

Введение в Ад



О курсе

Команда Физтех.Статистики





Проводимые нами учебные курсы

- ▶ Введение в анализ данных;
- ▶ DS-поток;
- ▶ Phystech@DataScience (в т.ч. мат. статистика на ЛФИ);
- ▶ Мат. статистика на ФБМФ и ФМХФ (семинары);
- ▶ Два курса в магистратуре кафедры X5 Group;
- ▶ Машинное обучение в ШАД (частично);
- ▶ АБ-тестирование в ШАД.

Организационная информация

Телеграм-бот

@miptstats_ds23_bot



Код регистрации **Pandas@2000!**

Сайт команды miptstats.github.io

Почта mipt.stats@yandex.ru

Цели курса

1. Дать представление об анализе данных;
2. Обучить базовым инструментам анализа данных;
3. Рассказать о практическом смысле объектов теории вероятностей;
4. Помочь определиться с кафедрой.

Яндекс

В курсе предполагается участие кафедры анализа данных:

- ▶ Гостевые лекции от Яндекса;
- ▶ Работа по курсу учитывается при отборе на кафедру.

План занятий первой части (обязательная)

Дата	Тема	Лектор	Дедлайн по ДЗ
4.02	<i>Введение, инструменты</i>		Никита Волков 10.02
11.02	<i>Сбор данных. Модели данных</i>		Роман Логинов 17.02
18.02	<i>Самостоятельная работа или гостевая лекция</i>		24.02
25.02	<i>Самостоятельная работа или гостевая лекция</i>		3.03
4.03	<i>Теория вероятностей на практике</i>		Екатерина Юшина 10.03
11.03	<i>Линейная регрессия</i>		Никита Волков 18.03

План занятий второй части (факультативная)

Дата	Тема	Лектор	Дедлайн по ДЗ
18.03	Самостоятельная работа или гостевая лекция		
25.03	Решающие деревья, случайные леса		Никита Волков 31.03
01.04	Основы статистики. Свойства оценок.		Никита Волков 5.04
18.03	Самостоятельная работа или гостевая лекция		
15.04	Байесовские классификаторы		Никита Волков 21.04
22.04	Введение в нейросети		Елизавета Дахова 28.04

Система оценивания первой части

Правила:

1. если $L > 25\%$ и ($B < 50\%$ или $C < 20\%$),
то $O = 3+4*L$;
2. если $L > 25\%$ и $B \geq 50\%$ и $C \geq 20\%$,
то $O = \max(3+4*L, 3+B+7*LC)$.
3. если $L \leq 25\%$,
то $O = 1+3*T$.

Обозначения:

- ▶ L — доля выполнения легких заданий (их немного);
- ▶ C — доля выполнения сложных заданий (их много);
- ▶ T — доля выполнения тестов;
- ▶ LC — доля выполнения всех заданий (кроме тестов);
- ▶ B — доля правильных ответов на вопросы в боте на занятии;
- ▶ O — итоговая оценка, округляется вверх.

Система оценивания первой части

Правила:

1. если $L > 25\%$ и ($B < 50\%$ или $C < 20\%$),
то $O = 3+4*L$;
2. если $L > 25\%$ и $B \geq 50\%$ и $C \geq 20\%$,
то $O = \max(3+4*L, 3+B+7*LC)$.
3. если $L \leq 25\%$,
то $O = 1+3*T$.

Следствия: для получения оценки

- ▶ уд(3) достаточно выполнить 33.4% тестов.
- ▶ хор(5) достаточно выполнить 25.1% легких заданий.
- ▶ отл(8) достаточно выполнить 50.1% всех заданий
и ответить на 50% вопросов.

Если списать: все участники списывания сдают устный зачет.



Для кого наш курс?

Первая часть курса:

- ▶ ПМИ, ПМФ КТ — курс обязательен.
- ▶ ИВТ, ПМФ МФ, иностр. группы — курс факультативен.
- ▶ Студенты других физтех-школ также могут сдавать курс.

Вторая часть курса:

- ▶ Факультативно для всех.

В силу ограниченных возможностей проверяющих
при большом количестве желающих возможность сдавать курс
может быть ограничена.



Правила комфорта

- ▶ Постарайтесь задавать вопросы на занятии в тот момент, когда это актуально, не перебивая на полуслове.

Другой вопрос лучше задать в перерыве или после занятия.

- ▶ Цените труд проверяющих :)

В каком из случаев проверяющему больше захочется пойти навстречу автору вопроса?

- ▶ "Объясните вашу претензию, почему вы мне сняли баллы, я же все сделал, я не согласен"
- ▶ "Добрый день! По такой-то задаче вы написали ..., но я считаю ..., потому что ..., и у меня в работе написано ..."

DS-поток



DS-поток

Семестр	DS-поток	Основной поток ПМИ
5	Математическая статистика	Математическая статистика
	Машинное обучение	Машинное обучение
	Практика	Практика по мат. статистике
	Основы прикладной статистики	Курс по выбору x 2
	Курс по выбору	
6	Дискр. случ. процессы и временные ряды	Случайные процессы
	Прикладная статистика и анализ данных	Курс по выбору x 3
	Практика	
	Курс по выбору	
Кафедра АД	Курс ШАД	Методы прикладной статистики
7	Байесовский подход в анализе данных	Курс по выбору x 2
	Практика	
8	Прикладные задачи машинного обучения	Курс по выбору x 2
	Практика	

Примечания.

1. По одним и тем же курсам лекторы разные.
2. По факту практика не является отдельным курсом.
3. Полужирным выделены курсы, по которым экзамен.
4. В процессе обучения перейти в DS-поток невозможно.



DS-поток

Чему мы учим

1. В меру глубокое математическое понимание статистики и машинного обучения.
2. Применение математических моделей на реальных данных, в том числе на реальных задачах.
3. Умение составлять полноценные выводы.

Реальная практика на DS-потоке

1. Реальные примеры из практики;
2. Соревнования на Kaggle;
3. Гостевые лекторы, применяющие анализ данных на практике;
4. Разбор статей на тему анализа данных;
5. Бонусы за участие в хакатонах, соревнованиях по анализу данных и прочую активность;

Отбор на DS-поток

Что будет учитываться?

1. Необходимое условие:

оценка не менее отл(8) по первой части курса

"Введение в анализ данных" и не менее хор(7) по второй части.

2. Работа в семестре по курсу, грамотное оформление ДЗ.

3. Оценка по теории вер., в меньшей степени — другие предметы.

Что нужно делать для отбора?

1. Трудиться в течение семестра.

2. В июне подать заявку.

3. Ждать. Результаты летом.

DS-поток адаптирован для ПМИ. Студенты других напр. могут попасть на DS-поток, но возможна доп. нагрузка и проблемы с расписанием.

Куда пойти?

2 курс

Введение в анализ данных

3-4 курсы

DS-поток

Кафедра
анализа данных

Школа анализа
данных (ШАД)

1. Если анализ данных интересен, то хорошее решение:
DS-поток + кафедра анализа данных.
2. Если выбираете кафедру анализа данных,
то кафедра рекомендует пойти на DS-поток.
3. По результатам нашего курса кафедра анализа данных
может зачесть тех. собеседование при отборе на кафедру.

Что такое анализ данных?



Кому нужен анализ данных

1. "Я математик, практическое применение не интересует".

Скорее всего АД не нужен,
но часто математики им начинают интересоваться.

2. "Я математик, но хочу применять свои знания на практике".

АД для вас, ждем вас на DS-потоке

3. "Я программист, и хочу писать только код".

Скорее всего АД в подробностях не нужен,
но стоит понимать, чем занимаются коллеги-аналитики.

4. "Я программист, но хочу глубоко разбираться
в тонкостях математических методов".

АД для вас, ждем вас на DS-потоке



Так что, анализ данных

это математика

или программирование?

Давайте разбираться...

Посмотрим на лекции по статистике

Статистика, прикладной поток 12. Контроль FWER и FDR. Критерий согласия

2) Метод Крамера.
Несколько наблюдений с $\alpha_{ij} = \frac{d}{m+n+8}$.
Тогда: Если $\forall j: P(X_j) \sim U[0, 1]$,
то $FWER(F) \leq \alpha$ \Rightarrow D.

D. $FWER(F) = P$

$\leq \sum_{j \in J_F} P(X_j \in \dots)$

Следует учесть, что

$P_0 = p\text{-значие}$

$P_1 = \text{критическое}$

Статистика, прикладной поток 11. p-значие. Практическая значимость. Множественный проверка гипотез

Критерии (напоминание)

Часто критерий имеет вид $S = \{T(x) \geq c_\alpha\}$, где $T(X)$ — статистика критерия.

α выбирается **до** эксперимента,

c_α вычисляется из условия $P_0(T(X) > c_\alpha) \leq \alpha$.

$S = \{T(x) > c_\alpha\}$ $S = \{T(x) < c_\alpha\}$ $S = \{|T(x)| > c_\alpha\}$

Статистика, прикладной поток 9. Бутстреп. Формы оценки плотности. Проверка статистических гипотез

Метод бутстрепа

Этап 2.
Процедуре генерации выборок повторить B раз:

$X_b^* = (X_{b1}^*, \dots, X_{bn}^*)$, где $1 \leq b \leq B$.

Далее по каждой выборке посчитаем значение статистики T , получив выборку значений $T_1^* = T(X_1^*), \dots, T_B^* = T(X_B^*)$.

Этап 3.
Полученную выборку использовать для аппроксимации значения оценки, которая называется **бутстрепной оценкой**.

Например, бутстрепная оценка дисперсии имеет вид

$$\hat{\sigma}_{boot}^2 = \frac{1}{B} \sum_{b=1}^B T_b^{*2} - \left(\frac{1}{B} \sum_{b=1}^B T_b^* \right)^2$$

Статистика, прикладной поток 15. Оптимальные оценки. Экспериментальные оценки. Доказательства первым способом

Тогда $\hat{F}_n(x) - F(x) \leq \hat{F}_n(U_{N+1}) - F(U_N) =$

$$= F_n\left(\frac{U_{N+1}}{N+1}\right) - F\left(\frac{U_N}{N}\right) + F\left(\frac{U_{N+1}}{N+1}\right) - F\left(\frac{U_N}{N}\right) \leq \frac{1}{N+1} + \frac{1}{N}$$

распишем P_0 в виде F .

(x) $\xrightarrow{\text{Р-знач.}}$ \circ

$\hat{F}_n(x) \geq \hat{F}_n(U_N) - \frac{1}{N}$

Аналогично

также $\Rightarrow \hat{F}_n(x) - F(x) \geq \hat{F}_n(U_N) - F(U_N) - \frac{1}{N}$

Пусть x — квантильный

$|F_n(x) - F(x)|$

Посмотрим на научные статьи

McInnes, L, Healy, J, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, ArXiv e-prints 1802.03426, 2018

The choice of non-expansive maps in Definition 6 is due to Spivak [33]. We note that it closely mirrors the work of Carlsson and Memoli in [14] for logical methods for clustering as applied to finite metric spaces. This is significant since pure isometries are too strict and do not provide local Hom-sets.

In [33] Spivak constructs a pair of adjoint functors, **Real** and **Sin**, from the categories **sFuzz** and **EPMet**. These functors are the natural extensions of the classical realization and singular set functors from algebraic topology. The functor **Real** is defined in terms of standard fuzzy simplices $\Delta_{\leq a}^n$ as

$$\text{Real}(\Delta_{\leq a}^n) \triangleq \left\{ (t_0, \dots, t_n) \in \mathbb{R}^{n+1} \mid \sum_{i=0}^n t_i = -\log(a), t_i \geq 0 \right\}$$

similarly to the classical realization functor $|\cdot|$. The metric on $\text{Real}(\Delta_{\leq a}^n)$ is simply inherited from \mathbb{R}^{n+1} . A morphism $\Delta_{\leq a}^n \rightarrow \Delta_{\leq b}^m$ exists only if $a \leq b$ and it is determined by a Δ morphism $\sigma : [n] \rightarrow [m]$. The action of **Real** on a morphism is given by the map

$$(x_0, x_1, \dots, x_n) \mapsto \frac{\log(b)}{\log(a)} \left(\sum_{i_0 \in \sigma^{-1}(0)} x_{i_0}, \sum_{i_0 \in \sigma^{-1}(1)} x_{i_0}, \dots, \sum_{i_0 \in \sigma^{-1}(m)} x_{i_0} \right).$$

Such a map is clearly non-expansive since $0 \leq a \leq b \leq 1$ implies that $\log(b)/\log(a) \leq 1$.

We then extend this to a general simplicial set X via colimits, defining

$$\text{Real}(X) \triangleq \text{colim}_{\Delta_{\leq a}^n \rightarrow X} \text{Real}(\Delta_{\leq a}^n).$$

Since the functor **Real** preserves colimits, it follows that there exists a right adjoint functor. Again, analogously to the classical case, we find the right adjoint denoted **Sing**, is defined for an extended pseudo metric space Y in terms of its action on the category $\Delta \times I$:

$$\text{Sing}(Y) : ([n], [0, a)) \rightsquigarrow \text{hom}_{\text{EPMet}}(\text{Real}(\Delta_{\leq a}^n), Y).$$

For our case we are only interested in finite metric spaces. To correspond with this we consider the subcategory of bounded fuzzy simplicial sets **Fin-sFuzz**. We therefore use the analogous adjoint pair **FinReal** and **FinSing**. Formally we define the finite fuzzy realization functor as follows:

Algorithm 2 Constructing a local fuzzy simplicial set

```
function LOCALFUZZYSIMPLICIALSET( $X, x, n$ )
    knn, knn-dists  $\leftarrow$  APPROXNEARESTNEIGHBORS( $X, x, n$ )
     $\rho \leftarrow$  knn-dists[1]                                 $\triangleright$  Distance to nearest neighbor
     $\sigma \leftarrow$  SMOOTHKNNDIST(knn-dists,  $n, \rho$ )           $\triangleright$  Smooth approximator to knn-distance
    fs-set0  $\leftarrow X$ 
    fs-set1  $\leftarrow \{([x, y], 0) \mid y \in X\}$ 
    for all  $y \in \text{knn}$  do
         $d_{x,y} \leftarrow \max\{0, \text{dist}(x, y) - \rho\}/\sigma$ 
        fs-set1  $\leftarrow$  fs-set1  $\cup$   $\{([x, y], \exp(-d_{x,y}))\}$ 
    return fs-set
```

Algorithm 3 Compute the normalizing factor for distances σ

```
function SMOOTHKNNDIST(knn-dists,  $n, \rho$ )
    Binary search for  $\sigma$  such that  $\sum_{i=1}^n \exp(-(\text{knn-dists}_i - \rho)/\sigma) = \log_2(n)$ 
    return  $\sigma$ 
```



Figure 8: Visualization of the full 3 million word vectors from the GoogleNews dataset as embedded by UMAP.

is contained in U , then g is constant in B and hence $\sqrt{\det(g)}$ is constant can be brought outside the integral. Thus, the volume of B is

$$\sqrt{\det(g)} \int_B dx^1 \wedge \dots \wedge dx^n = \sqrt{\det(g)} \frac{\pi^{n/2} r^n}{\Gamma(n/2 + 1)},$$

where r is the radius of the ball in the ambient \mathbb{R}^n . If we fix the volume of the ball to be $\frac{\pi^{n/2}}{\Gamma(n/2 + 1)}$ we arrive at the requirement that

$$\det(g) = \frac{1}{r^{2n}}.$$

Since g is assumed to be diagonal with constant entries we can solve for g as

$$g_{ij} = \begin{cases} \frac{1}{r^2} & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

geodesic distance on M under g from p to q (where $p, q \in B$) is defined as

$$\inf_{c \in C} \int_a^b \sqrt{\langle \dot{c}(t), \dot{c}(t) \rangle} dt,$$

where C is the class of smooth curves c on M such that $c(a) = p$ and $c(b) = q$, and \dot{c} denotes the first derivative of c on M . Given that g is as defined in (2) we see that this can be simplified to

$$\begin{aligned} & \frac{1}{r} \inf_{c \in C} \int_a^b \langle \sqrt{\dot{c}(t)}, \dot{c}(t) \rangle dt \\ &= \frac{1}{r} \inf_{c \in C} \int_a^b \langle \|\dot{c}(t)\|, \dot{c}(t) \rangle dt \\ &= \frac{1}{r} d_{\mathbb{R}^n}(p, q). \end{aligned} \quad (3)$$

□

B Proof that FinReal and FinSing are adjoint

Theorem 2. The functors $\text{FinReal} : \text{Fin-sFuzz} \rightarrow \text{FinEPMet}$ and $\text{FinSing} : \text{FinEPMet} \rightarrow \text{Fin-sFuzz}$ form an adjunction with FinReal the left adjoint and FinSing the right adjoint.

Посмотрим на научные статьи

Diederik P Kingma, Max Welling: Auto-Encoding Variational Bayes, ArXiv 1312.6114, 2014

2.2 The variational bound

The marginal likelihood is composed of a sum over the marginal likelihoods of individual datapoints $\log p_{\theta}(\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(N)}) = \sum_{i=1}^N \log p_{\theta}(\mathbf{x}^{(i)})$, which can each be rewritten as:

$$\log p_{\theta}(\mathbf{x}^{(i)}) = D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\theta}(\mathbf{z}|\mathbf{x}^{(i)})) + \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) \quad (1)$$

The first RHS term is the KL divergence of the approximate from the true posterior. Since this KL-divergence is non-negative, the second RHS term $\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$ is called the (variational) *lower bound* on the marginal likelihood of datapoint i , and can be written as:

$$\log p_{\theta}(\mathbf{x}^{(i)}) \geq \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) = \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}[-\log q_{\phi}(\mathbf{z}|\mathbf{x}) + \log p_{\theta}(\mathbf{x}, \mathbf{z})] \quad (2)$$

which can also be written as

$$\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) = -D_{KL}(q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})||p_{\theta}(\mathbf{z})) + \mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x})}$$

We want to differentiate and optimize the lower bound $\mathcal{L}(\theta, \phi; \mathbf{x})$ parameters ϕ and generative parameters θ . However, the gradient is a bit problematic. The usual (naive) Monte Carlo gradient estimator is: $\nabla_{\phi} \mathbb{E}_{q_{\phi}(\mathbf{z})} [f(\mathbf{z})] = \mathbb{E}_{q_{\phi}(\mathbf{z})} [f(\mathbf{z}) \nabla_{q_{\phi}(\mathbf{z})} \log q_{\phi}(\mathbf{z})] \approx \frac{1}{L} \sum_{l=1}^L f(\mathbf{z}^{(l)}) \sim q_{\phi}(\mathbf{z}|\mathbf{x}^{(l)})$. This gradient estimator exhibits very high variance and is impractical for our purposes.

2.3 The SGVB estimator and AEVB algorithm

In this section we introduce a practical estimator of the lower bound parameters. We assume an approximate posterior in the form q_{ϕ} can be applied to the case $q_{\phi}(\mathbf{z})$, i.e. where we do not consider variational Bayesian method for inferring a posterior over the parameters.

Under certain mild conditions outlined in section 2.4 for a chosen approximation we can reparameterize the random variable $\bar{\mathbf{z}} \sim q_{\phi}(\mathbf{z}|\mathbf{x})$ using a difference of an (auxiliary) noise variable ϵ :

$$\bar{\mathbf{z}} = g_{\phi}(\epsilon, \mathbf{x}) \quad \text{with} \quad \epsilon \sim p(\epsilon)$$

See section 2.4 for general strategies for choosing such an appropriate $g_{\phi}(\epsilon, \mathbf{x})$. We can now form Monte Carlo estimates of expectation $\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})}$ as follows:

$$\mathbb{E}_{q_{\phi}(\mathbf{z}|\mathbf{x}^{(i)})} [f(\mathbf{z})] = \mathbb{E}_{p(\epsilon)} \left[f(g_{\phi}(\epsilon, \mathbf{x}^{(i)})) \right] \approx \frac{1}{L} \sum_{l=1}^L f(g_{\phi}(\epsilon^{(l)}, \mathbf{x}^{(i)}))$$

We apply this technique to the variational lower bound (eq. 2), yielding our generic Stochastic Gradient Variational Bayes (SGVB) estimator $\tilde{\mathcal{L}}^A(\theta, \phi; \mathbf{x}^{(i)}) \approx \mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$:

$$\tilde{\mathcal{L}}^A(\theta, \phi; \mathbf{x}^{(i)}) = \frac{1}{L} \sum_{l=1}^L \log p_{\theta}(\mathbf{x}^{(i)}, \mathbf{z}^{(l, l)}) - \log q_{\phi}(\mathbf{z}^{(l, l)}|\mathbf{x}^{(i)})$$

$\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)})$ is the variational lower bound of the marginal likelihood of datapoint i :

$$\mathcal{L}(\theta, \phi; \mathbf{x}^{(i)}) = \int q_{\phi}(\mathbf{z}|\mathbf{x}) (\log p_{\theta}(\mathbf{x}^{(i)}|\mathbf{z}) + \log p_{\theta}(\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})) d\mathbf{z} \quad (16)$$

The expectations on the RHS of eqs (14) and (16) can obviously be written as a sum of three separate expectations, of which the second and third component can sometimes be analytically solved, e.g. when both $p_{\theta}(\mathbf{x})$ and $q_{\phi}(\mathbf{z}|\mathbf{x})$ are Gaussian. For generality we will here assume that each of these expectations is intractable.

Under certain mild conditions outlined in section (see paper) for chosen approximate posteriors parameterize conditional samples $\bar{\mathbf{z}} \sim q_{\phi}(\mathbf{z}|\mathbf{x})$ as

$$\bar{\mathbf{z}} = g_{\phi}(\epsilon, \mathbf{x}) \quad \text{with} \quad \epsilon \sim p(\epsilon) \quad (17)$$

and a function $g_{\phi}(\epsilon, \mathbf{x})$ such that the following holds:

$$\mathbf{z}|\mathbf{x}) (\log p_{\theta}(\mathbf{x}^{(i)}|\mathbf{z}) + \log p_{\theta}(\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x})) d\mathbf{z}$$

$$+ \left. \left(\log p_{\theta}(\mathbf{x}^{(i)}|\mathbf{z}) + \log p_{\theta}(\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x}) \right) \right|_{\mathbf{z}=g_{\phi}(\epsilon, \mathbf{x}^{(i)})} d\epsilon \quad (18)$$

: approximate posterior $q_{\phi}(\theta)$:

$$\theta = h_{\phi}(\zeta) \quad \text{with} \quad \zeta \sim p(\zeta) \quad (19)$$

e, choose a prior $p(\zeta)$ and a function $h_{\phi}(\zeta)$ such that the following

$$q_{\phi}(\theta) (\log p_{\theta}(\mathbf{X}) + \log p_{\alpha}(\theta) - \log q_{\phi}(\theta)) d\theta$$

$$p(\zeta) (\log p_{\theta}(\mathbf{X}) + \log p_{\alpha}(\theta) - \log q_{\phi}(\theta)) \Big|_{\theta=h_{\phi}(\zeta)} d\zeta \quad (20)$$

we introduce a shorthand notation $f_{\phi}(\mathbf{x}, \mathbf{z}, \theta)$:

$$p_{\theta}(\mathbf{x}|\mathbf{z}) + \log p_{\theta}(\mathbf{z}) - \log q_{\phi}(\mathbf{z}|\mathbf{x}) + \log p_{\alpha}(\theta) - \log q_{\phi}(\theta) \quad (21)$$

(18), the Monte Carlo estimate of the variational lower bound, given

$$\mathcal{L}(\phi; \mathbf{X}) = \frac{1}{L} \sum_{l=1}^L f_{\phi}(\mathbf{x}^{(l)}, g_{\phi}(\epsilon^{(l)}, \mathbf{x}^{(l)}), h_{\phi}(\zeta^{(l)})) \quad (22)$$

where $\epsilon^{(l)} \sim p(\epsilon)$ and $\zeta^{(l)} \sim p(\zeta)$. The estimator only depends on samples from $p(\epsilon)$ and $p(\zeta)$ which are obviously not influenced by ϕ , therefore the estimator can be differentiated w.r.t. ϕ .

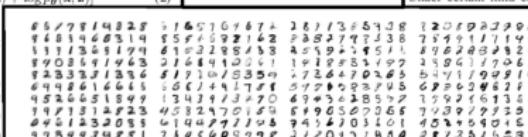
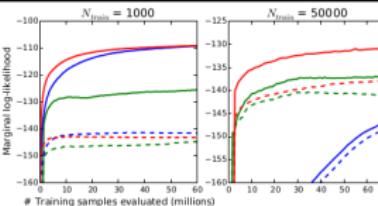


Figure 5: Random samples from learned generative models of MNIST for different dimensionalities of latent space.



Wake-Sleep (train)

Wake-Sleep (test)

MCEM (train)

AEVB (train)

AEVB (test)

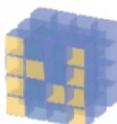


Посмотрим на примеры использования библиотек

Variational Autoencoders

<https://github.com/pyro-ppl/pyro/blob/dev/examples/vae/vae.py>

Какие вообще инструменты могут потребоваться



NumPy

pandas
data analysis



SciPy

matplotlib



PyTorch



Yandex
CatBoost



Numba



plotly



Вывод:
и математика
и программирование

Анализ данных это

процесс поиска закономерностей

в данных при помощи

- ▶ средств визуализации данных,
- ▶ математических методов,
- ▶ программных алгоритмов.



Искусственный интеллект

Отличительная особенность:

нет четко зафиксированного ответа на каждый входящий объект.

Что можно почитать:

Анализ данных — основы и терминология

<https://habr.com/ru/post/352812/>

Всё, что вам нужно знать об ИИ — за несколько минут

<https://habr.com/ru/post/416889/>

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 девушкиам: какой-то – 5, а остальным – по 3 цветка? [ОКТЧ 2019]

Ровно один правильный ответ.

Анализ данных

Задача: Имеются данные $(x_1, y_1), \dots, (x_n, y_n)$.

Восстановите по ним функцию $f : x \mapsto y$.

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

Пример — распознавание рукописных цифр

Вход:

5

Ожидается на выходе: 5

Но как четко алгоритмически определить границу между 6 и 8?

6 8

2 — 2 или 9?

4 — 4 или 7?

Актуальность в научной среде

Число статей по запросам в Google Scholar с 2016:

- ▶ *statistics* \approx 2 010 000 статей
- ▶ *machine learning* \approx 1 360 000 статей
- ▶ *artificial intelligence* \approx 595 000 статей
- ▶ *neural network* \approx 970 000 статей
- ▶ *computer vision* \approx 854 000 статей

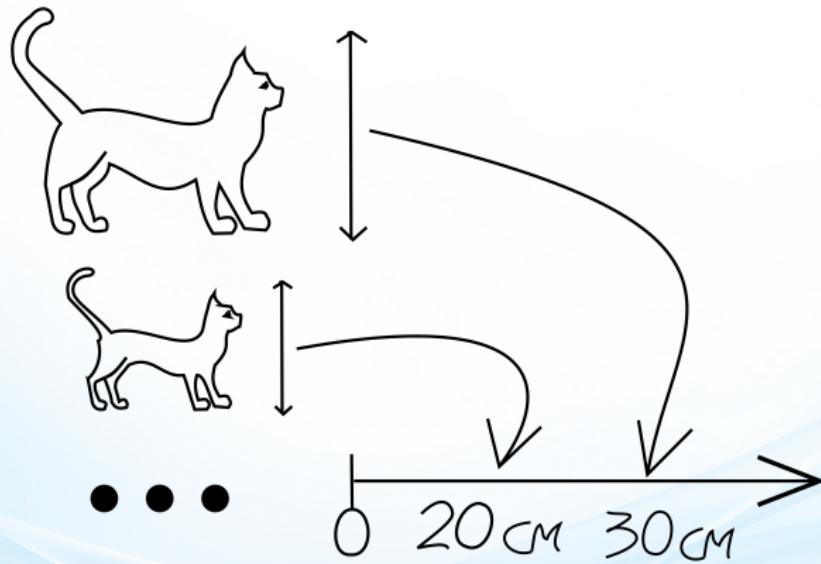
- ▶ *статистика* \approx 60 200 статей
- ▶ *машинное обучение* \approx 15 200 статей
- ▶ *искусственный интеллект* \approx 16 600 статей
- ▶ *нейронные сети* \approx 16 700 статей
- ▶ *компьютерное зрение* \approx 12 300 статей

Обзор задач анализа данных

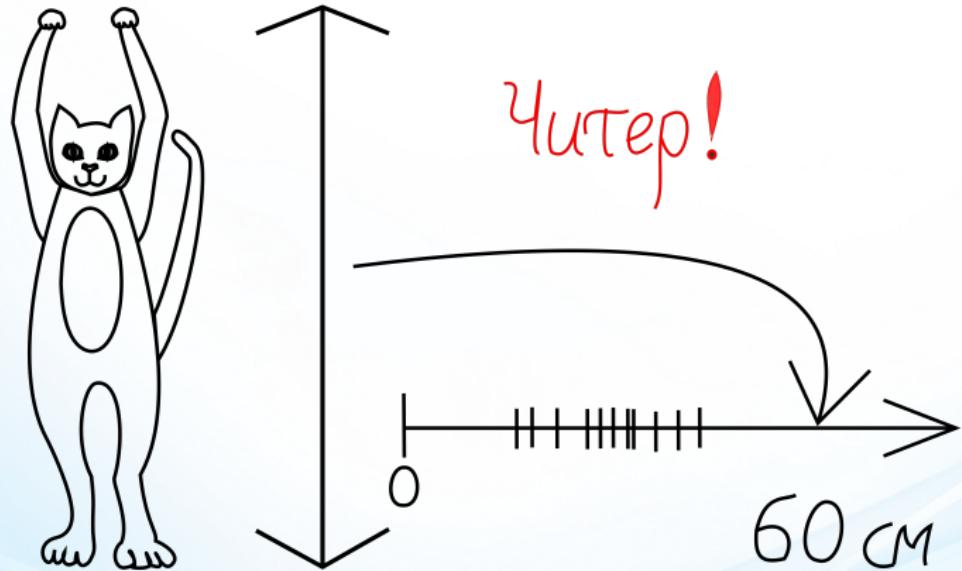
Мурмурландия



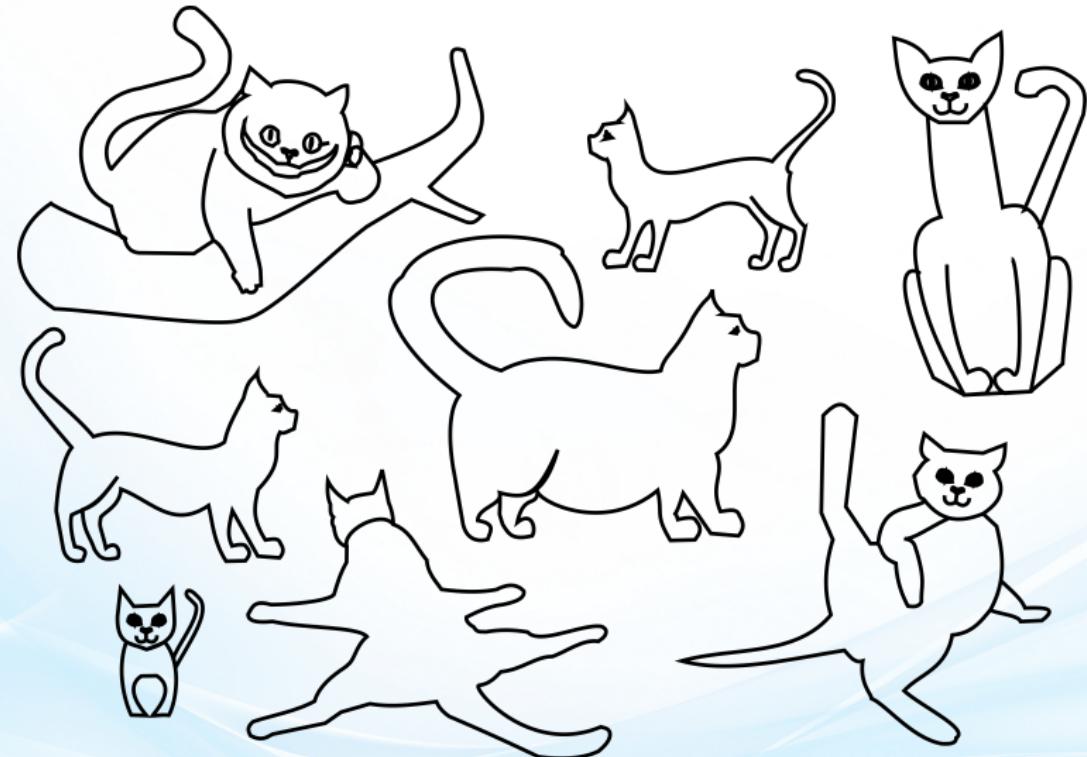
Каков средний рост котиков?



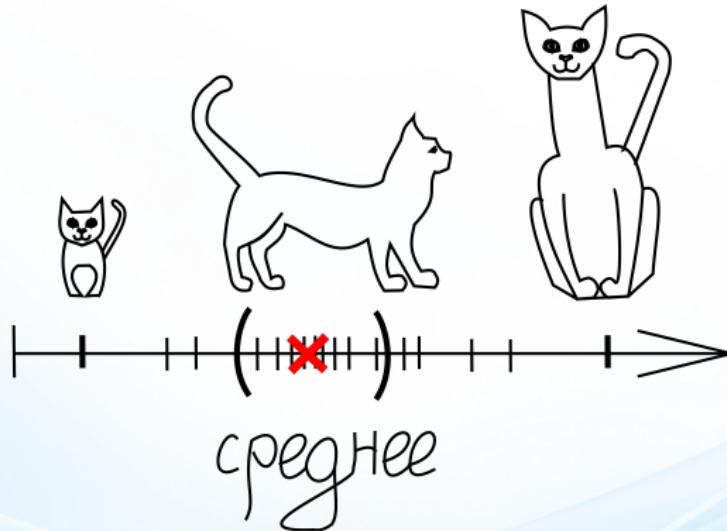
Точечное оценивание



Выбросы

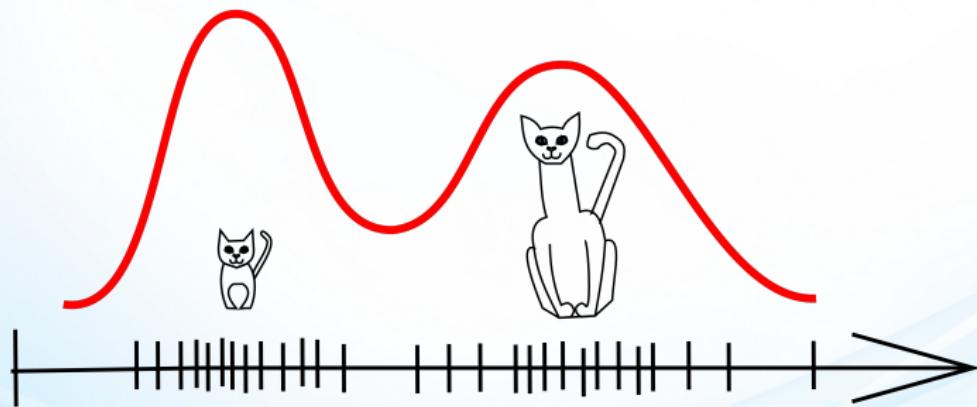


Среднее определяется неточно



Интервальное оценивание

Характер распределения



Непараметрическое оценивание

НИЗКИЕ

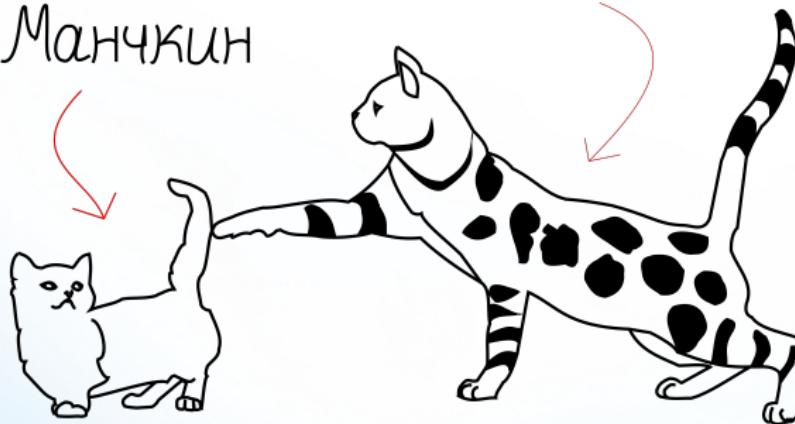


ВЫСОКИЕ



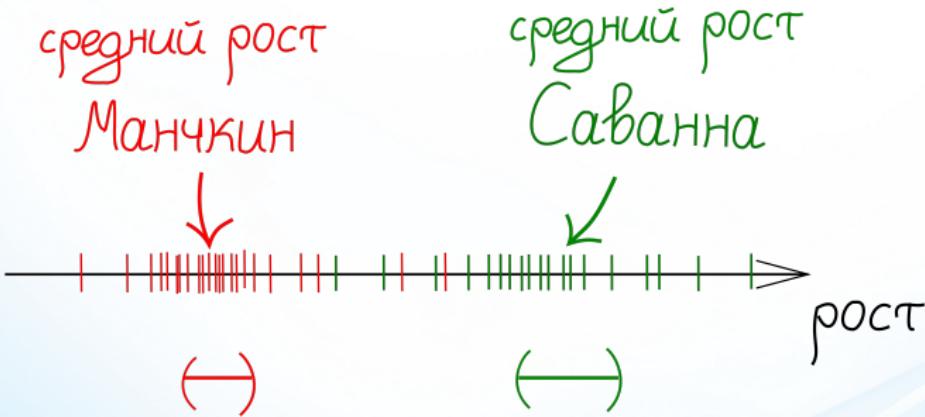
Саванна

Манчкин



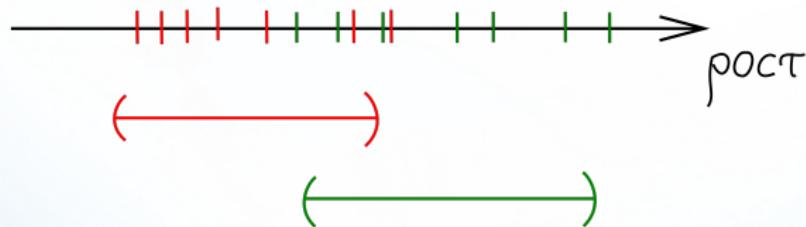
Отличается ли их средний рост?

Собираем данные



отличается

Если данных мало

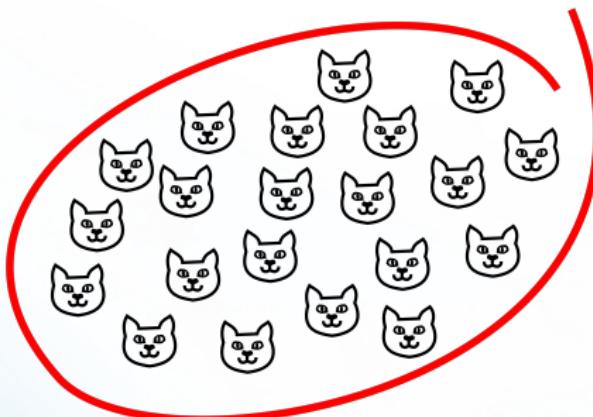


непонятно

Статистические гипотезы, АВ-тесты



счастье

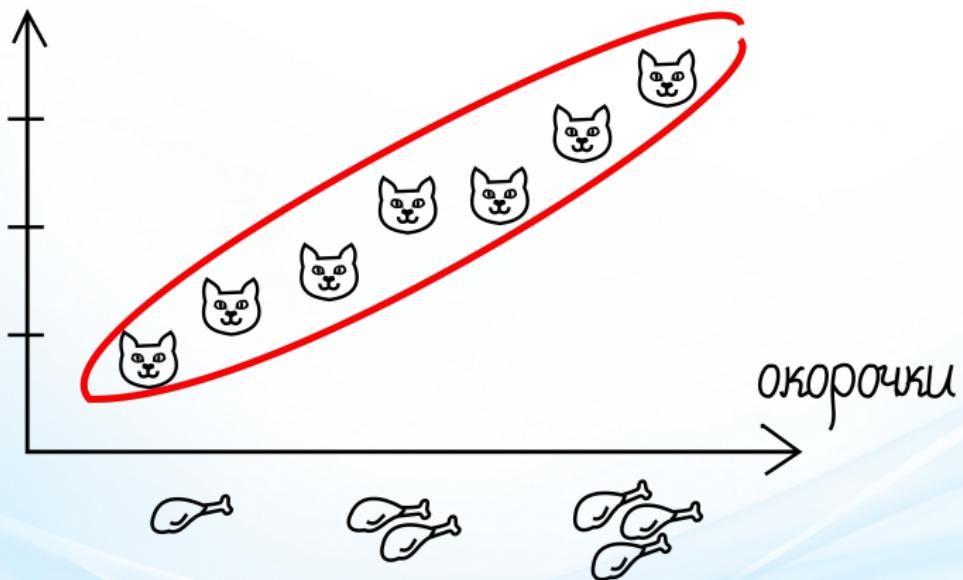


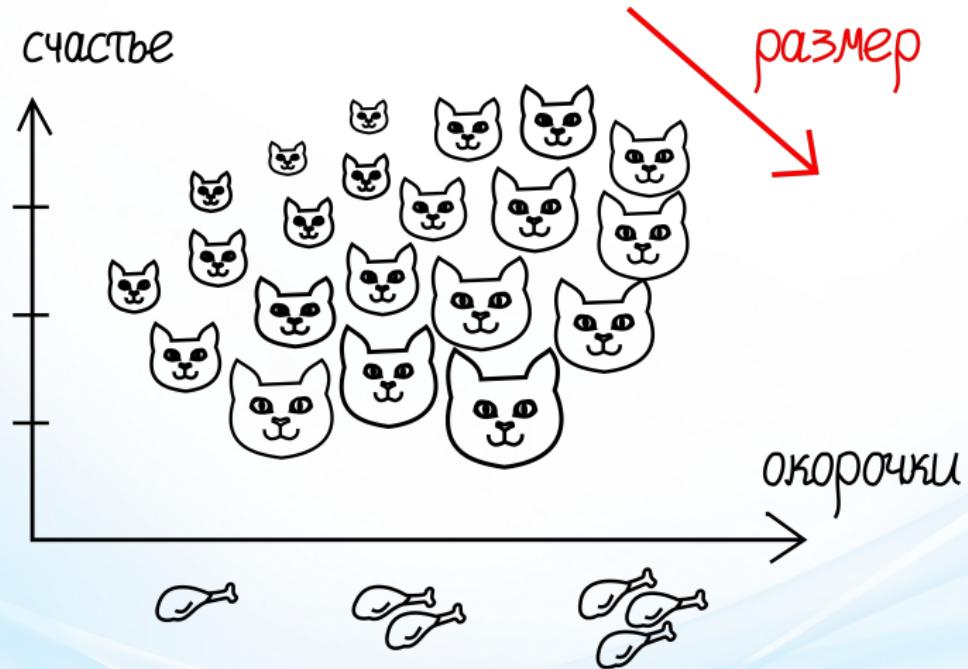
окорочки





счастье





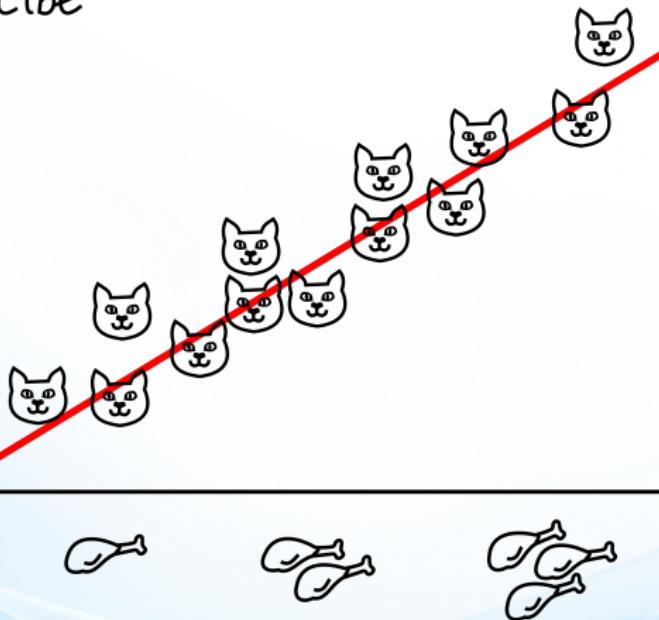
Корреляционный анализ



счастье



окорочки



Формула счастья


$$\text{счастье} = \theta_0 + \theta_1 \times \text{кал-бо} + \text{погрешности}$$



Больше окорочков

счастье



Формула счастья



$$= \theta_0 + \theta_1 \times \text{кал-бо} - \theta_2 \times (\text{кал-бо})^2$$

+ погрешности



Другие факторы

$$\begin{aligned} &= \theta_0 + \theta_1 \times \text{чили} - \theta_2 \times (\text{чили})^2 \\ &\quad + \theta_3 \times \text{шарф} \\ &\quad + \theta_4 \times \text{диван} \\ &\quad + \text{погрешности} \end{aligned}$$

Регрессионный анализ

Классификация котиков

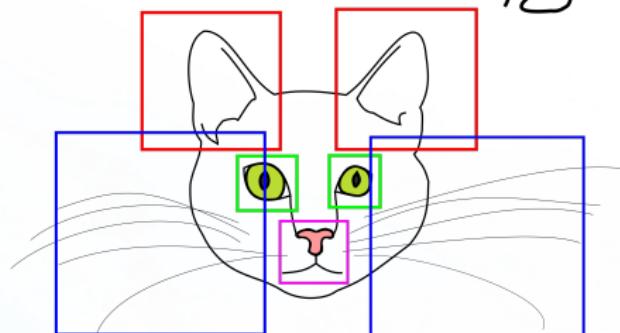


Классификация

Собираем данные

котик	порода	рост	шерсть
	Саванна	50 см	да
	Сфинкс	30 см	нет
	Манчкин	15 см	да
	Саванна	40 см	да

Распознавание мордочек



Нейронные сети

Theta

Художник курса



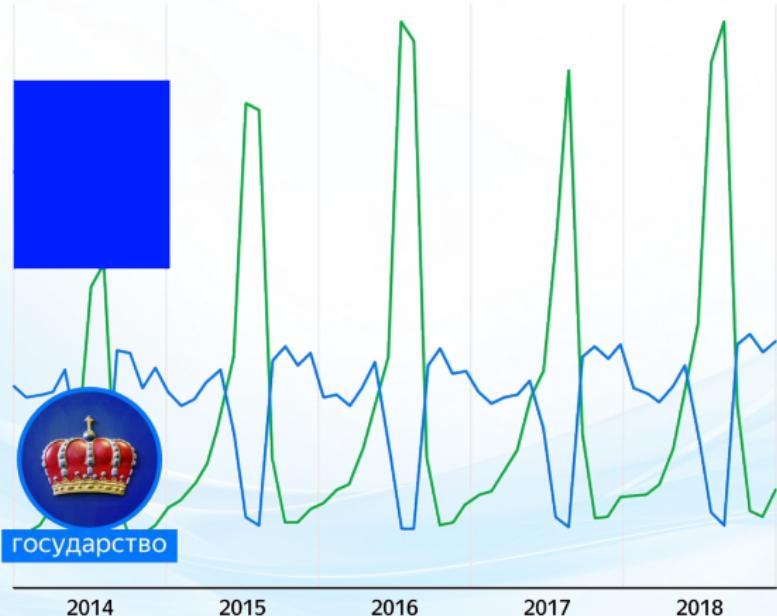
Гаврилова



Книга с похожим содержанием



Когда в Поиске растёт интерес
к [REDACTED] снижается доля запросов
со словом **государство**



Когда в Поиске растёт интерес
к **огурцам**, снижается доля запросов
со словом **государство**



Когда в Поиске растёт интерес к **тату**,
снижается доля запросов со словом **[REDACTED]**



Когда в Поиске растёт интерес к **тату**,
снижается доля запросов со словом **смысл**



Когда в Поиске растёт интерес
к **жилью**, снижается доля запросов
со словом [REDACTED]



Когда в Поиске растёт интерес
к **жилью**, снижается доля запросов
со словом **перспектива**



Президентские выборы в США в 1948 г.

Гарри Труман (демократы) vs. Томас Дьюи (республиканцы)

В ночь на оглашение результатов газета Chicago Tribune опубликовала заголовок: **DEWEY DEFEATS TRUMAN**



После закрытия участков газета провела опрос, обзвонив большое число избирателей, все предвещало оглушительную победу Дьюи.

Президентские выборы в США в 1948 г.

Смеющийся Труман, победитель выборов 1948 года.



Что же пошло не так?

В 1948 году телефон был доступен только людям определенного достатка и редко встречался у людей с небольшим заработком.

Выборка не учитывала достаточно широкий пласт избирателей Трумана, т.к. как правило демократы имеют большую долю голосов среди бедного населения, которым телефон в свою очередь был недоступен.

Попробуем решить задачу

А ты кто?

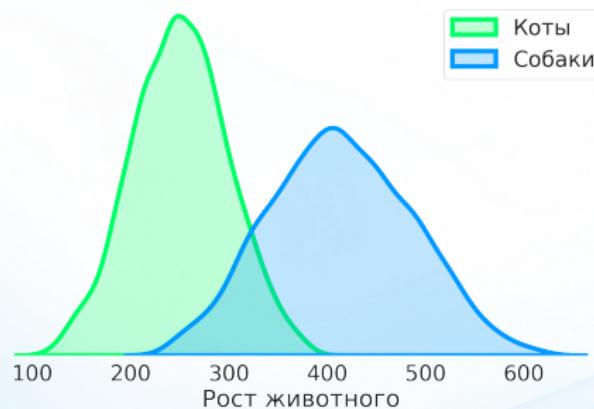
Перед нами домашнее животное. Кто это — собака или кот?



Классификация: собака vs кот

Попробуем сначала извлечь какой-то признак.

Построим вероятностные плотности для каждого класса.



При каких-то значениях роста мы уже можем

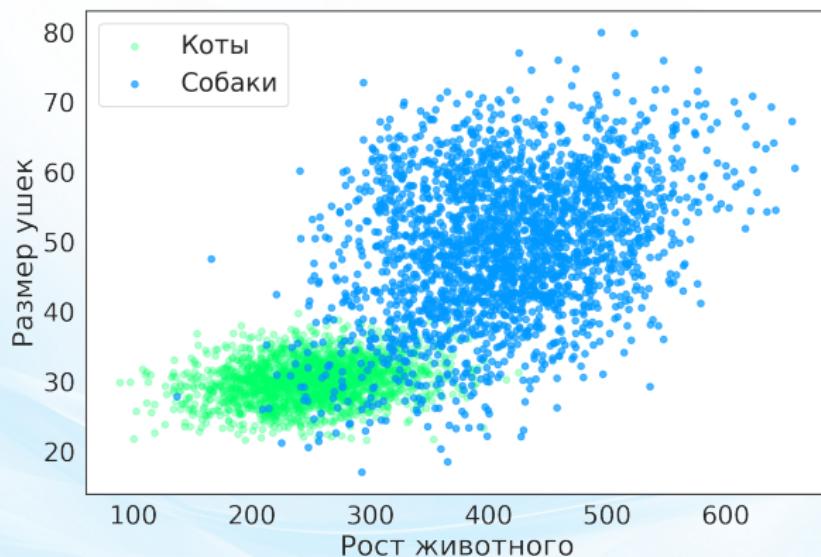
с большой уверенностью сказать ответ.

Но есть большое пересечение, это не очень здорово.

Классификация: собака vs кот

Извлечем еще один признак — размер ушек.

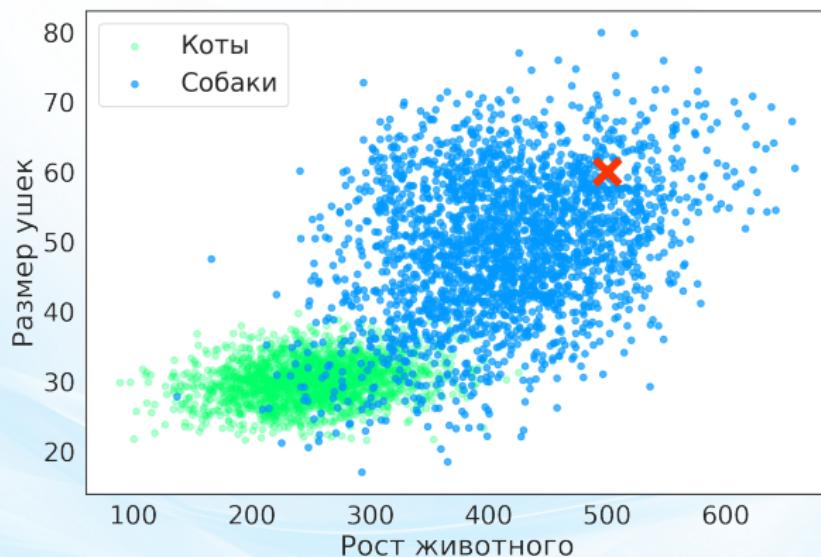
Теперь классы лучше разделяются.



Классификация: собака vs кот

Попробуем классифицировать новое животное.

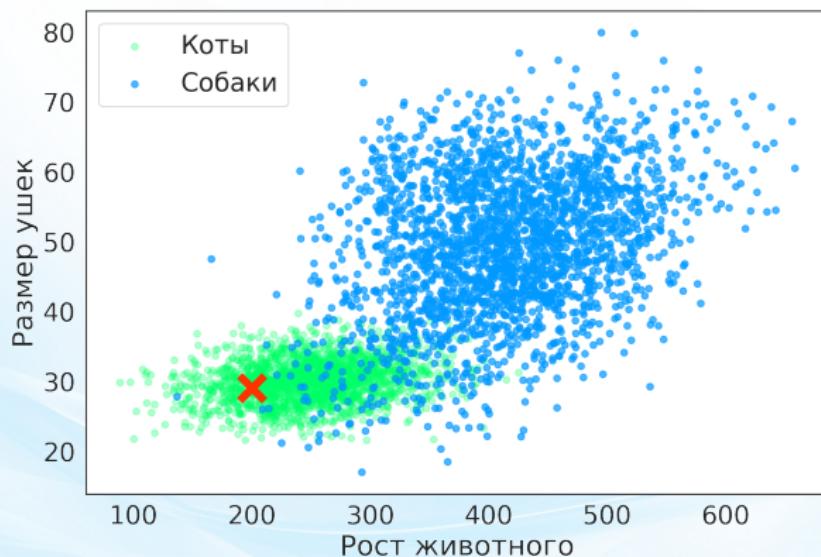
Кто отмечен красным?



Классификация: собака vs кот

Попробуем классифицировать новое животное.

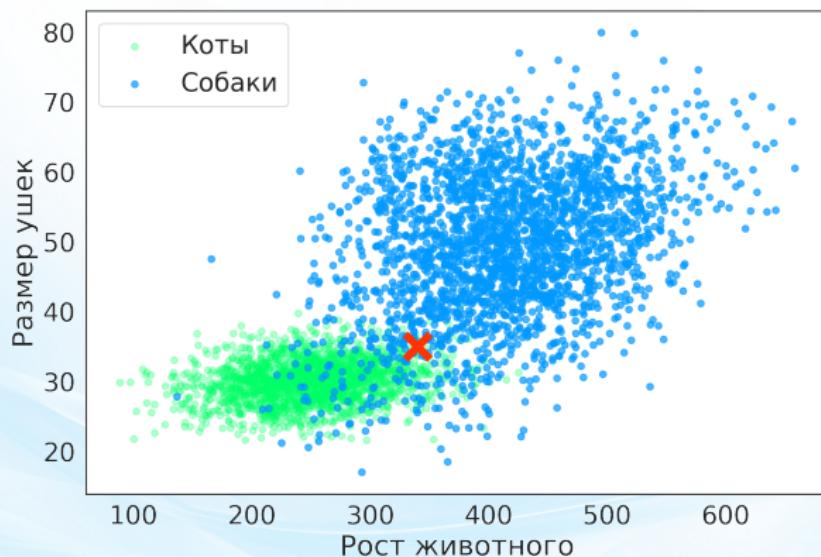
Кто отмечен красным?



Классификация: собака vs кот

Попробуем классифицировать новое животное.

Кто отмечен красным?



На основе чего вы сделали все выводы?

Метод ближайших соседей (kNN)

Дано:

X_1, \dots, X_n — набор размеченных объектов.

Y_1, \dots, Y_n — соответствующие метки класса.

Задача:

Пусть x — исследуемый объект. Какого он класса?

Решение:

Будем смотреть на свойства k ближайших соседей.

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

Y_1, \dots, Y_k — соответствующие им классы.

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

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.

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

Y_1, \dots, Y_k — соответствующий класс.

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

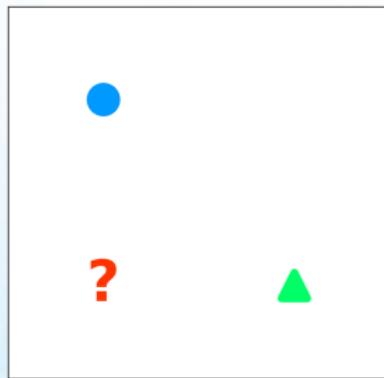
Способы определения веса:

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

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

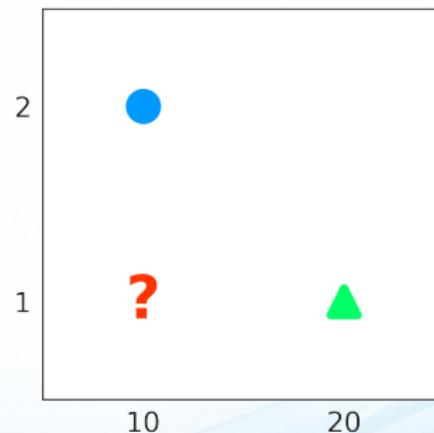
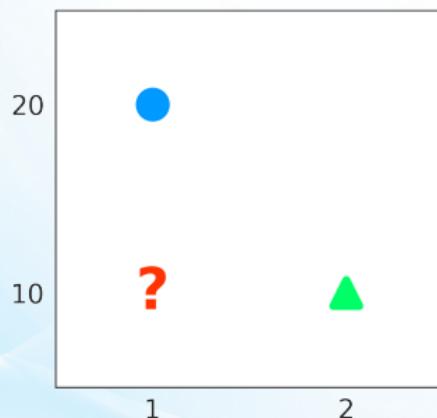
Особенности

Классифицируйте объект "?".



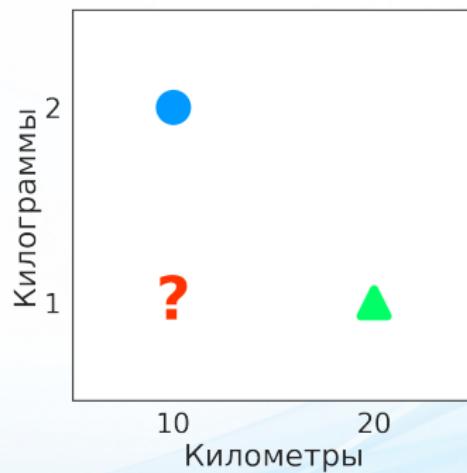
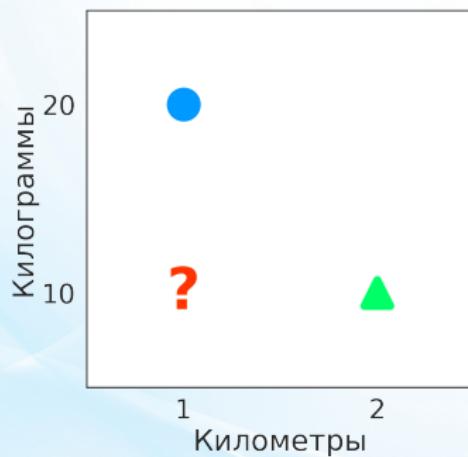
Особенности

Классифицируйте объект "?".



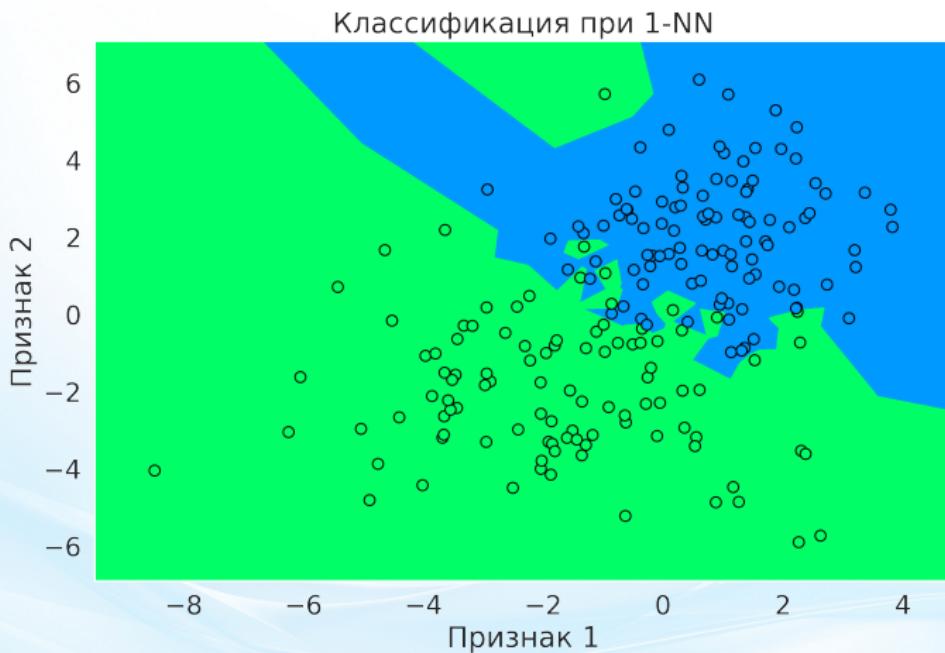
Особенности

Классифицируйте объект "?".

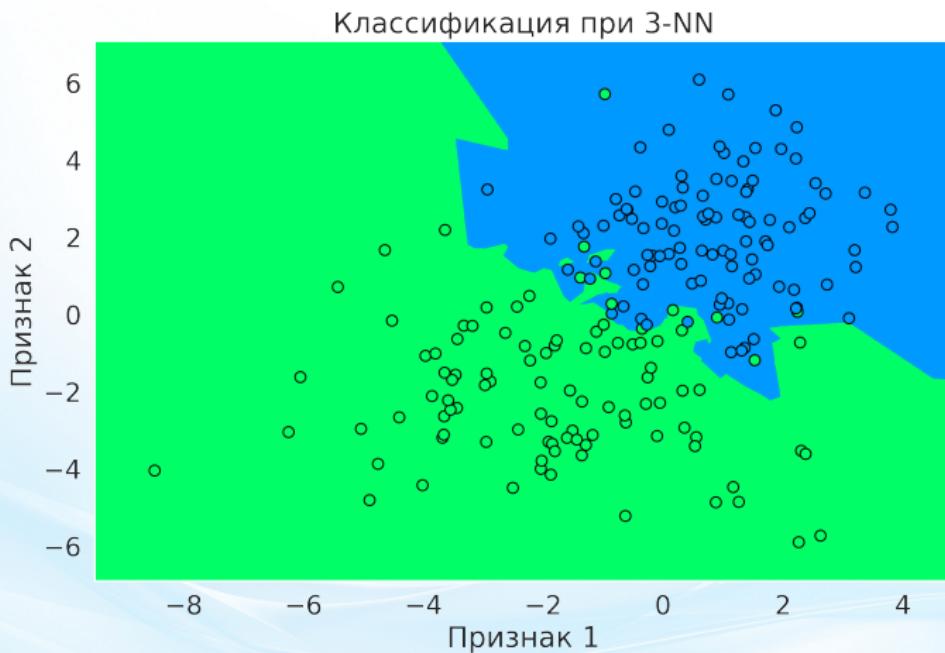


Вывод: результат сильно зависит от используемой метрики между точками в пространстве. Не складывайте кг с км!

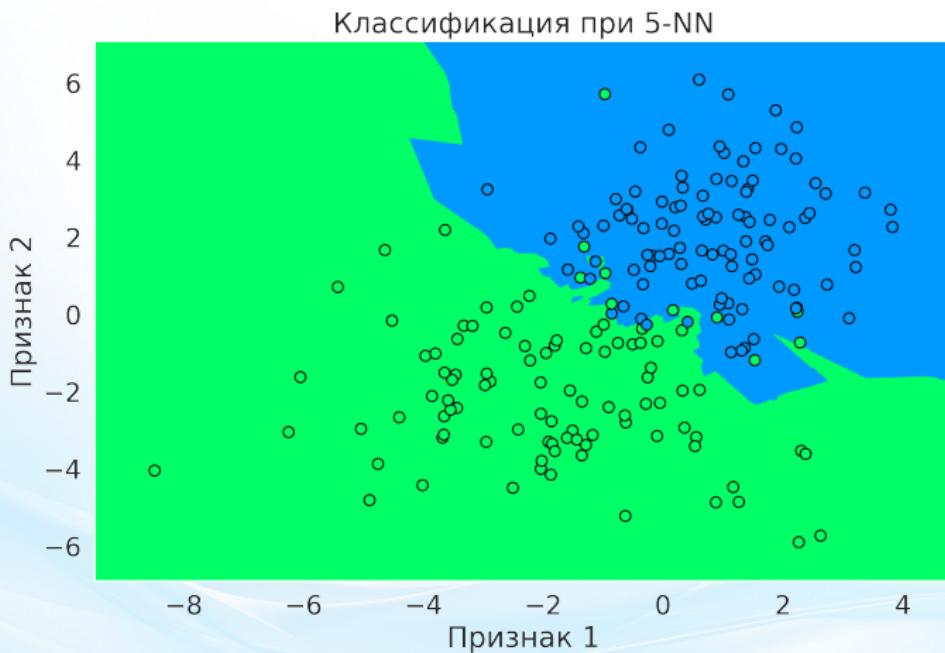
Что происходит при разных k ?



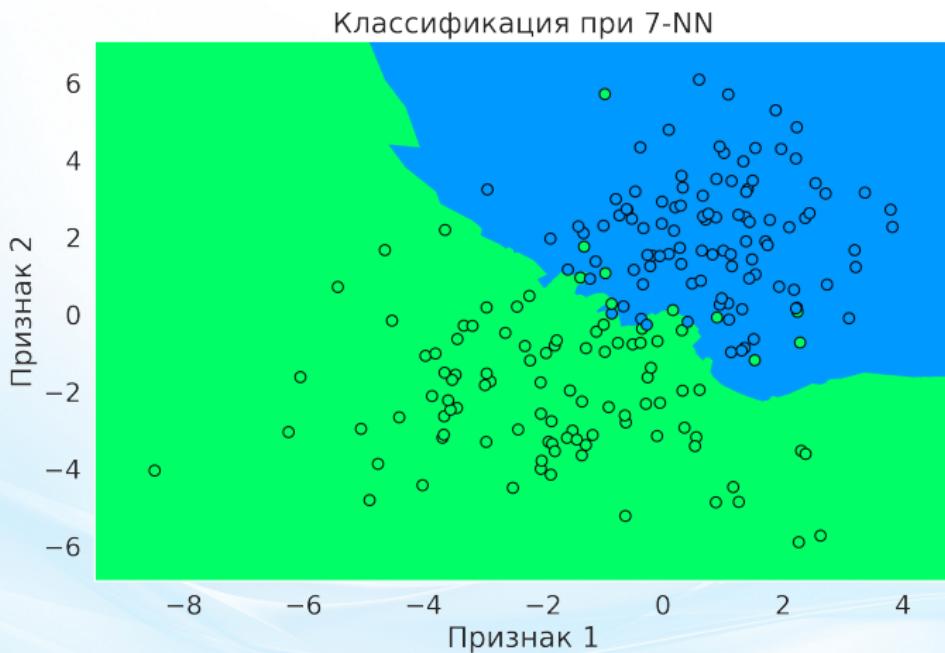
Что происходит при разных k ?



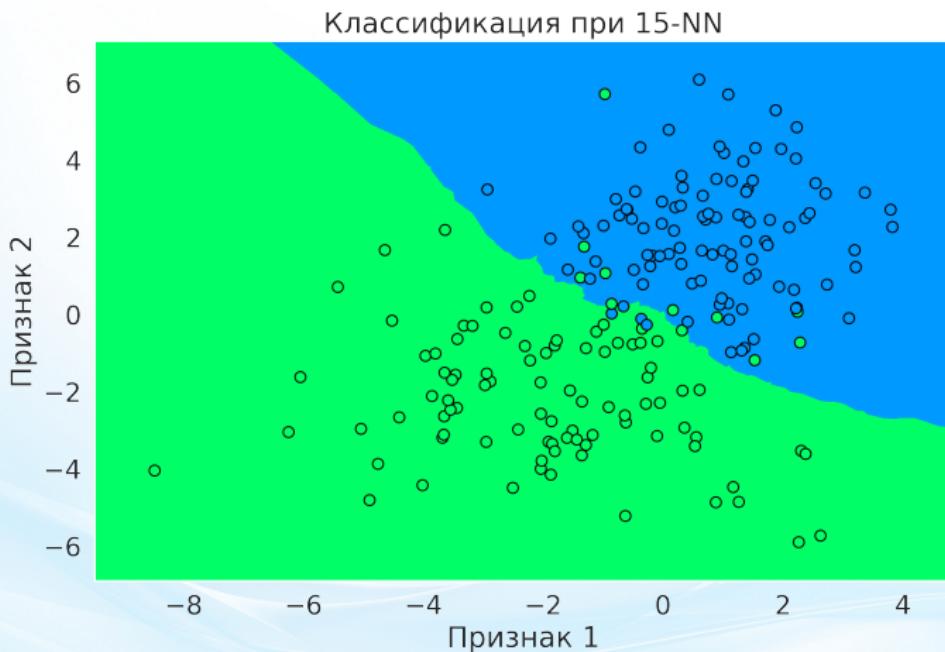
Что происходит при разных k ?



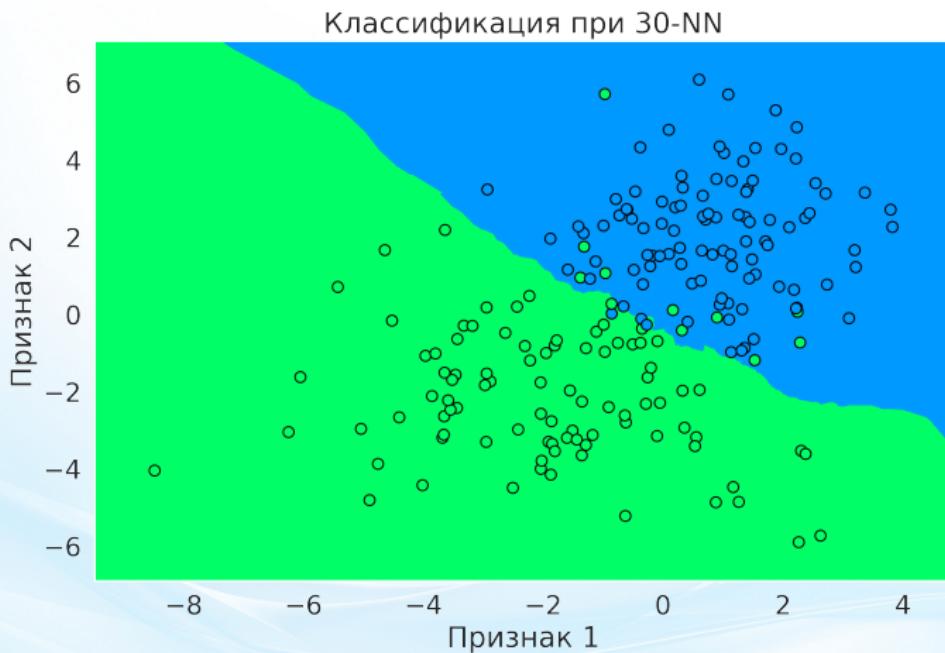
Что происходит при разных k ?



Что происходит при разных k ?



Что происходит при разных k ?



Как оценить качество классификации?

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

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

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

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

Какое число соседей оптимизирует эту метрику?

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

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

- на одной определяют правило классификации,
- 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,

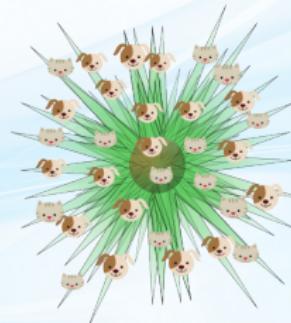
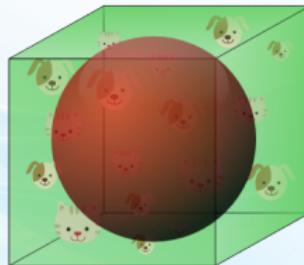
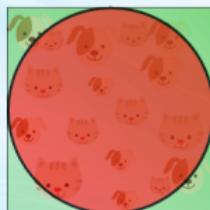
в каждом по 3 числа. Какой размерности получается объект?

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

Проблема:

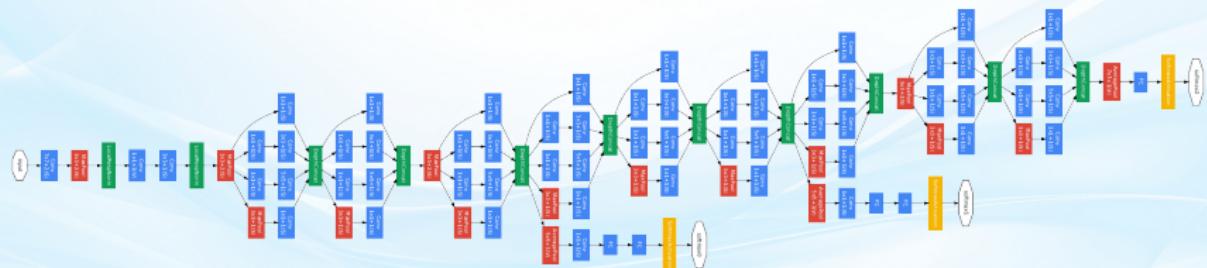
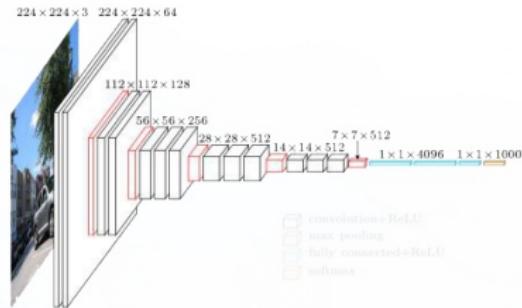
в пр-ве больших размерностей расстояния неинформативны.

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

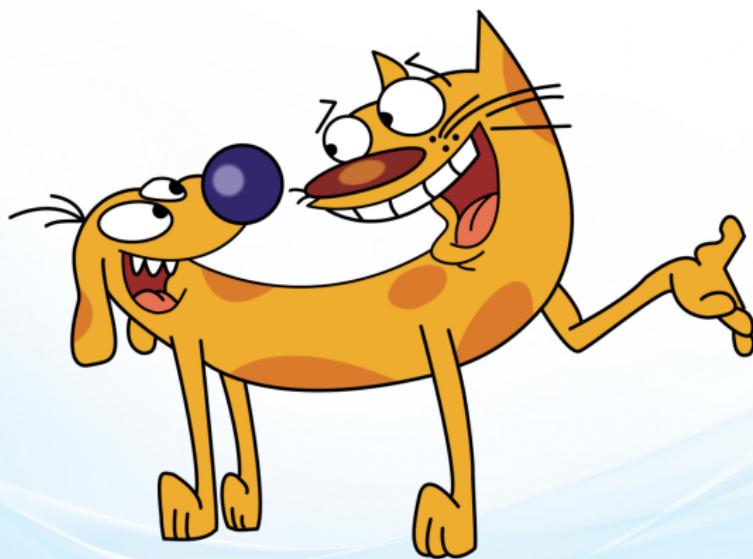


А что в сложных случаях?

Нейросети! Но об этом позже :)



А потом приходит кто?





Примеры прикладных задач

Рекомендательная система

Продуктовая аналитика

Синтез речи

Где еще?



Рекомендательная система

Постановка задачи:

Спроектировать рекомендательную систему для муз. сайта.

Система должна рекомендовать музыку каждому пользователю в соответствии с его предпочтениями.

Данные:

- ▶ оценки пользователей различным трекам;
- ▶ прослушивание треков пользователями.

Яндекс Музыка

Трек, альбом, исполнитель, подкаст

Главное • Подкасты Жанры Радио

исполнитель
David Guetta
Нравится слушателям: Avicii, Rihanna, Calvin Harris

Слушать 2577 575

ГЛАВНОЕ ТРЕКИ АЛЬБОМЫ КЛИПЫ ПОХОЖИЕ ИНФО

Популярные треки

Смотреть всё > Недавний релиз

Memories (2021 Remix) feat. Kid Cudi — David Guetta, K... 2:42

Let's Love — David Guetta, Sia 3:20

Премьера

Премьера
Только новинки, подобранные по вашим предпочтениям
Обновлён 7 февраля

Рекомендательная система

Пусть U — множество треков, I — множество пользователей.

$R = (r_{ui})_{u \in U, i \in I}$ — матрица лайков/прослуш. пользователями треков.

Идея: T — небольшое множество интересов.

$P = (p_{tu})_{t \in T, u \in U}$ — матрица интересов пользователей.

$Q = (q_{ti})_{t \in T, i \in I}$ — матрица соответствия треков интересам.

Если величины p_{tu} и q_{ti} неотрицательны, то их можно нормировать по темам, получив вероятности:

- ▶ $p_{tu} / \sum_{t \in T} p_{tu} = P(\text{пользователю } u \text{ интересно } t),$
- ▶ $q_{ti} / \sum_{t \in T} q_{ti} = P(\text{трек } i \text{ соответствует } t).$

Рекомендательная система

Неотрицательные матричные разложения

Решаем поочередно для каждой t .

$$\begin{cases} \|R_t - p_t^T q_t\|^2 \rightarrow \min_{p_t} \\ p_t \geq 0 \end{cases} \implies p_t = \left(\frac{q_t R_t^T}{q_t q_t^T} \right)^+$$

$$\begin{cases} \|R_t - p_t^T q_t\|^2 \rightarrow \min_{q_t} \\ q_t \geq 0 \end{cases} \implies q_t = \left(\frac{p_t R_t^T}{p_t p_t^T} \right)^+$$

Что нужно помнить при реализации

- ▶ Пользователей и треков миллионы;
- ▶ Матрица R сильно разрежена — лайков очень малая доля.



Примеры прикладных задач

Рекомендательная система

Продуктовая аналитика

Синтез речи

Где еще?



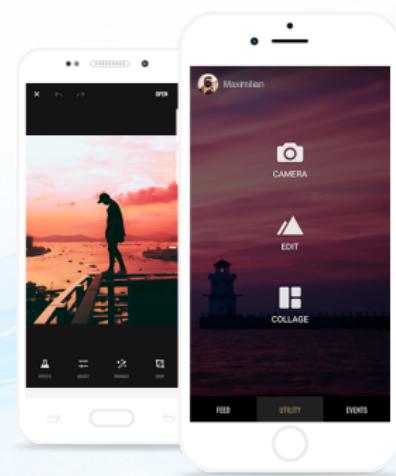
Продуктовая аналитика

Представьте, что вы продуктовый аналитик в фоторедакторе.

Вы всегда хотите, чтобы продукт приносил больше денег.

Реальный вопрос:

пользователи какой страны платят больше?





Продуктовая аналитика

Постановка задачи:

Оказывает ли страна пользователя влияние на его LifeTime Value?

LifeTime Value — доход, который принес пользователь за все время жизни в продукте.

Данные:

- ▶ страна пользователя — фактор;
- ▶ LifeTime Value каждого пользователя.



Продуктовая аналитика

Решение задачи:

Однофакторный дисперсионный анализ и Post-Hoc анализ.

Данные методы помогают понять какие значения фактора влияют на исследуемую величину и как именно.

Т.е., можно узнать, пользователи каких стран платят больше.

Как же выглядят эти методы?

Однофакторный дисперсионный анализ

Независимые выборки

1	2	...	k
X_{11}	X_{12}	...	X_{1d}
X_{21}	X_{22}	...	X_{2d}
...
$X_{n_1 1}$	$X_{n_2 2}$...	$X_{n_k k}$

$$X_{ij} = \underbrace{\mu}_{\mu_j} + \beta_j + \varepsilon_{ij},$$

$i = 1, \dots, n_j$ — номер наблюдения в выборке

$j = 1, \dots, k$ — номер выборки

μ — неизвестное общее среднее

β_j — неизвестный эффект воздействия фактора для j -й выборки

ε_{ij} — случайная ошибка

Предположение:

ε_{ij} независимы и имеют одинаковое непрерывное распределение.

$$H_0: \mu_1 = \dots = \mu_k \quad vs. \quad H_1: \exists j_1, j_2 \text{ т.ч. } \mu_{j_1} \neq \mu_{j_2}$$

Оценка контраста

Пусть H_{rs} отвергается \Rightarrow оцениваем контраст $\Delta_{rs} = \mu_r - \mu_s$.

$V_{rs} = med\{X_{ir} - X_{is}, i = 1..n_r, j = 1..n_s\}$ — первичная оценка

$$W_r = \frac{1}{N} \sum_{s=1}^k n_s V_{rs}, \text{ где } V_{rr} = 0$$

$\hat{\Delta} = W_r - W_s$ — уточненная оценка контраста

Свойства:

1. Первичные оценки могут быть несогласованными: $V_{12} \neq V_{13} + V_{32}$
2. Уточненные оценки согласованы и состоятельны.
3. Уточненные оценки зависят от всех выборок.

Для применения данных методов нужно знать
теорию вероятностей и математическую статистику.

Результат

Теперь мы знаем, в какой стране завести маркетинговую кампанию!

Мы принесли компании деньги, мы молодцы!





Примеры прикладных задач

Рекомендательная система

Продуктовая аналитика

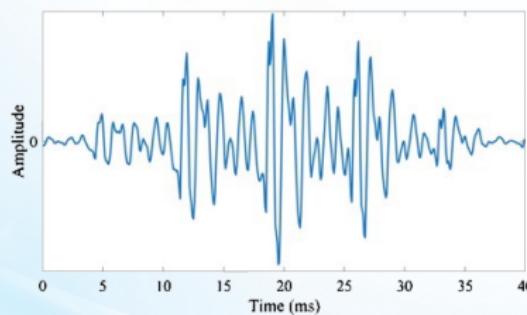
Синтез речи

Где еще?

Синтез речи

Что такое звук?

Звук



Волна



Звук — композиция волн с разными амплитудами и частотой.

Волна — периодич. ф-ия, имеющая амплитуду, период и частоту.

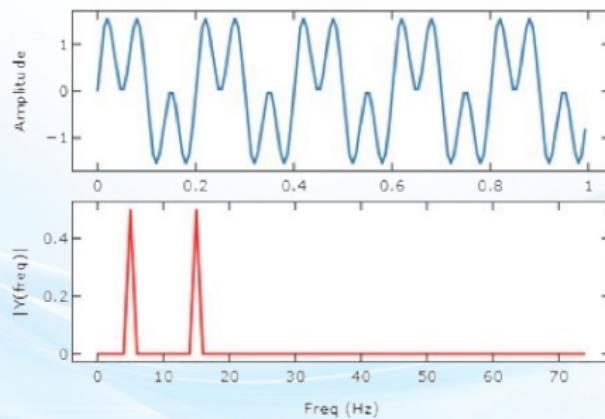
Синтез речи

Как работать со звуком?

Посмотрим на распределение частот волн в звуке.

Для этого применяется дискретное преобразование Фурье:

$$X_k = \sum_{n=0}^{N-1} x_n \cdot e^{-\frac{i2\pi}{N} kn} = \sum_{n=0}^{N-1} x_n \left[\cos\left(\frac{2\pi}{N} kn\right) - i \sin\left(\frac{2\pi}{N} kn\right) \right]$$



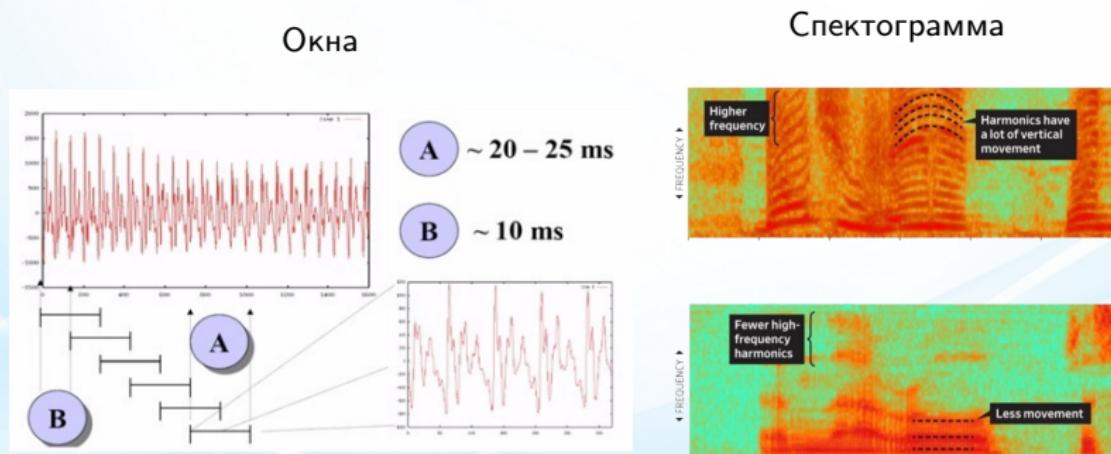
Пример для звука из двух волн

Синтез речи

Как работать со звуком?

У результата преобразования Фурье ко всем данным нет времени.

Посчитаем распределение на окнах из звука.



- ▶ Берем маленькие окна (20-25 мс) от исходных данных.
- ▶ К каждому окну применяем преобразования Фурье.
- ▶ Стакаем распределения частот вместе, получаем спектrogramму.

Синтез речи

Постановка задачи синтеза речи

Для введенного получить соответствующее аудио.

Другие задачи, связанные со звуком:

- ▶ Определить, была ли речь на записи.
- ▶ Споттер — распознавание определенной фразы или определенного события.

Например, “Ok, Google”, “Алиса” или определение выстрела.

- ▶ Идентификация говорящего или его признаков.

Например, мужчина/женщина, ребенок/взрослый.

- ▶ Распознавание речи

По аудио вернуть полную его расшифровку в виду текста.

Синтез речи

1. Акустическая модель по тексту

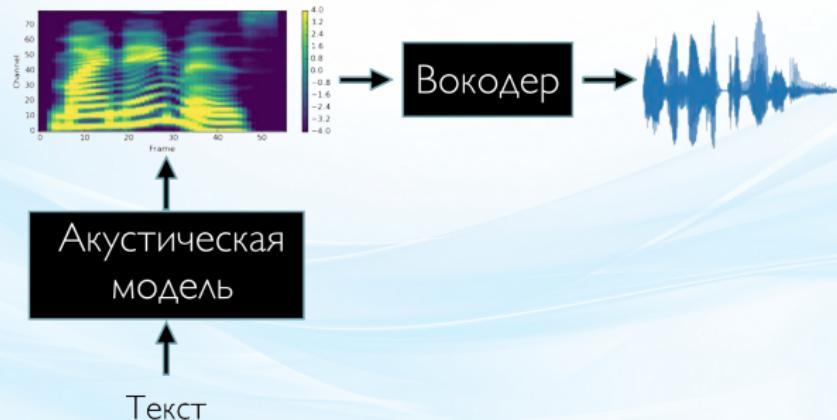
предсказывает мел-спектограмму.

Акустическая модель — это какая-то нейросеть.

2. Вокодер по мел-спектограмме предсказывает аудио.

Вокодер может быть как алгоритмом, так и нейросетью.

Некоторые известные вокодеры используют байесовские методы.





Примеры прикладных задач

Рекомендательная система

Продуктовая аналитика

Синтез речи

Где еще?



- ▶ Беспилотные автомобили
- ▶ Ранжирование результатов в поисковой системе
- ▶ Поиск по картинкам и видео на основе содержания
- ▶ Распознавание спама/флуда
- ▶ Прогноз погоды
- ▶ Прокладывание маршрутов в навигаторах
- ▶ AI-камеры в телефонах
- ▶ Генерация картин подобно художникам
- ▶ Оплата проезда в метро взглядом
- ▶ Решение о выдаче кредита
- ▶ Персонализированная реклама
- ▶ Определение причин оттока клиентов
- ▶ Прогнозирование спроса на товар
- ▶ Определение месторасположения новой торговой точки
- ▶ Расстановка полок в магазинах
- ▶ Автоопределение свежести товаров в магазинах
- ▶ Распознавание патологий на медицинских снимках
- ▶ Персонализированная медицина
- ▶ Прогнозирование поломок оборудования
- ▶ Автоматическая система оптимального управления оборудованием

Theta



BCE !