Введение в АД

ThetaHat

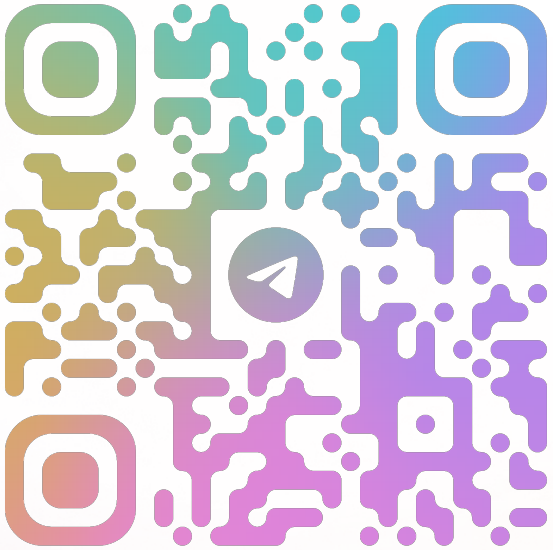
Проводимые нами учебные курсы

- Введение в анализ данных;
- DS-поток;
- Phystech@DataScience (в т.ч. статистика на ЛФИ);
- Статистика и машинное обучение на ФБМФ БТ;
- Мат. статистика на ФБМФ и ФЭФМ (семинары);
- Машинное обучение в ШАД;
- АБ-тестирование в ШАД.

Организационная информация

Телеграм-бот

@thetahat_ds26_bot



Код регистрации MAPE=0.072

Сайт команды thetahat.ru

Почта thetahat@yandex.ru

Чем отличаются задачи АД?

Сравним задачи

Алгоритмы и структуры данных

Задача: дан массив x, нужно его отсортировать.

Ровно один правильный ответ, можно
получить с помощью четких алгоритмов.

Комбинаторика

Задача: Сколько имеется способов раздать 11 разных цветков,
трём девушкам: какой-то – 5, а остальным – по 3 цветка? [ОКТЧ 2019]

Ровно один правильный ответ.

Анализ данных

Задача: Имеются данные $(x_1, y_1), ..., (x_n, y_n)$.
Восстановите по ним функцию $f : x \mapsto y$.

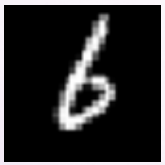
Особенности: нет четкого ответа, требуется только приближение,
но есть критерии качества.

Пример — распознавание рукописных цифр

Вход: 

Ожидается на выходе: **5**

Но как четко алгоритмически определить границу между 6 и 8?




 — 2 или 9?

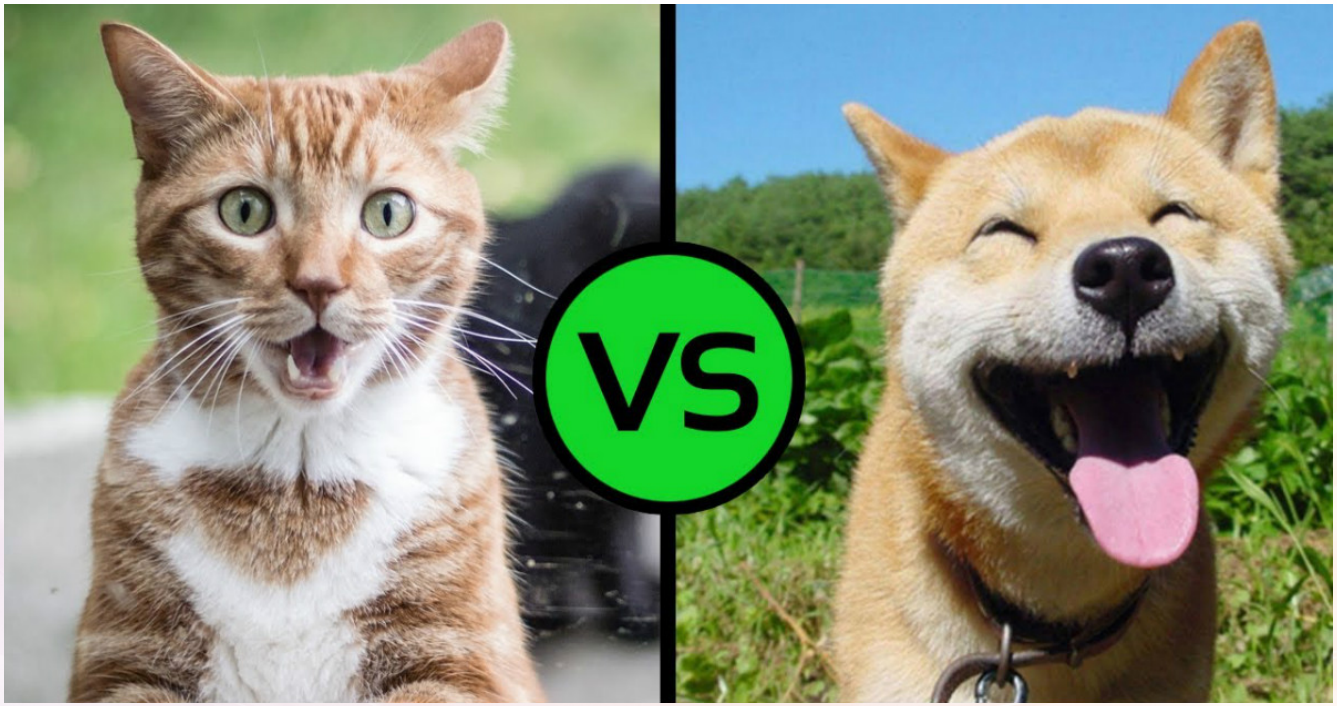
 — 4 или 7?

Кратко и на пальцах

Подробнее сегодня чуть позже

А ты кто?

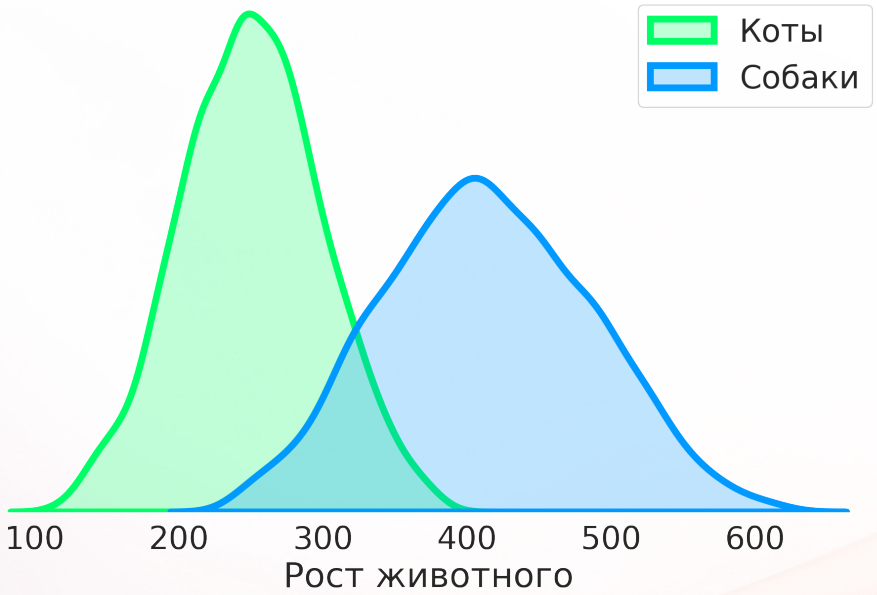
Перед нами домашнее животное. Кто это — собака или кот?



Классификация: собака vs кот

Попробуем сначала извлечь какой-то *признак*.

Построим вероятностные плотности для каждого класса.



При каких-то значениях роста мы уже можем

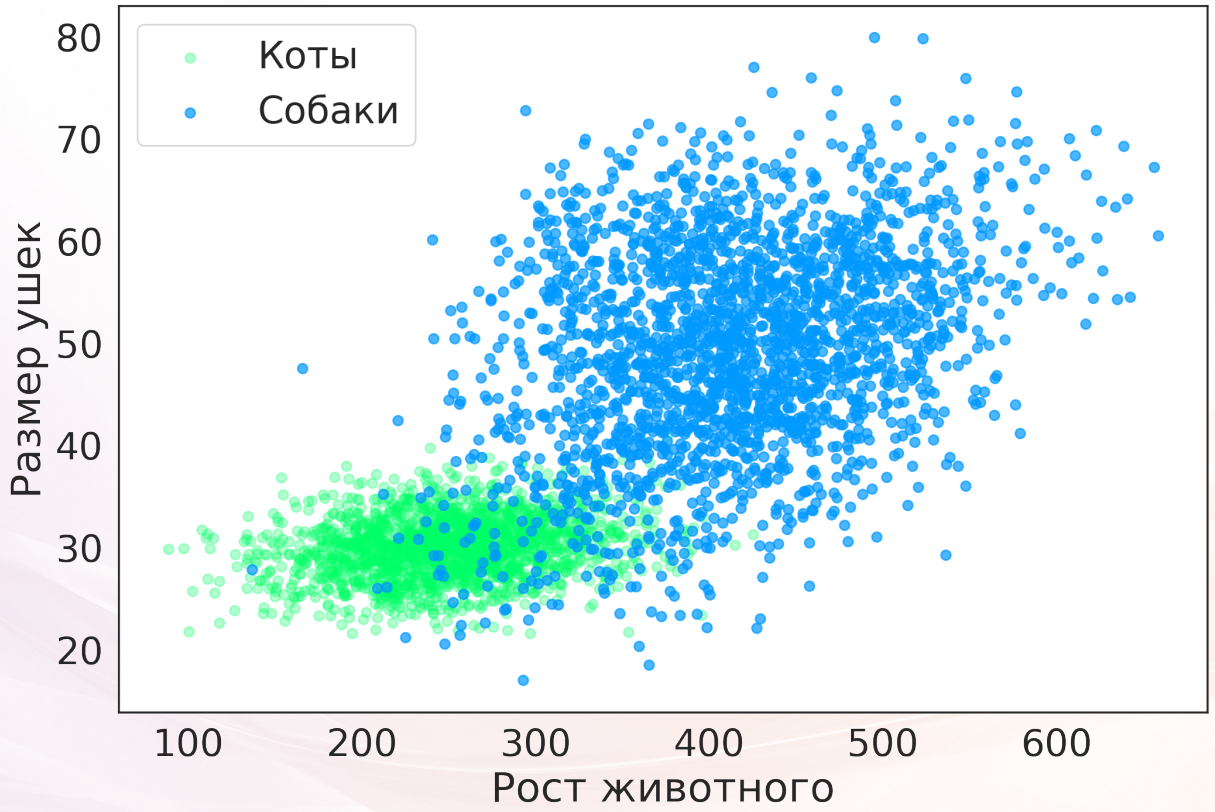
с большой уверенностью сказать ответ.

Но есть большое пересечение, это не очень здорово.

Классификация: собака vs кот

Извлечем еще один признак — размер ушек.

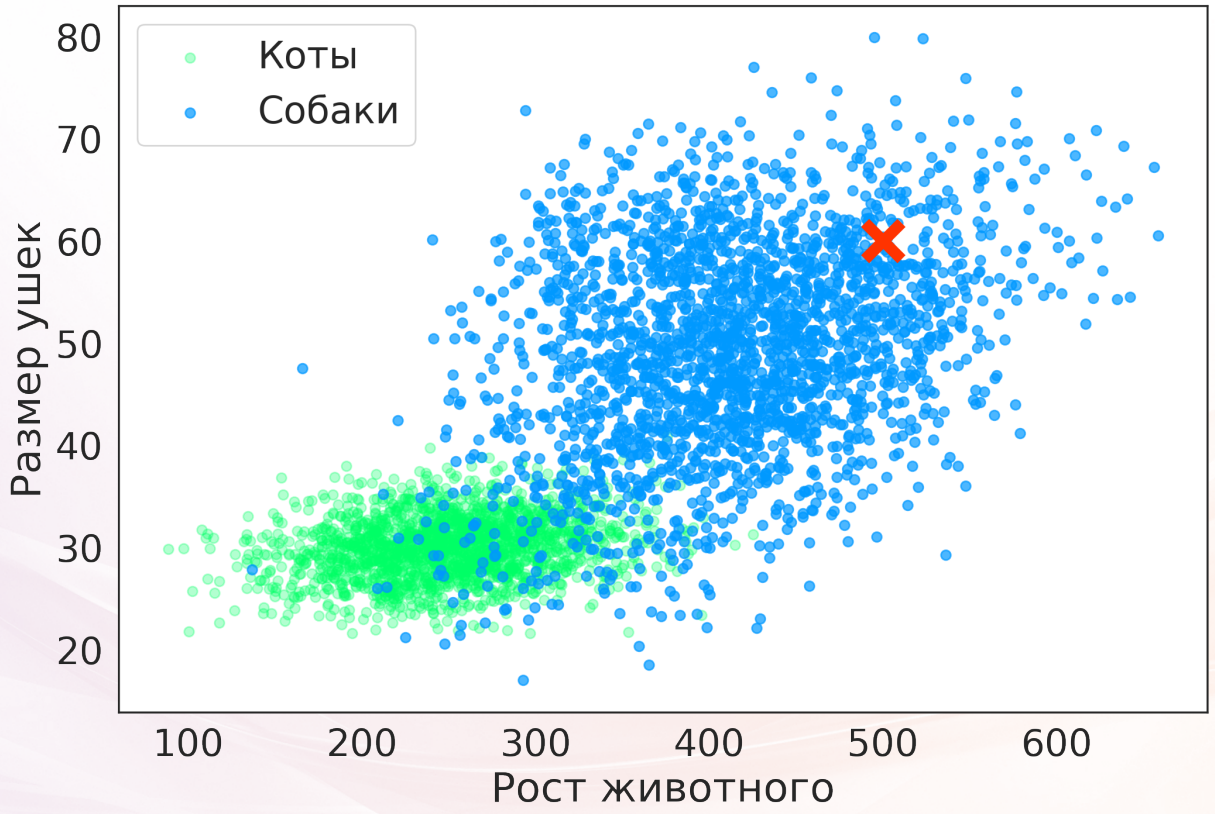
Теперь классы лучше разделяются.



Классификация: собака vs кот

Попробуем классифицировать новое животное.

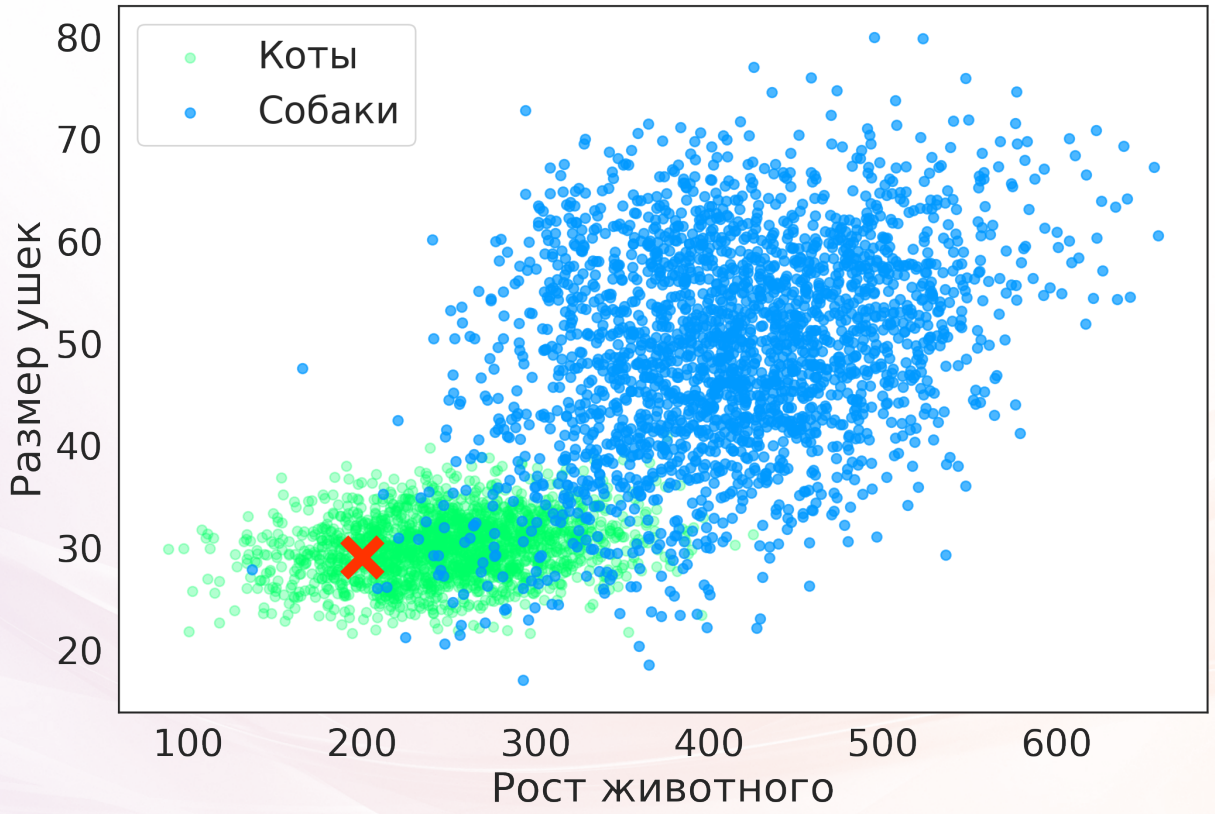
Кто отмечен красным?



Классификация: собака vs кот

Попробуем классифицировать новое животное.

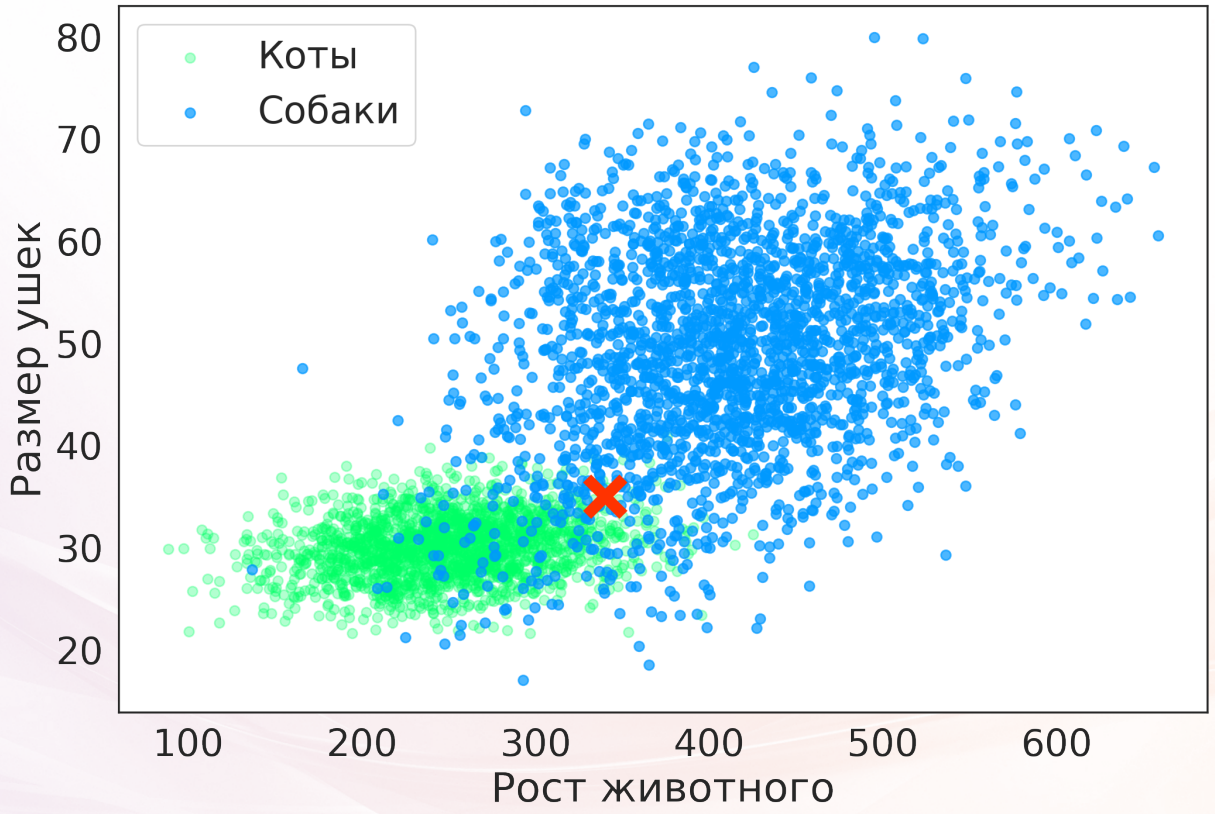
Кто отмечен красным?



Классификация: собака vs кот

Попробуем классифицировать новое животное.

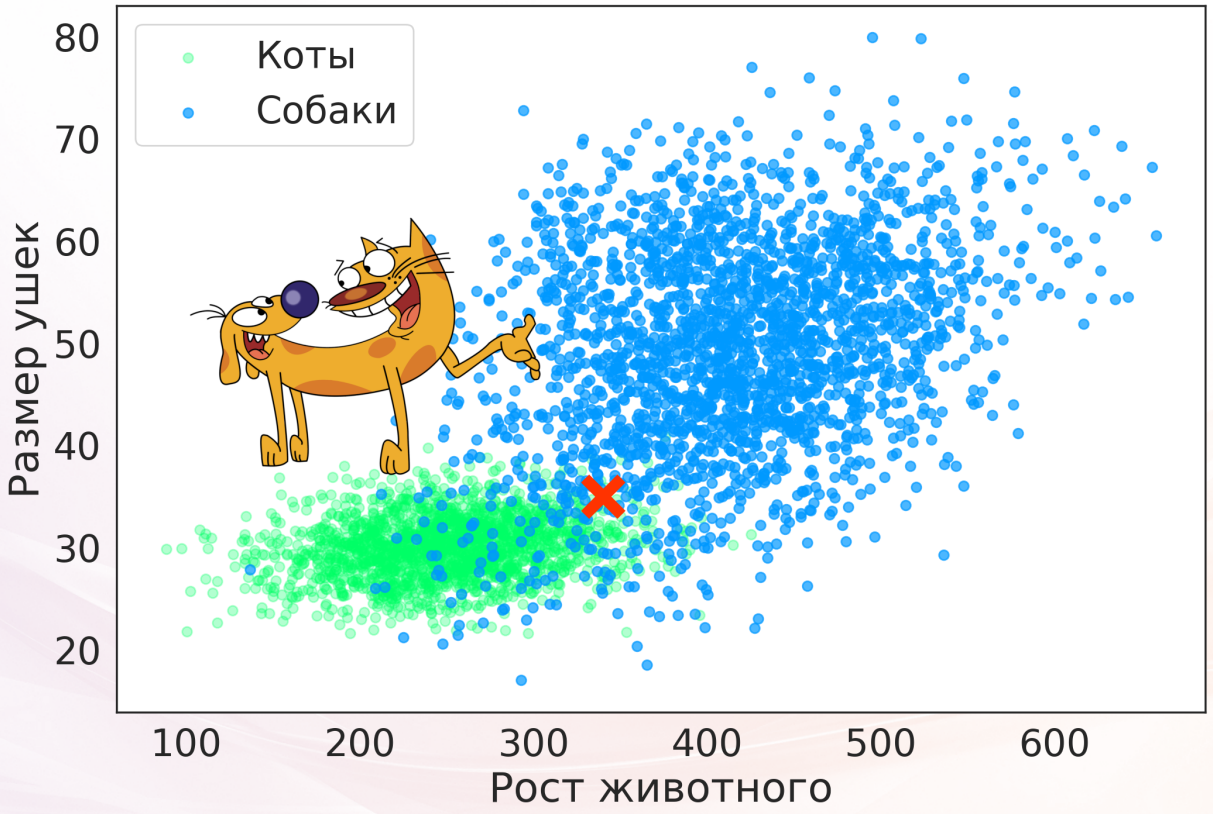
Кто отмечен красным?



Классификация: собака vs кот

Попробуем классифицировать новое животное.

Кто отмечен красным?



На основе чего вы сделали все выводы?

Метод ближайших соседей (kNN)

Дано:

$X_1, ..., X_n$ — набор размеченных объектов.

$Y_1, ..., Y_n$ — соответствующие метки класса.

Задача:

Пусть x — исследуемый объект. Какого он класса?

Решение:

Будем смотреть на свойства k ближайших соседей.

$x_{(1)}, ..., x_{(k)}$ — k его соседей в порядке удаления от x.

$Y_1, ..., Y_k$ — соответствующие им классы.

Ответ — наиболее часто встречаемый класс среди $x_{(1)}, ..., x_{(k)}$.

Свойства:

1. k — гиперпараметр модели;
2. Не редко на практике показывает хорошие результаты.
3. **Дорогое применение:**
   для каждого x результат вычисляется за $O(n \ln n)$.

Взвешенный метод ближайших соседей

Пусть x — исследуемый объект.

$x_{(1)}, ..., x_{(k)}$ — k его соседей в порядке удаления от x.

$Y_1, ..., Y_k$ — соответствующий класс.

$w_1, ..., w_k$ — вклад k-го соседа, определяемый пользователем.

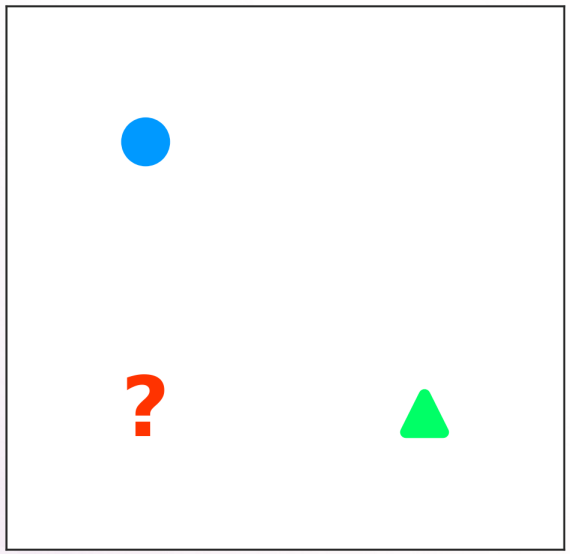
Способы определения веса:

- $w_j = 1 - j/k$ — зависящий от номера соседа;
- $w_j = \|x - x_{(j)}\|^{-1}$ — зависящий от расстояния до соседа.

$$\widehat{y}(x) = \arg\max_y \sum_{j=1}^{k} w_j I\{Y_j = y\} \text{ — классификация}$$

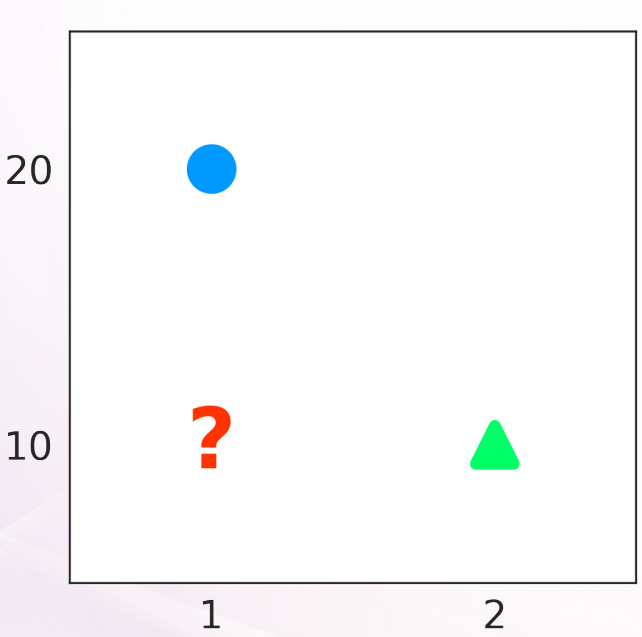
Особенности

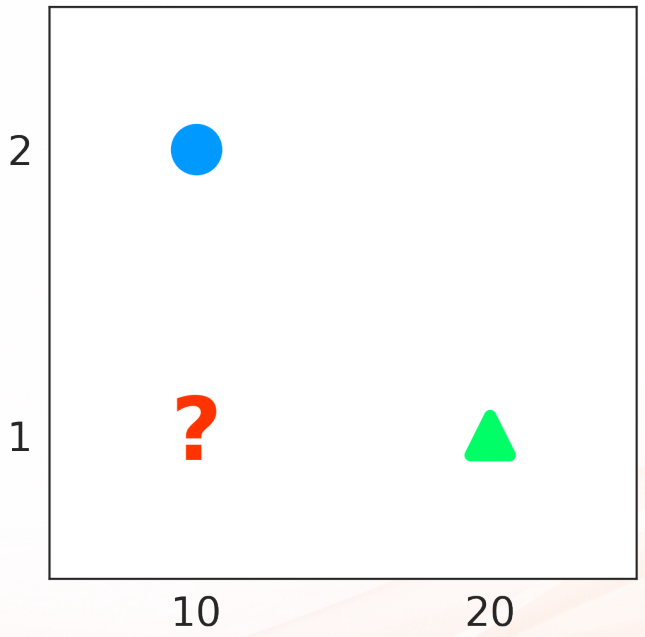
Классифицируйте объект "?".



Особенности

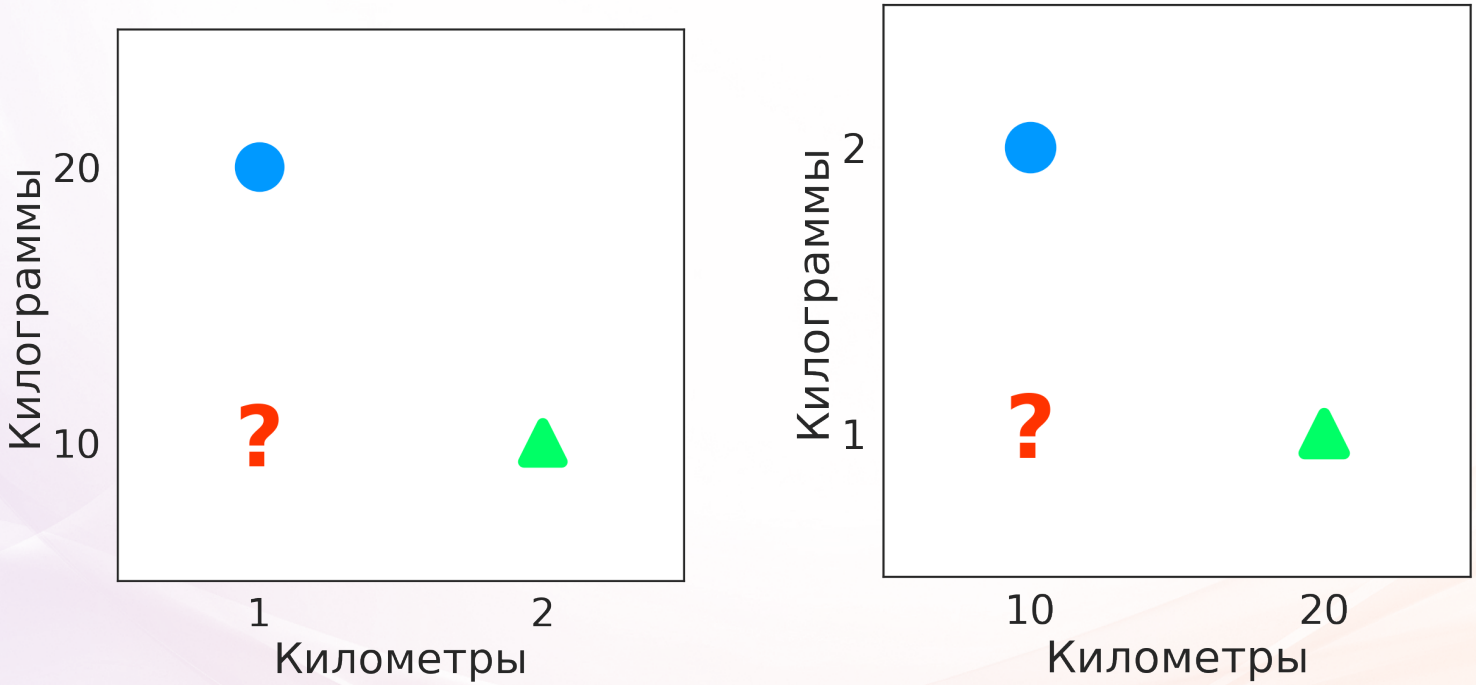
Классифицируйте объект "?".





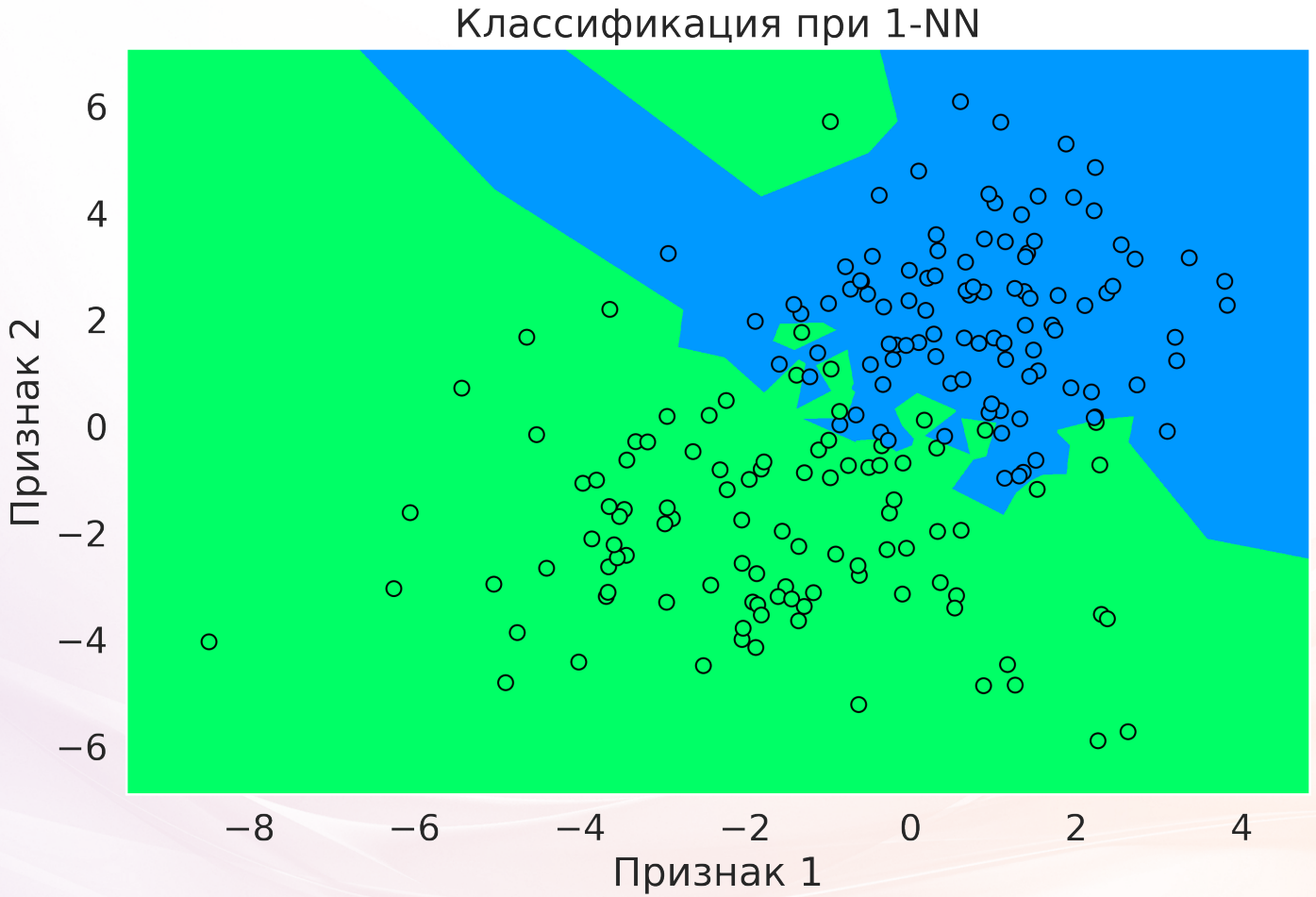
Особенности

Классифицируйте объект "?".

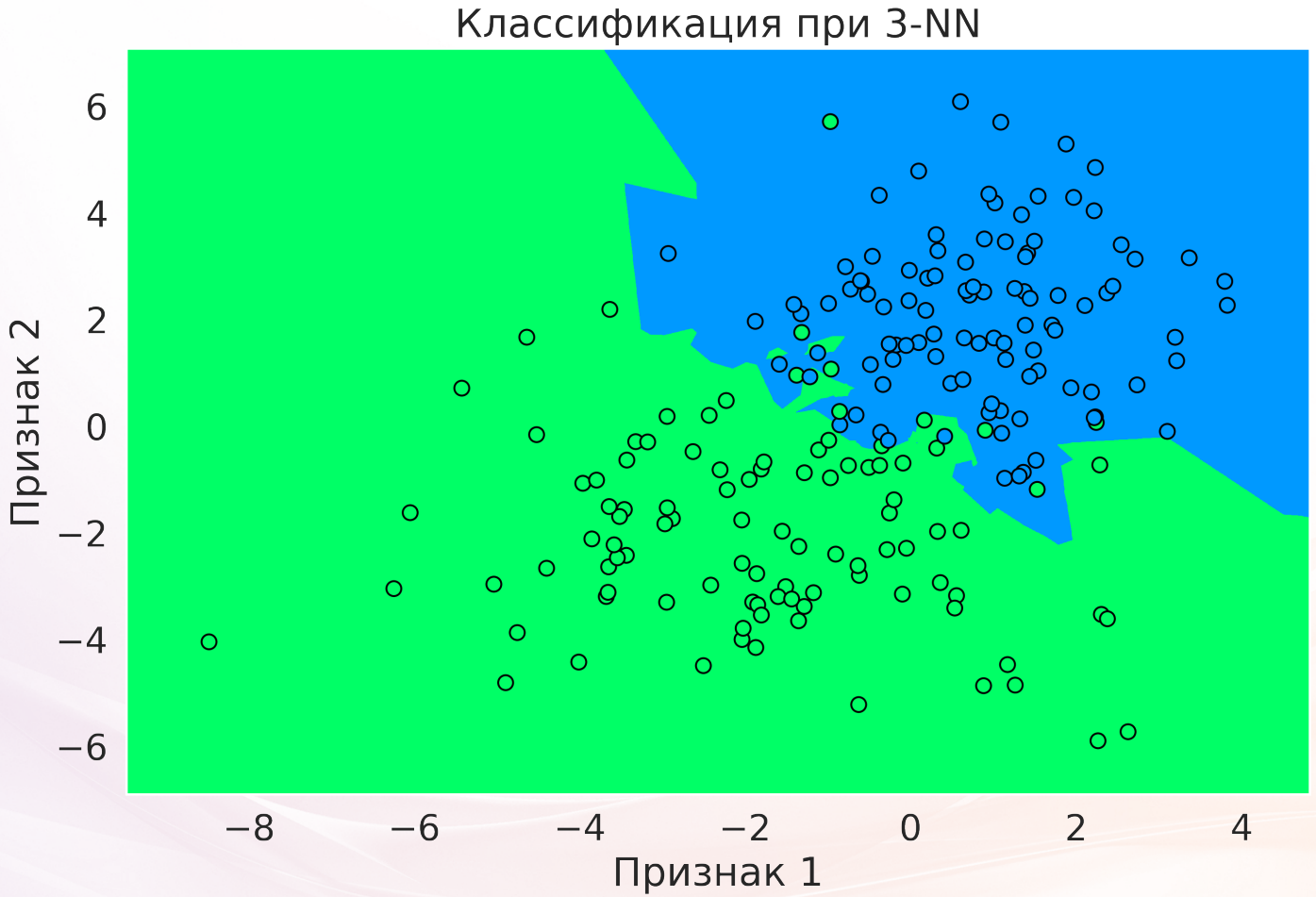


Вывод: результат сильно зависит от используемой метрики между точками в пространстве. Не складывайте *кг* с *км*!

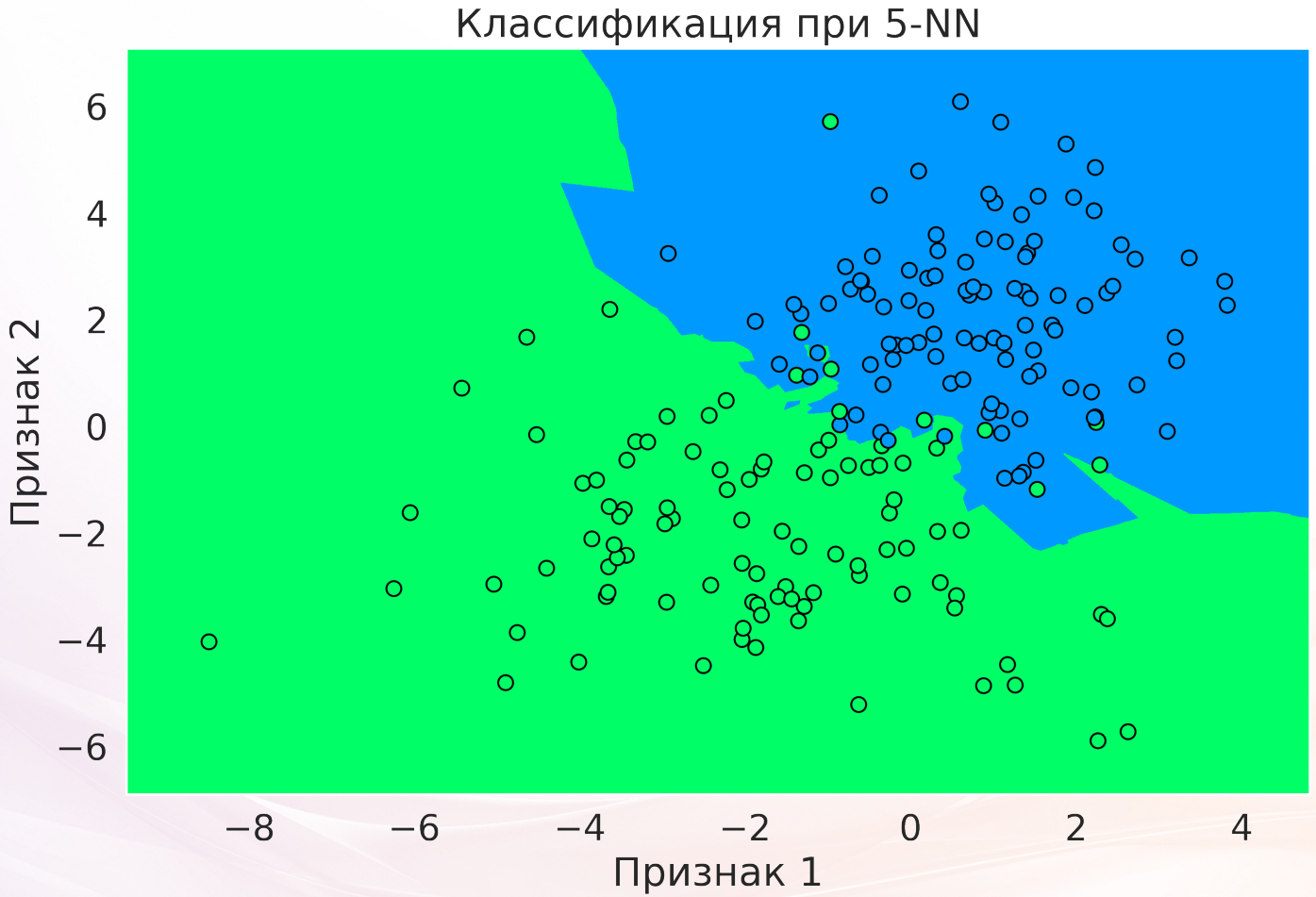
Что происходит при разных k?



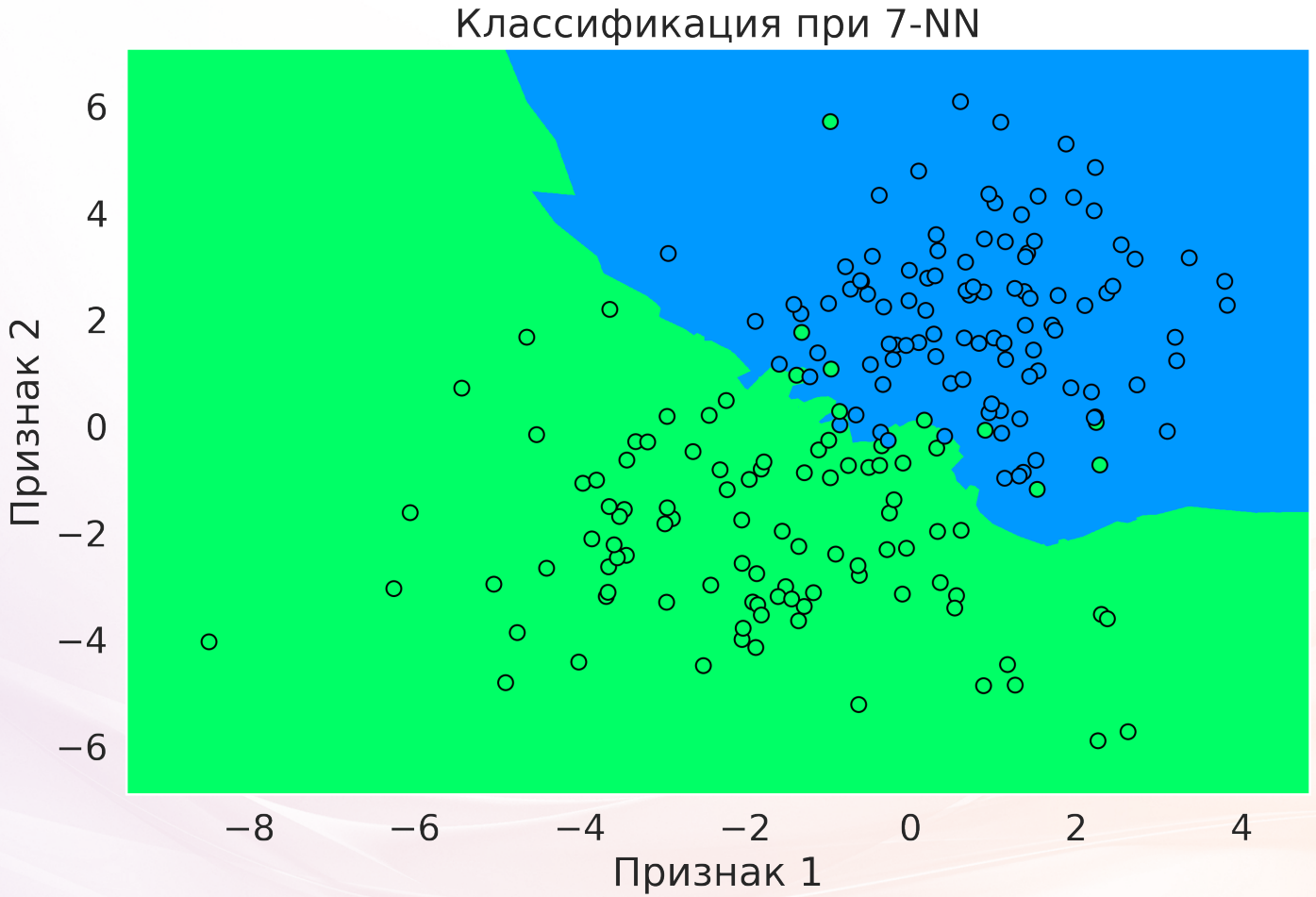
Что происходит при разных k?



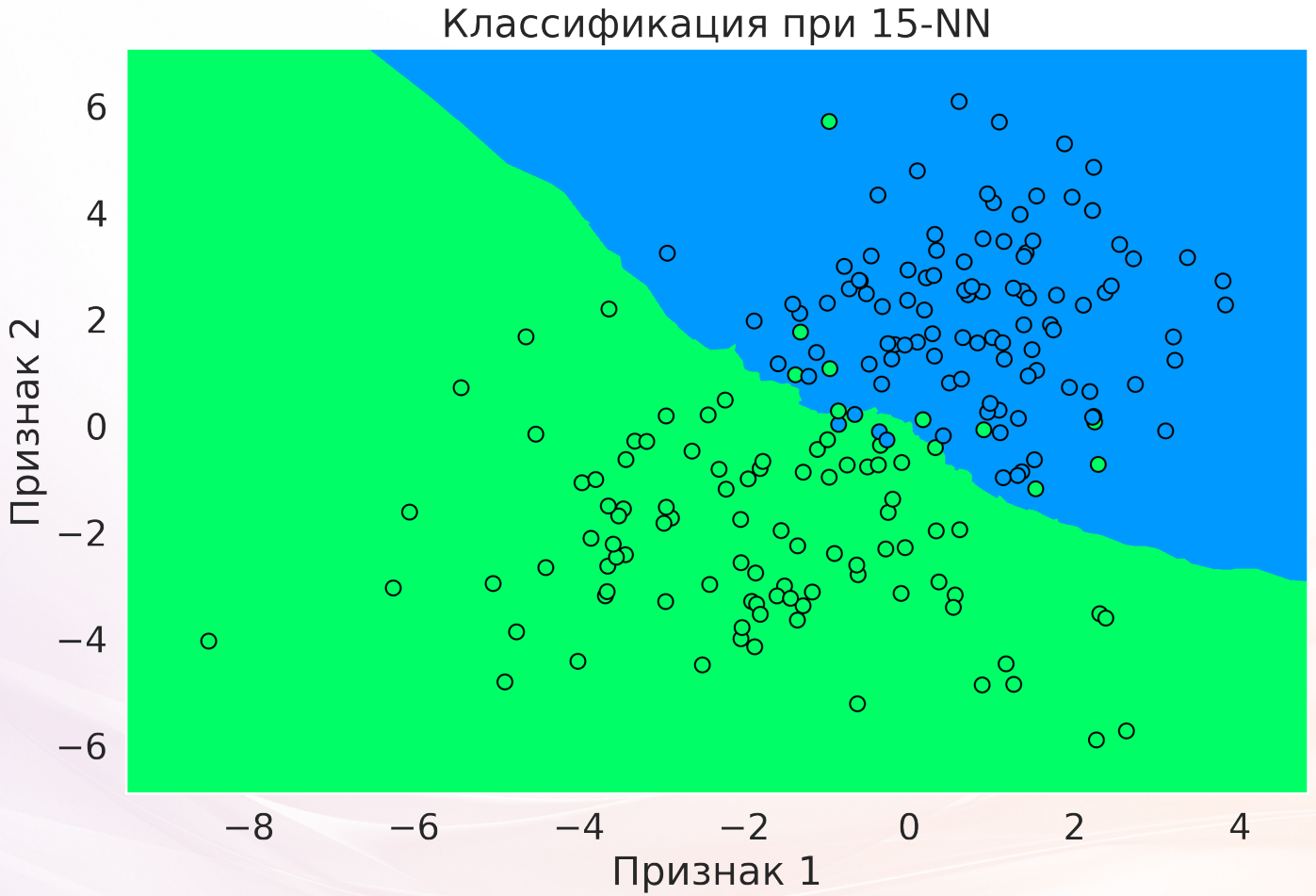
Что происходит при разных k?



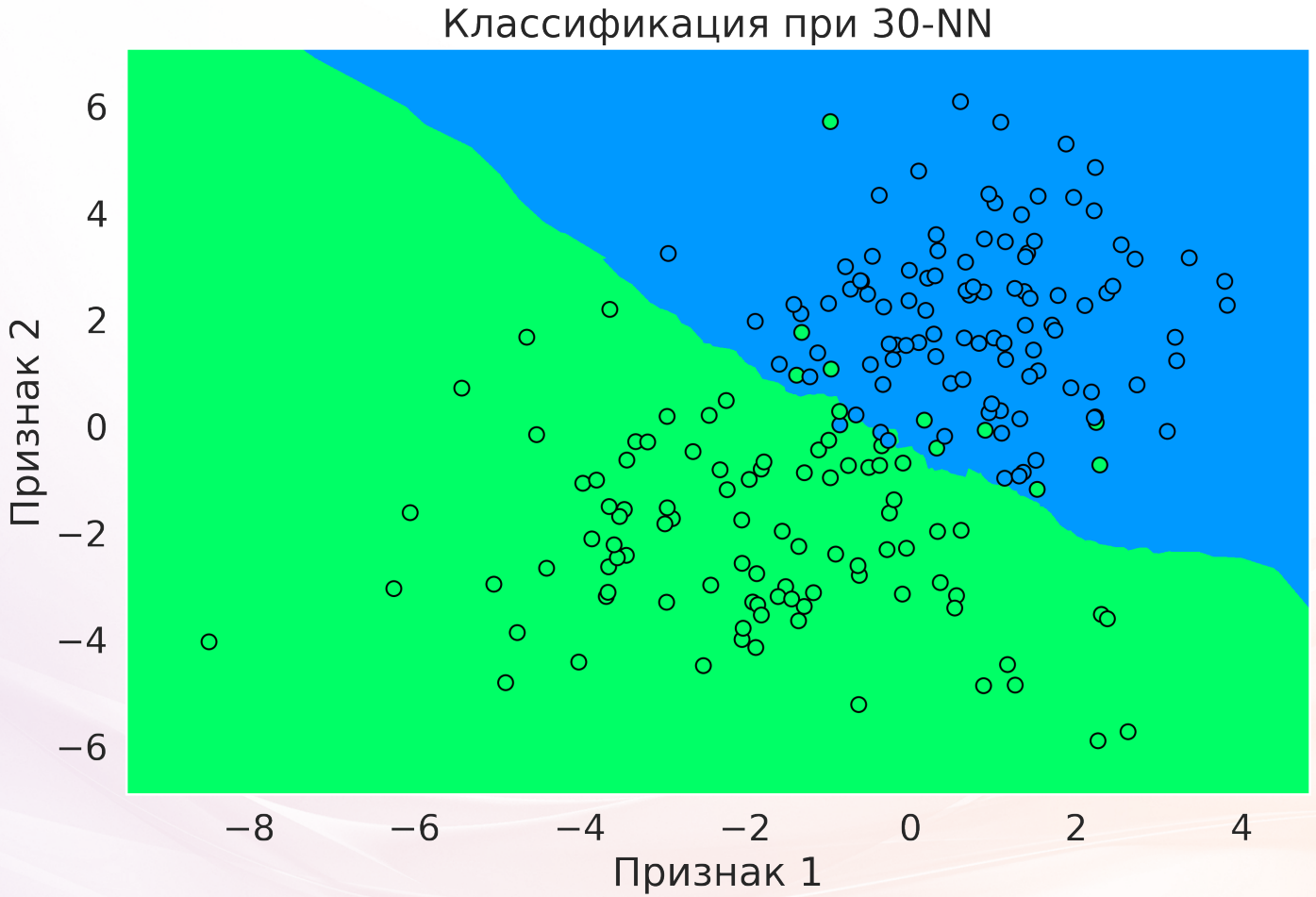
Что происходит при разных k?



Что происходит при разных k?



Что происходит при разных k?



Как оценить качество классификации?

Пусть $\widehat{y}(x)$ — оценка класса для объекта x.

Можем посчитать **точность** — доля правильно угаданных классов

$$A = \frac{1}{n}\sum_{i=1}^{n} I\{Y_i = \widehat{y}(x_i)\}$$

Оценка качества называется **метрикой** (*не путать с метр. пр-вами*).

Какое число соседей оптимизирует эту метрику?

Ответ: $k = 1$, т.к. при вычислении $\widehat{y}(x_i)$ берем сам Y_i.

Поэтому данные делят случайно на **две непересекающие части**:

1. на одной определяют правило классификации,
2. на другой — считают оценку качества классификации.

Точность 90% это много или мало?

Кажется, круто. А если в данных 85% котов? Тогда отвечая всегда "кот" сможем добиться точности 85%, и 90% уже не так круто...

А что если по картинке?

Хорошо, но что если объект — изображение кота или собаки?

Изображение 100×100 состоит из 10^4 пикселей,

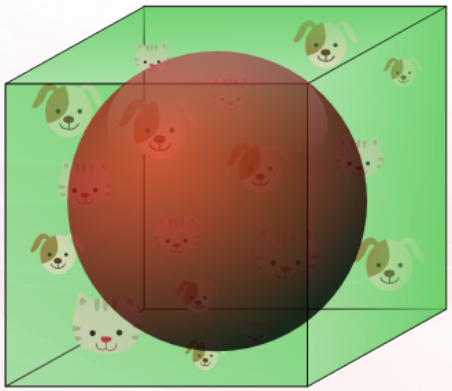
в каждом по 3 числа. Какой размерности получается объект?

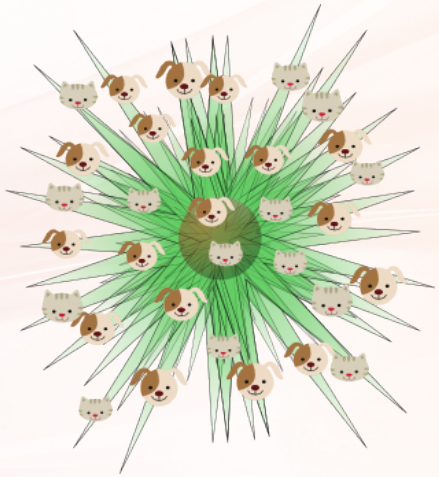
Ответ: $100 \times 100 \times 3 = 30\,000$ чисел в одной картинке.

Проблема:

в пр-ве больших размерностей расстояния неинформативны.

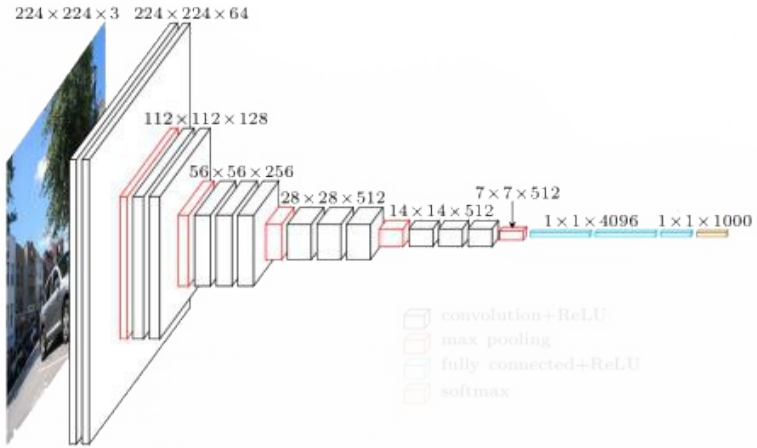
Например, среди фиксированного количества случайных точек
в единичном кубе в пространстве большой размерности
почти все точки будут лежать около границы куба.

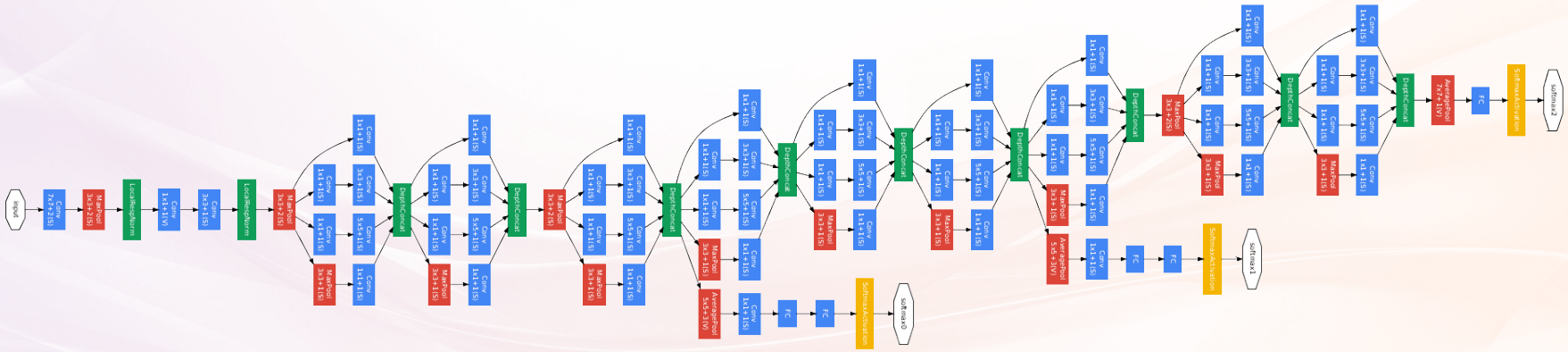




А что в сложных случаях?

Нейросети! Но об этом позже :)





Сбор и разметка данных

Откуда получаются обучающие датасеты?

1. **Данные о продажах**

   Собираем статистику покупок по товарам.

2. **Работа оборудования, инженерные системы**

   Ставим датчики, пишем информацию.

3. **Рекомендательные системы**

   Собираем информацию о действиях пользователей.

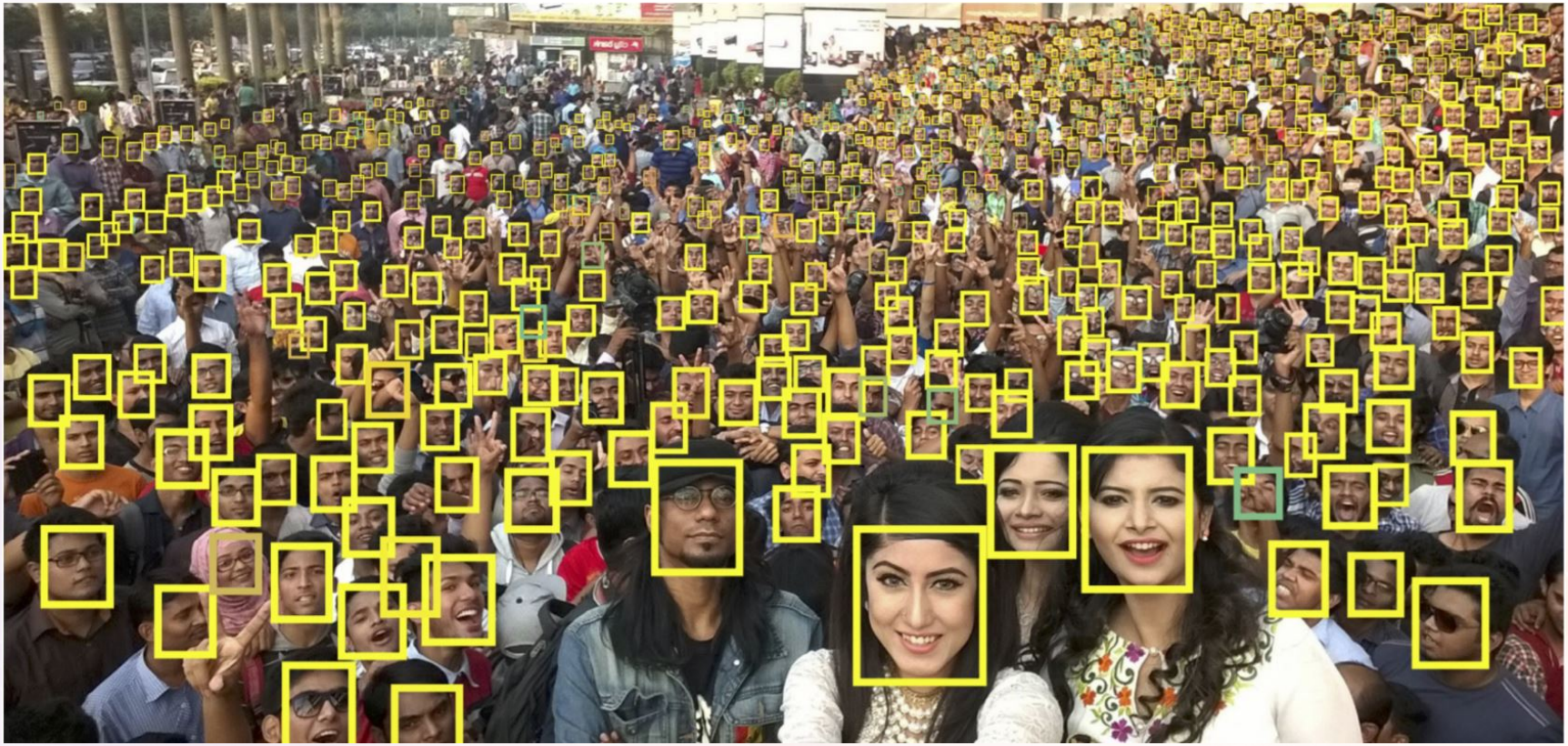
4. **Лабораторные испытания, AB-тестирование**

   Проводим эксперименты над разными группами.

Откуда получаются обучающие датасеты?

Распознавание лиц

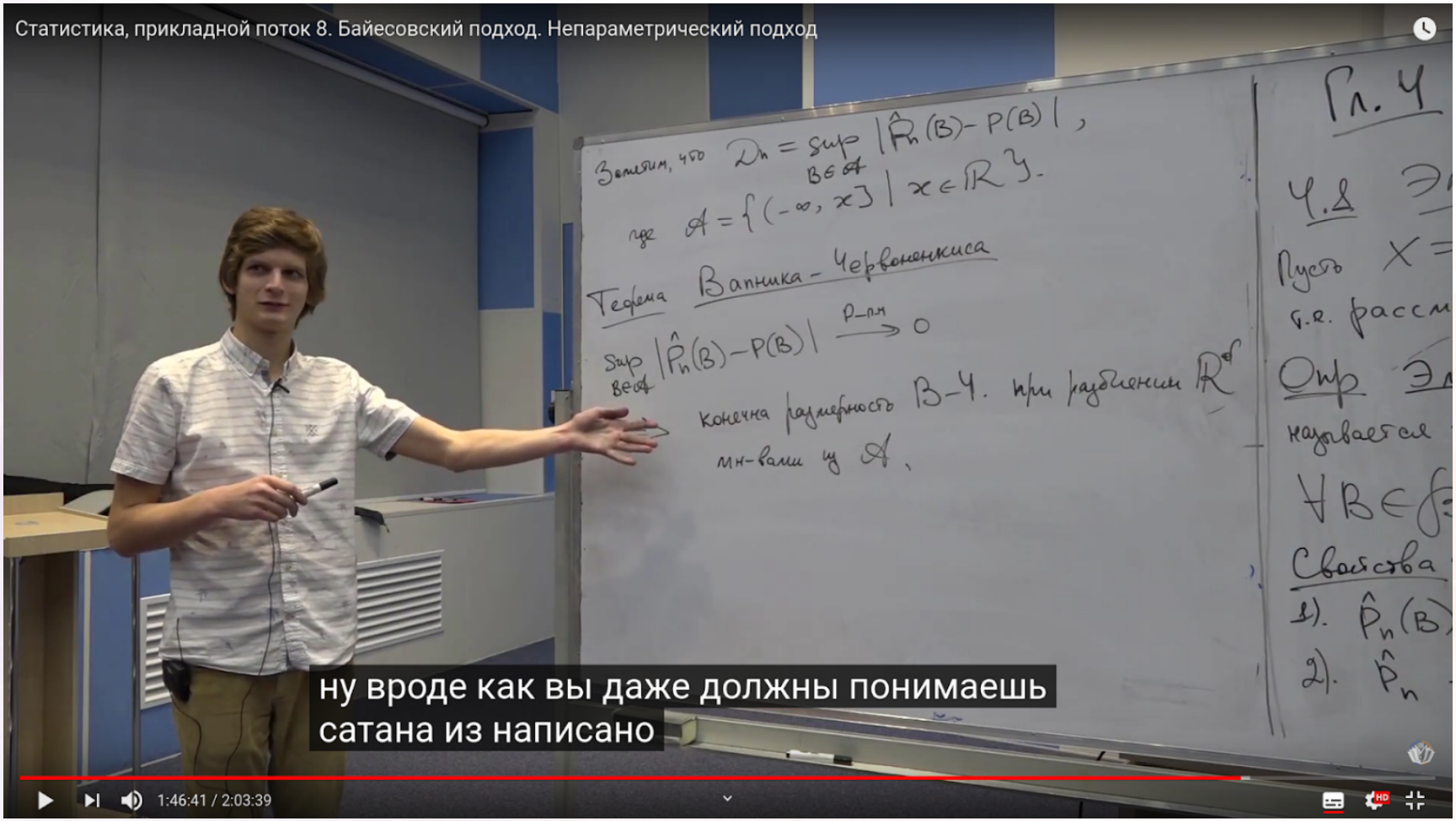
Нужно разметить лица вручную.



Откуда получаются обучающие датасеты?

Распознавание речи

Нужно прослушать и законспектировать много текстов.



Данные для обучения

Пример 1. Одна из основных задач **компьютерного зрения** — задача распознавания объектов. Рассмотрим эту задачу в контексте беспилотного транспорта.

Откуда мы возьмем данные?

Открытые источники: KITTI, nuScenes, Cityscapes и др...

Плюсы

- бесплатно
- достаточно хорошее качество

Минусы

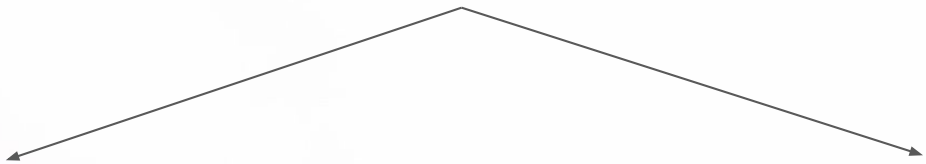
- другая классификация объектов
- сильно отличаются от продовых данных

nuScenes

Данные для обучения

Собственные данные решают проблемы открытых датасетов, но требуют много ресурсов на сбор данных и разметку.

Разметка собственных данных



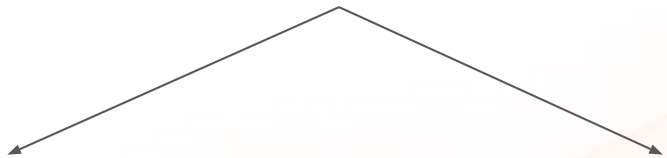
Самостоятельная разметка

Среда для разметки:

- Label Studio
- CVAT
- …

Разметчики и валидаторы

Аутсорс разметка



Сторонняя компания

~ те же методы, что и в случае самост. разметки

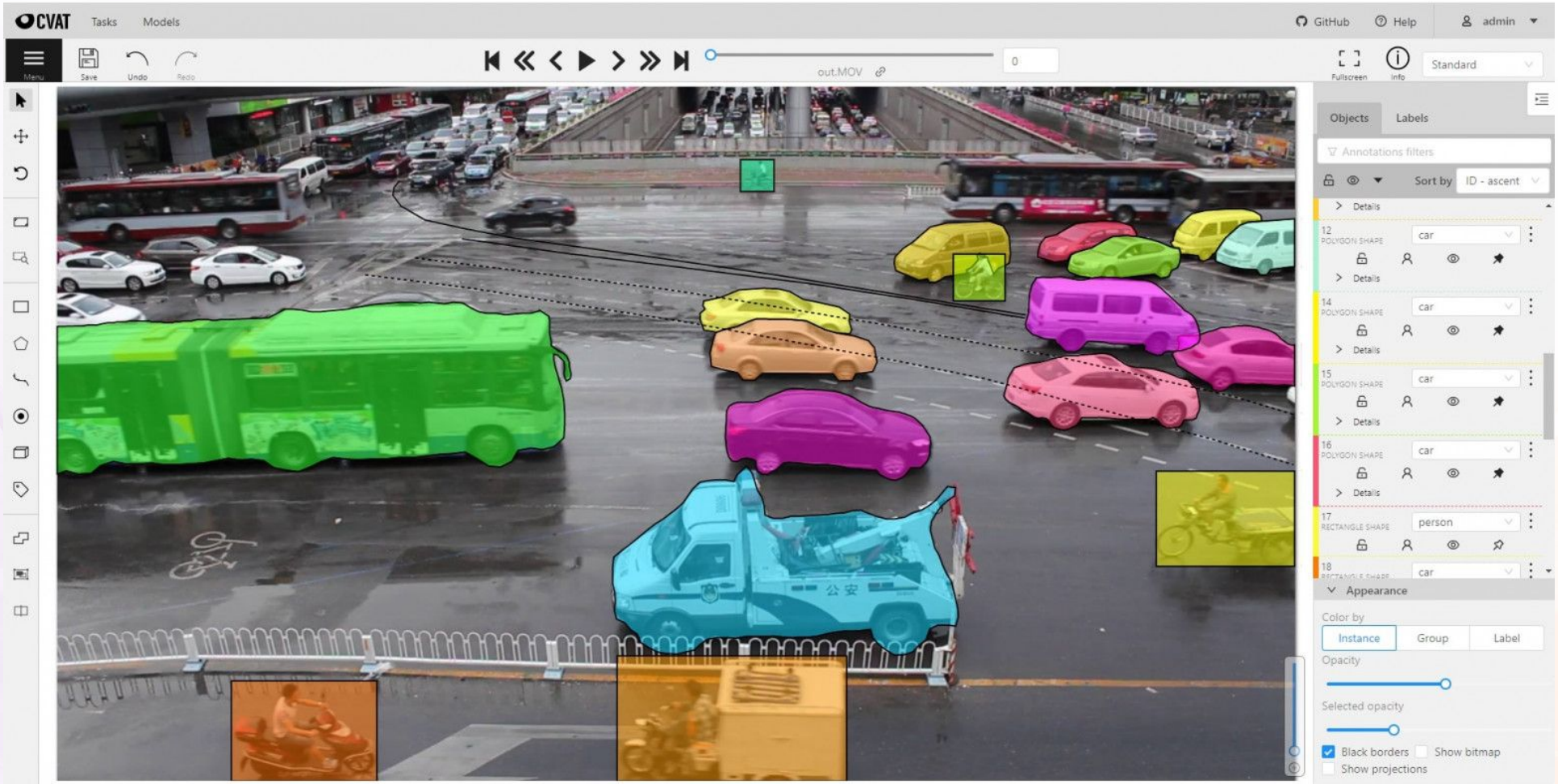
Краудсорсинг

Краудсорс. платформа:

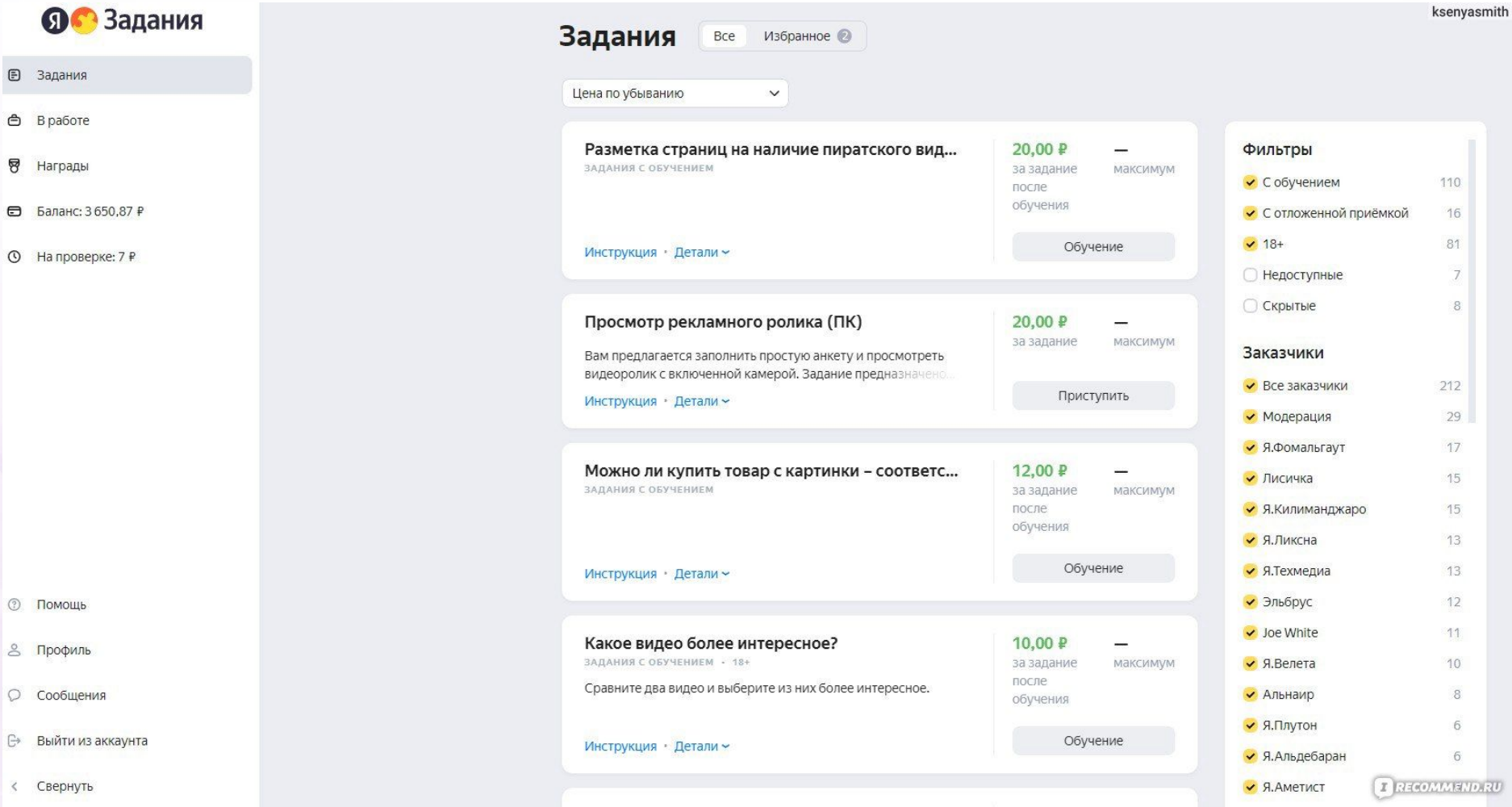
- Яндекс Задания
- …

Дешево, но нужна более сложная валидация, т.к. не знаем исполнителей

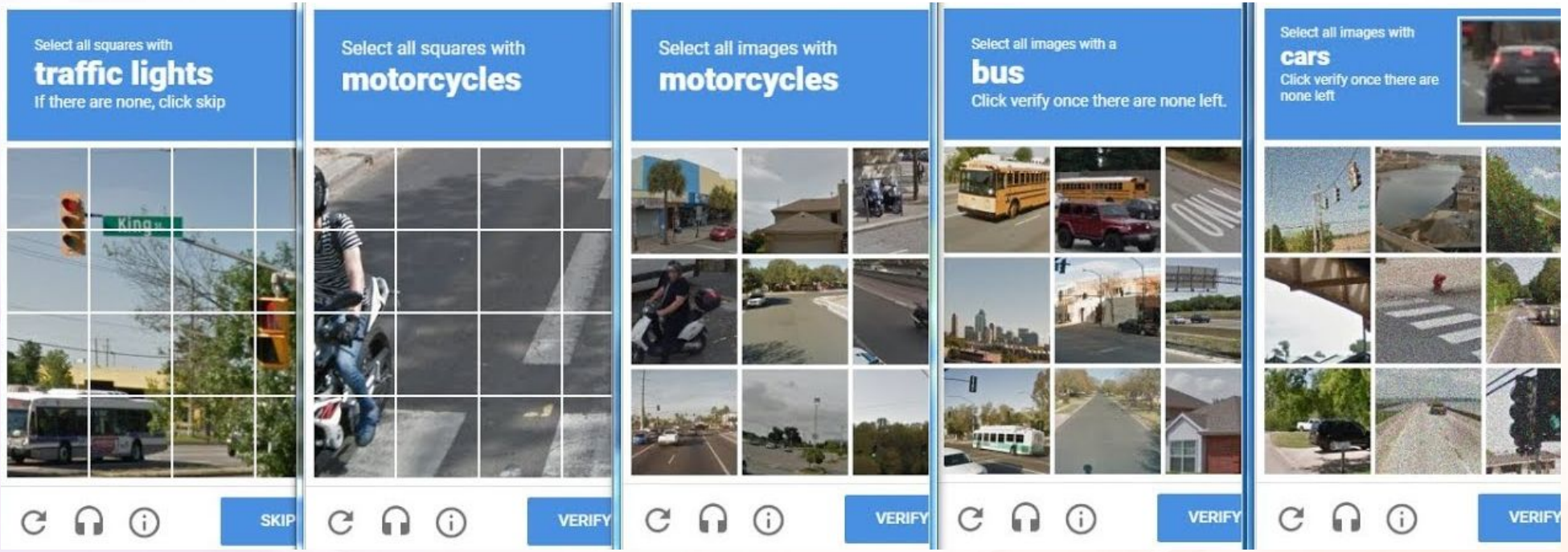
Разметка в CVAT



Пример задач в Яндекс Заданиях



Капча


Select all squares with
traffic lights
If there are none, click skip
SKIP
Select all squares with
motorcycles
VERIFY
Select all images with
motorcycles
VERIFY
Select all images with a
bus
Click verify once there are none left.
VERIFY
Select all images with
cars
Click verify once there are none left
VERIFY

Данные для обучения

Пример 2. Рассмотрим область **обработки естественного языка.**
Задача: научить ИИ понимать жалобы клиентов.

Данные для обучения

Собственные данные

Источники

- переписки службы поддержки
- письма клиентов
- обращения из приложения или сайта

Примеры

«Посылка не пришла»
«Мне списали деньги два раза»
«Где мой заказ?»

Кто собирает

сотрудники компании, аналитики, ML-инженеры

Особенности

- очень точные и релевантные данные
- полностью соответствуют нашей задаче
- обычно таких данных немного

Внешние данные

Источники

- форумы, сайты с отзывами, маркетплейсы
- открытые датасеты

Затем применяем фильтры:

- оставляем только тексты про заказы, доставку и оплату
- убираем рекламу, спам и нерелевантные сообщения

Кто собирает и чистит

- ML-инженеры делают автоматические фильтры
- обычные люди на краудсорсинг-платформах проверяют и отбрасывают мусор

В итоге получаем:

- большие объёмы
- достаточно хорошее качество

Данные для обучения

Разметка — добавление специальных меток к тексту:

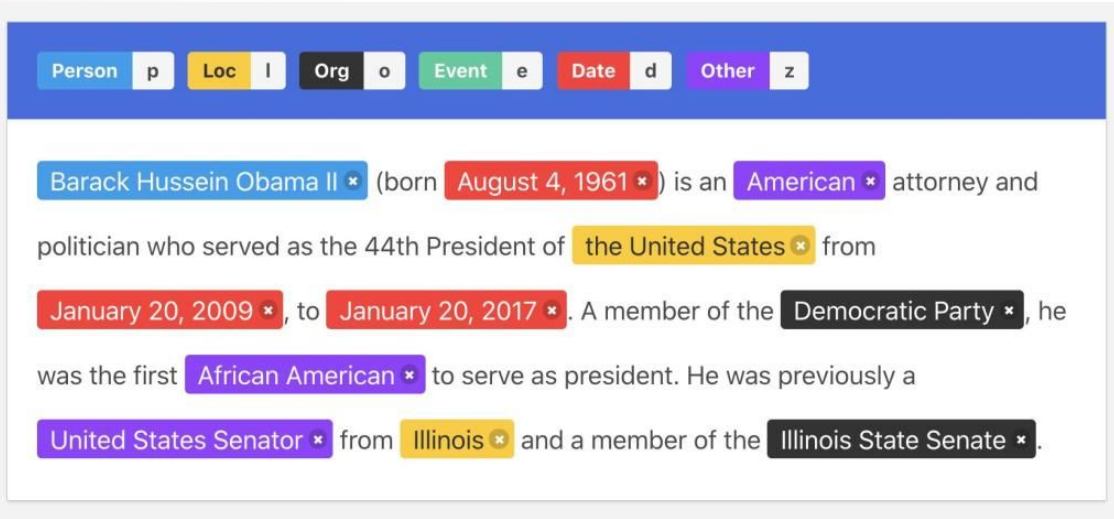
- это негативный комментарий или позитивный
- вопрос про доставку или про оплату
- есть ли тут имя, дата, адрес и т. д.

На таких данных можно обучить модель предсказывать нужные нам результаты

Пример: «Курьер опоздал, я недоволен»

Разметка:

- тип сообщения → **жалоба**
- эмоция → **негатив**
- тема → **доставка**



Сбор и разметка данных

Краудсорсинг

Как составлять задания?

- Разбить сложное задание на простые.

  Например, задача: разметить на картинке дорожные знаки.

  Тогда делаем три последовательных задачи:

  1. Есть ли на изображении дорожные знаки?
  2. Разметьте все знаки на изображении.
  3. Правильно ли на изображении размечены знаки?

- Написать подробную инструкцию, составить примеры.
- Дать сначала обучающие задания.
- Выставить разумную цену.

Агрегация результатов

Зачем нужно делать пост-обработку?

Разметчики — различные типы людей, не все всегда качественно выполняют задания, могут быть боты.

Принцип выдачи заданий.

Обычно выдаются с перекрытием. Например, каждый объект должны разметить не менее 3 человек.

Простой метод: голос большинства.

Для каждого объекта взять самый популярный ответ.

Недостатки

1. Не учитывает способности пользователей.
2. Не учитывает сложность объекта.

Метод Дэвида-Скина (для справки)

n_{ik}^u — кол-во раз, при кот. разметчик u поставил класс k объекту i.

$Y_{ik} = I\{\text{объект } i \text{ класса } k\}$.

$\pi_{k\ell}^u$ — вер-ть того, что разметчик u поставил класс ℓ вместо правильного класса k.

ρ_k — вероятность класса k.

Задача оценки параметров имеет вид

$$\prod_{i\in I}\prod_{k\in K}\left(\rho_k \prod_{u\in U}\prod_{\ell\in K}(\pi_{k\ell}^u)^{n_{i\ell}^u}\right)^{Y_{ik}} \longrightarrow \max_{\pi,\rho}$$

Преимущество:

учитывает то, что разные люди размечают лучше разные классы.

Недостаток: не учитывает сложности объектов.

Метод Дэвида-Скина (для справки)

n_{ik}^u — кол-во раз, при кот. разметчик u поставил класс k объекту i.

$\pi_{k\ell}^u$ — вер-ть того, что разметчик u поставил класс ℓ вместо правильного класса k.

ρ_k — вероятность класса k.

Повторять до сходимости следующие шаги:

$$\gamma_{ik} = \frac{p_k \prod\limits_{u \in U} \prod\limits_{\ell \in K} \left(\pi_{k\ell}^u\right)^{n_{i\ell}^u}}{\sum\limits_{t \in K} p_t \prod\limits_{u \in U} \prod\limits_{\ell \in K} \left(\pi_{t\ell}^u\right)^{n_{i\ell}^u}};$$

$$p_k = \frac{1}{|I|} \sum_{i \in I} \gamma_{ik}; \qquad \pi_{k\ell}^u = \frac{\sum\limits_{i \in I} \gamma_{ik} n_{i\ell}^u}{\sum\limits_{t \in K} \sum\limits_{i \in I} \gamma_{ik} n_{it}^u};$$

Примеры прикладных задач

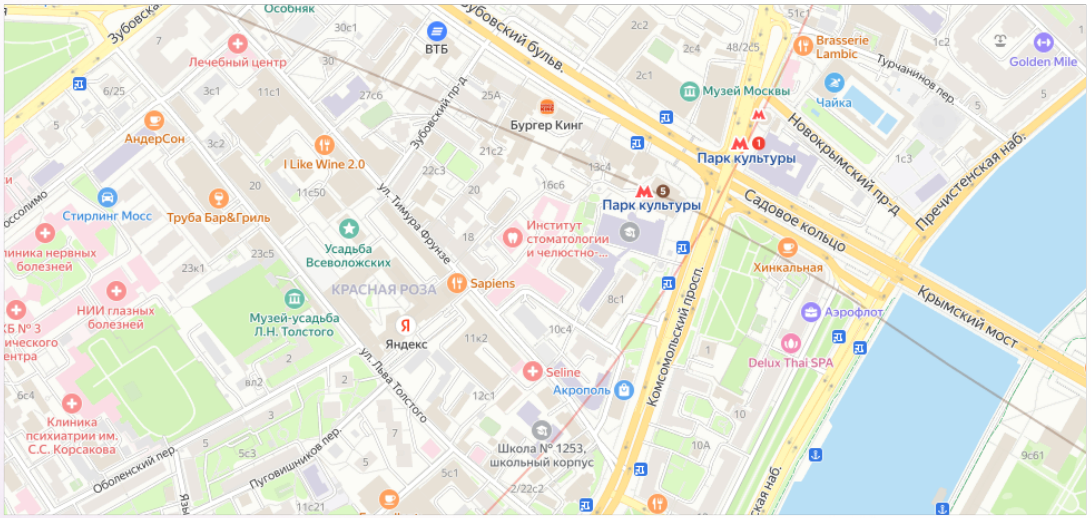
АБ-тестирование

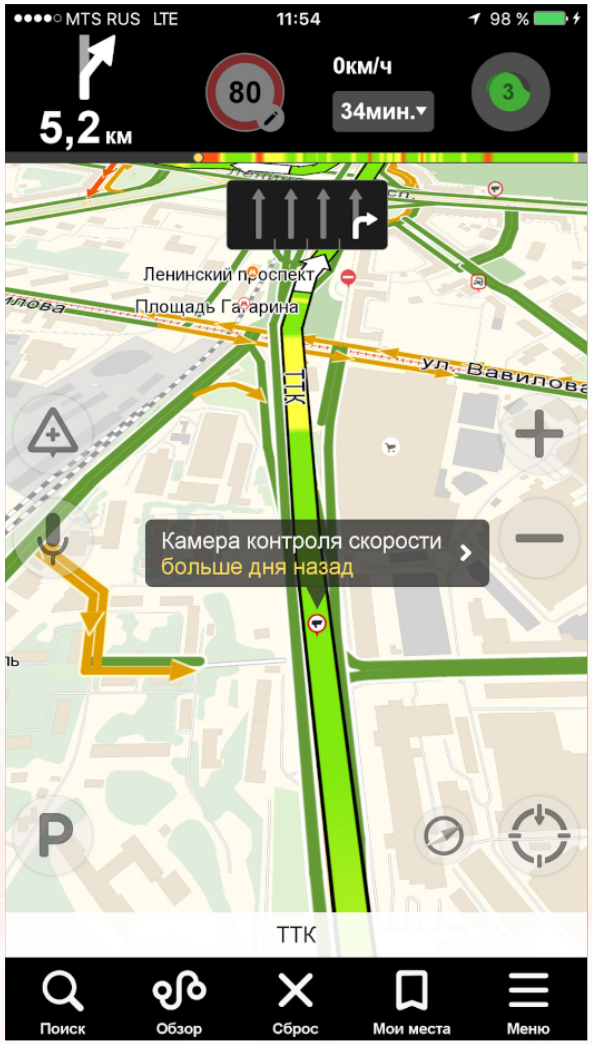
Распознавание лиц

Генерация изображений

В этом разделе только примеры,
понимать все формулы не требуется.

Какой стиль карт удобнее пользователям?





Какая концепция магазина приносит больше выручки?

Контрольная группа
Группа A

Пользователи видят
прежнюю версию сервиса

Тестовая группа
Группа B

Пользователи видят
новую версию сервиса

Классический способ проверки

Пусть $X_1, ..., X_n$ и $Y_1, ..., Y_m$ — значения целевой метрики (выручка, клики, рейтинг и т.д.) для контрольной и тестовой групп.

Постановка задачи:

$\mathrm{H}_0 : \mathrm{E}X_1 = \mathrm{E}Y_1$ *vs.* $\mathrm{H}_1 : \mathrm{E}X_1 \ \{<, \neq, >\} \ \mathrm{E}Y_1$

Справедлива сходимость

$$T(X, Y) = \frac{\overline{X} - \overline{Y}}{\sqrt{S_X^2 / n + S_Y^2 / m}} \xrightarrow{d_0} \mathcal{N}(0, 1).$$

Статистический критерий $S = \{|T(X, Y)| > z_{1-\alpha/2}\}$

Доверительный интервал для $\mathrm{E}X_1 - \mathrm{E}Y_1$ уровня доверия $1 - \alpha$

$$\left(\overline{X} - \overline{Y} \pm z_{1-\alpha/2}\sqrt{S_X^2 / n + S_Y^2 / m}\right).$$

При наличии дополнительных данных

Введем обозначения

$$Y = \begin{pmatrix} Y_{11} \\ \dots \\ Y_{n1} \\ Y_{12} \\ \dots \\ Y_{m2} \end{pmatrix}, \quad X = \begin{pmatrix} 1 & 0 \\ \dots & \\ 1 & 0 \\ 1 & 1 \\ \dots & \\ 1 & 1 \end{pmatrix}, \quad \theta = \begin{pmatrix} \theta_0 \\ \theta_1 \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_{n+m} \end{pmatrix}.$$

Линейная регрессия предполагает зависимость $Y = X\theta + \varepsilon$.

Тогда $Y_{i1} = \theta_0 + \varepsilon_i$ и $Y_{i2} = \theta_0 + \theta_1 + \varepsilon_{n+i}$, следовательно

- θ_0 — среднее в группе A,
- θ_1 — эффект от эксперимента.

Вывод: для проверки AB-теста нужно построить интервал для θ_1 и проверить гипотезу $H_0 : \theta_1 = 0$ критерием Стьюдента.

Важно использовать оценку дисперсии, устойчивую к гетероскедастичности, т.к. группы могут иметь разную дисперсию.

Примеры прикладных задач

АБ-тестирование

Распознавание лиц

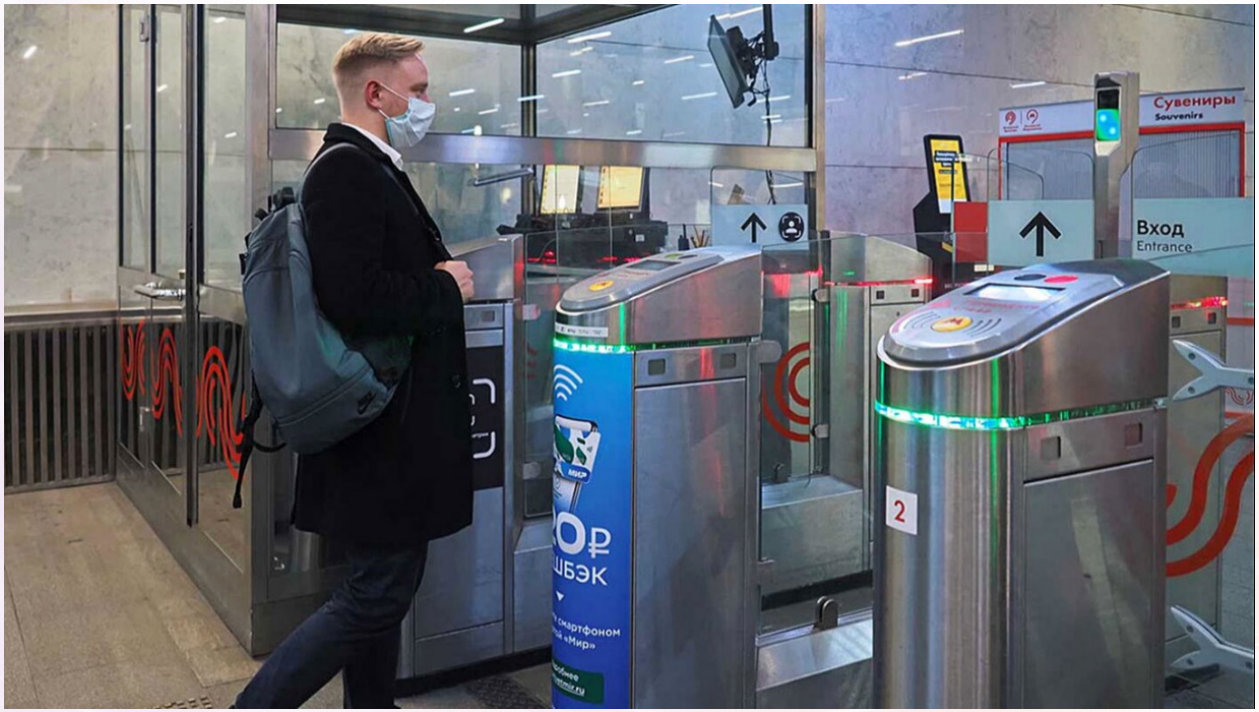
Генерация изображений

В этом разделе только примеры,
понимать все формулы не требуется.

Распознавание лиц

Фото человека → Модель детекции лиц → Координаты лица→

Обработка фото → Сиамская сеть → Идентификатор человека



Распознавание лиц

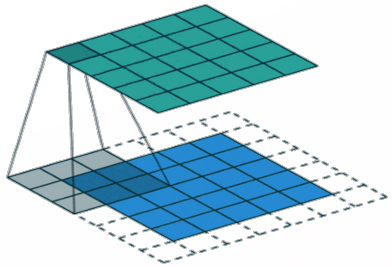
Пример **модели детекции лиц**: TinaFace.

- Основу модели составляют сверточные слои.

  Формула свертки для изображения X

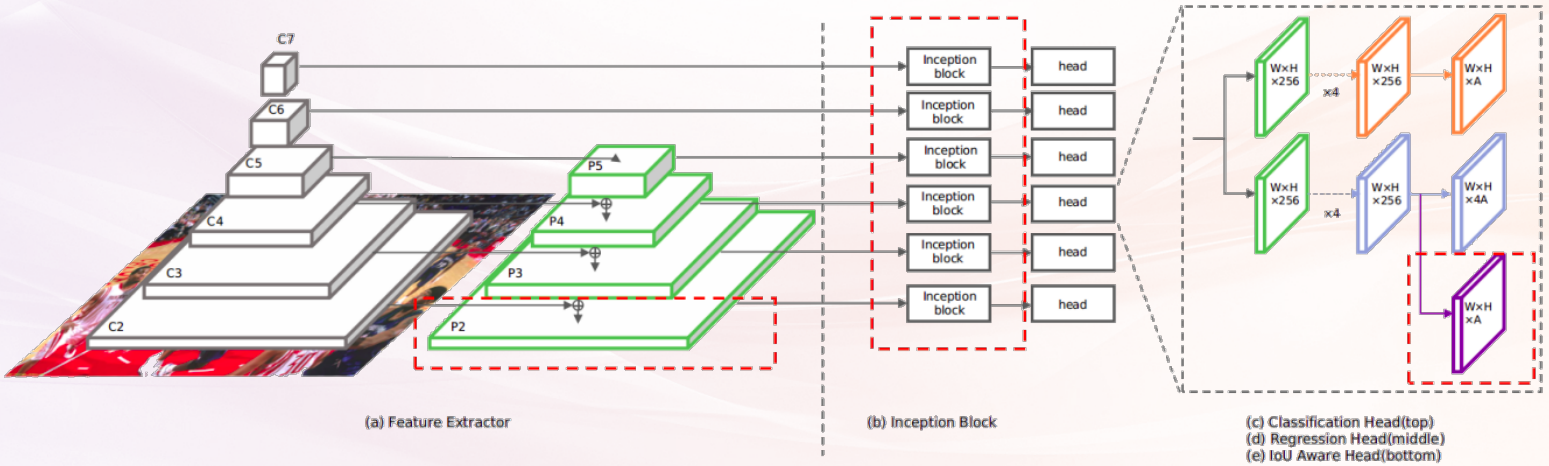
  и фильтра с весами W и сдвигом b:

$$\sum_{i=1}^{M}\sum_{j=1}^{M} w_{ij} \cdot x_{m+i-1,n+j-1} + b$$



- Архитектура модели содержит множество блоков

  из сверточных слоев, функций активаций и т.д.



Распознавание лиц

- **Сиамская сеть** для двух объектов X_1 и X_2 определяет, принадлежат ли они одному классу, оценивая близость между ними.
- Архитектура модели представляет собой сверточную сеть.
- Для оптимизации параметров модели минимизируется **contrastive loss**.

Пусть Y_1 и Y_2 — классы объектов X_1 и X_2 соотв.,
d — функция расстояния,
L_{sim} и L_{dissim} — функции штрафующие за близость объектов одного класса и дальность объектов разных классов соотв.
Тогда лосс равен:

$$L(X_1, X_2, Y_1, Y_2) = I\{Y_1 = Y_2\} L_{sim}\Big(d\left(f(X_1), f(X_2)\right)\Big)$$

$$+ I\{Y_1 \neq Y_2\} L_{dissim}\Big(d\left(f(X_1), f(X_2)\right)\Big)$$

Примеры прикладных задач

АБ-тестирование

Распознавание лиц

Генерация изображений

В этом разделе только примеры,
понимать все формулы не требуется.

Генерация изображений

Задача — научиться генерировать разнообразные правдоподобные изображения, например котов.

Генерация изображений

Для этого построим **диффузионную модель**.

Она моделирует 2 процесса:

- Прямой процесс — постепенно добавляем шум ко входу.
- Обратный процесс — модель постепенно восстанавливает данные из шума.

Генерация изображений

Прямой диффузионный процесс

$$X_t = \sqrt{1-\beta_t} X_{t-1} + \sqrt{\beta_t}\varepsilon, \text{ где } \varepsilon \sim \mathcal{N}(0, \mathrm{I})$$

$$X_t \mid X_{t-1} \sim \mathcal{N}(\sqrt{1-\beta_t} X_{t-1}, \beta_t \mathrm{I})$$

Генерация изображений

Обратный диффузионный процесс

Обучается нейросетевая модель таким образом,

чтобы минимизировать *ELBO*:

$$ELBO = \mathsf{E}_q \log \frac{p_\theta(X_0, \ldots, X_T)}{q(X_1, \ldots, X_T \mid X_0)}$$

$$\simeq const - \sum_{t=2}^{T} \frac{\widetilde{\alpha}_{t-1}\beta_t^2}{2\widetilde{\beta}_t(1-\alpha_t)^2} \mathsf{E}_q \|X_0 - x_\theta(X_t, t)\|^2$$


ВСЁ!