: признаки

$$x = (x_1, x_2, ..., x_d)$$

Описание дома:

- Этаж
- Количество комнат
- Общая площадь
- Район
- Расстояние до метро

- ...



https://www.kaggle.com/c/zillow-prize-1/overview/competition-overview

Обозначения: обучающая выборка

Набор объектов с известными ответами (Training Dataset)

$$(x_i, y_i)_{i=1}^N$$

N - размер выборки

Обозначения: алгоритм

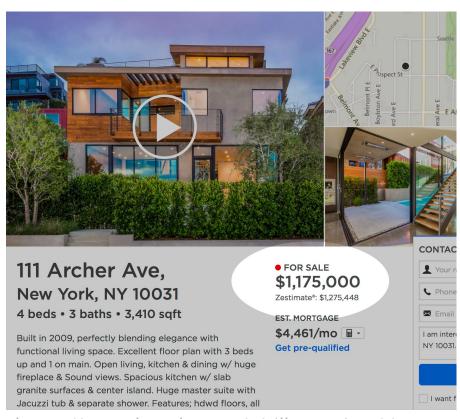
Функция, предсказывающая ответ для любого объекта - алгоритм (модель)

$$a(x): \mathbb{X} \to \mathbb{Y}$$

Обозначения: алгоритм

$$a(x): \mathbb{X} \to \mathbb{Y}$$

a(x) = 1′000′000 + 200′000 * (площадь) - 50′000 * (расстояние до метро)



https://www.kaggle.com/c/zillow-prize-1/overview/competition-overview

Обозначения: функция потерь

Функция потерь (Loss function) измеряет как "далеко" наше предсказание от правды.

$$L(a,x) = (a(x) - y)^2$$

Обозначения: функционал качества

Функционал качества, метрика качества — мера качества работы алгоритма на выборке

Среднеквадратичная ошибка (Mean Squared Error, MSE):

$$Q(a,X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

Обозначения: функционал качества

Функционал качества, метрика качества — мера качества работы алгоритма на выборке

Среднеквадратичная ошибка (Mean Squared Error, MSE):

$$Q(a,X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

- Должен соответствовать бизнес-требованиям
- Одна из самых важных составляющих анализа данных

Обозначения: обучение алгоритма

• Есть обучающая выборка и функционал качества

- Семейство алгоритмов 🖋
 - Из чего выбираем алгоритм
 - Пример: все линейные модели

$$\mathcal{A} = \{ w_0 + w_1 x_1 + \dots + w_d x_d \mid w_0, w_1, \dots, w_d \in \mathbb{R} \}$$

 Обучение: поиск оптимального алгоритма с точки зрения функционала качества

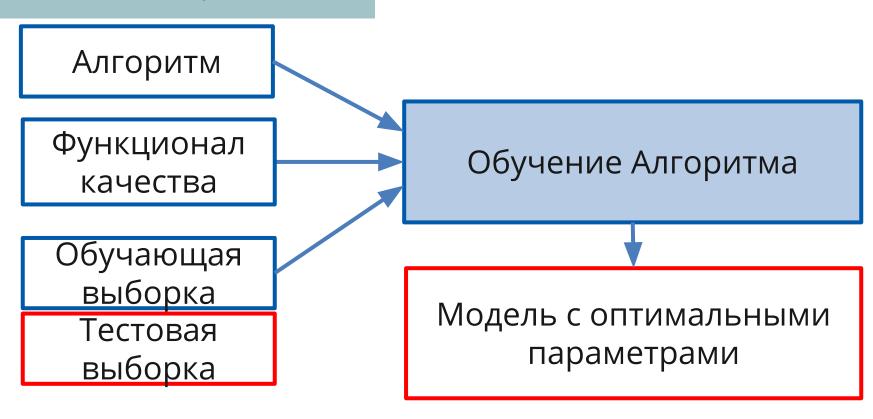
Алгоритм

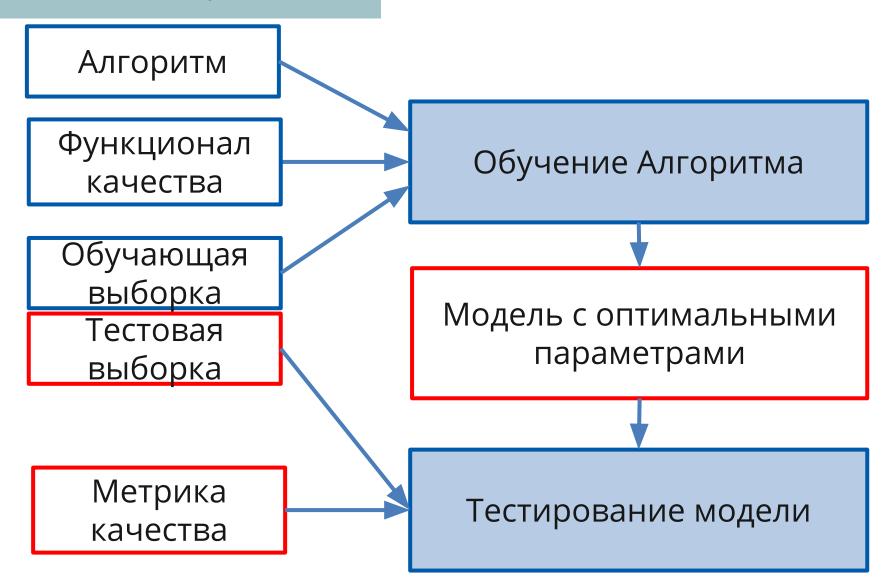
Функционал качества

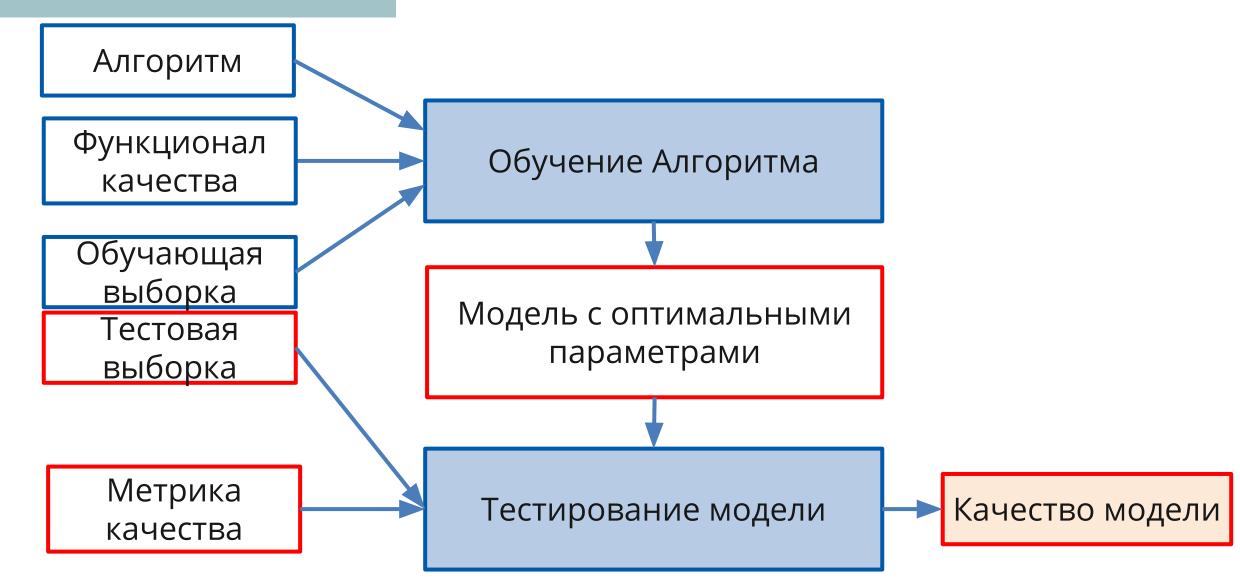
Обучение Алгоритма

Обучающая выборка

> Тестовая выборка





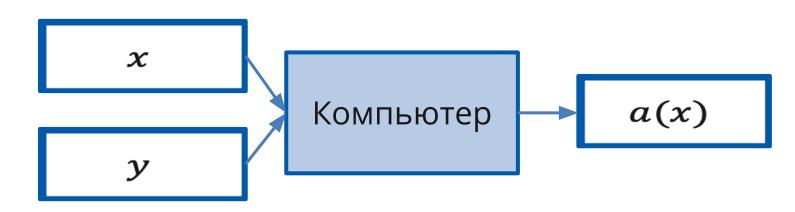


Типы Задач

Обучение с учителем и без

Supervised learning

• У нас есть "правильные ответы"



Unsupervised learning

• У нас есть только объекты



Классификация $Y = \{1, ..., K\}$





Asirra is a human interactive proof that asks users to identify photos of cats and dogs. It's powered by over three million photos from our unique partnership with Petfinder.com. Protect your web site with Asirra * free!



Score Test

You're a bot!

Computer accuracy = 60% Probability to guess= 0.6^{12} = 0.00217



Completed • Swag • 215 teams

Dogs vs. Cats

Wed 25 Sep 2013 - Sat 1 Feb 2014 (8 months ago)

Dashboard

Private Leaderboard - Dogs vs. Cats

This competition has completed. This leaderboard reflects the final standings.

See someone

#	Δ1w	Team Name * in the money	Score @	Entries	Last Submission UTC (Best - Las
1	-	Pierre Sermanet *	0.98914	5	Sat, 01 Feb 2014 21:43:19 (
2	↑26	orchid *	0.98309	17	Sat, 01 Feb 2014 23:52:30
3	-	Owen	0.98171	15	Sat, 01 Feb 2014 17:04:40 (
4	new	Paul Covington	0.98171	3	Sat, 01 Feb 2014 23:05:20

Computer accuracy = 98% Probability to guess = $0.98^{12} = 0.875$

Идеи задач классификации?

Пример 1: Credit scoring

https://www.kaggle.com/c/home-credit-default-risk

Пример 2: Sentiment analysis

https://pypi.org/project/dostoevsky/

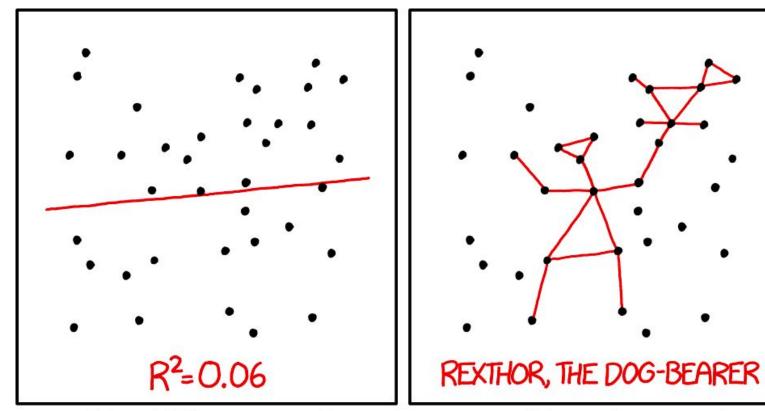
https://www.kaggle.com/c/tweet-sentiment-extraction

Пример 3: Image Segmentation

https://ai.facebook.com/blog/-detectron2-a-pytorch-based-modular-object-detection-library-/https://www.med.upenn.edu/cbica/brats2020/tasks.html

Регрессия

$$\mathbb{Y} = \mathbb{R}$$



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Идеи задач регрессии?

Пример 1: Predicting age

https://www.kaggle.com/c/trends-assessment-prediction https://www.how-old.net

Пример 2: Detection

https://ai.facebook.com/blog/-detectron2-a-pytorch-based-modular-object-detection-library-/

Кластеризация

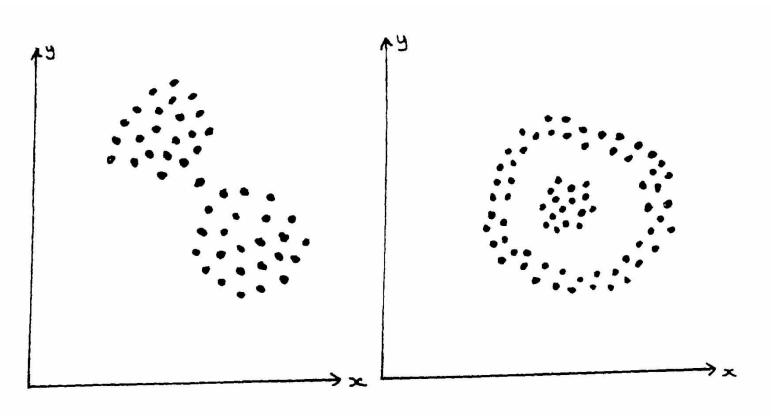
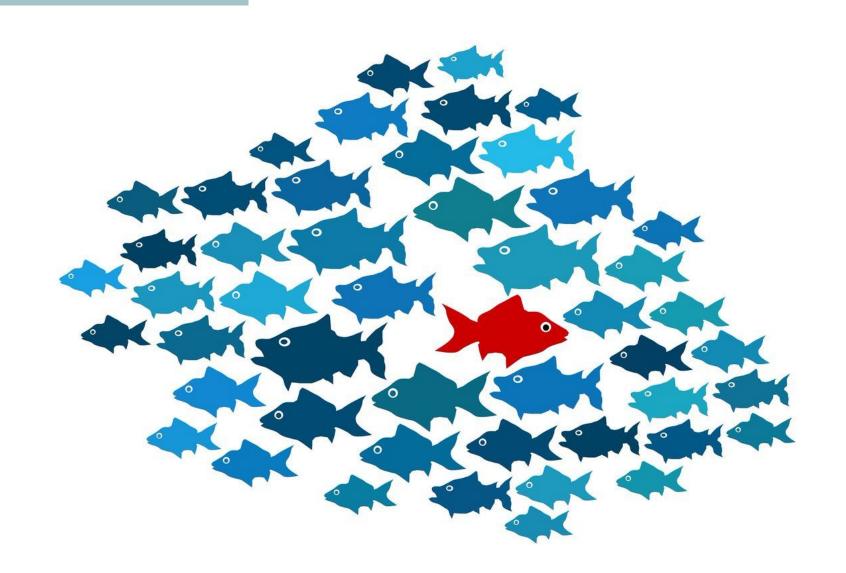


Figure 3.4 Further Examples of Clusters

Поиск Аномалий



Ранжирование



strong artificial intelligence



Найти





Поиск Картинки Видео Карты Маркет Новости Переводчик Эфир Кью

W Strong AI - Wikipedia

en.wikipedia.org > Strong Al ▼

Strong artificial intelligence or, True Al, may refer to: Artificial general intelligence, a hypothetical machine that exhibits behavior at least as skillful and flexible as humans do, and the research program of building such an artificial general i... Читать ещё >

W Сильный и слабый искусственные интеллекты...

ru.wikipedia.org > Сильный и слабый искусственные интеллекты v Сильный и слабый искусственные интеллекты — гипотеза в философии искусственного интеллекта, согласно которой некоторые формы искусственного интеллекта могут...

Deep learning programming, bitcoin, blockchain.

strongartificialintelligence.com v

Strong Artificial Intelligence is the born of new era for programming machines. Supercomputers need new language and different algorithms and we give the key for... Читать ещё >

What is Strong Al? | IBM

ibm.com > cloud/learn/strong-ai ▼

Strong artificial intelligence (AI), also known as artificial general intelligence (AGI) or general AI, is a theoretical form of AI used to describe a certain mindset of AI development. If researchers are able to develop **Strong** AI, the machine would require... Читать ещё >

Нашлось 3 млн результатов

21 показ в месяц

Дать объявление

В следующих сериях

- Типы признаков
- Линейная регрессия
- Решение задач оптимизации для обучения алгоритмов МО