Математическая статистика

и основы анализа данных

Основы машинного обучения

План занятия

- Задача регрессии
  - Линейная регрессия
  - Метрики качества для задачи регрессии
- Задача классификации
  - Классификатор kNN
  - Метрики качества для задачи классификации

Линейная регрессия

Метод наименьших квадратов

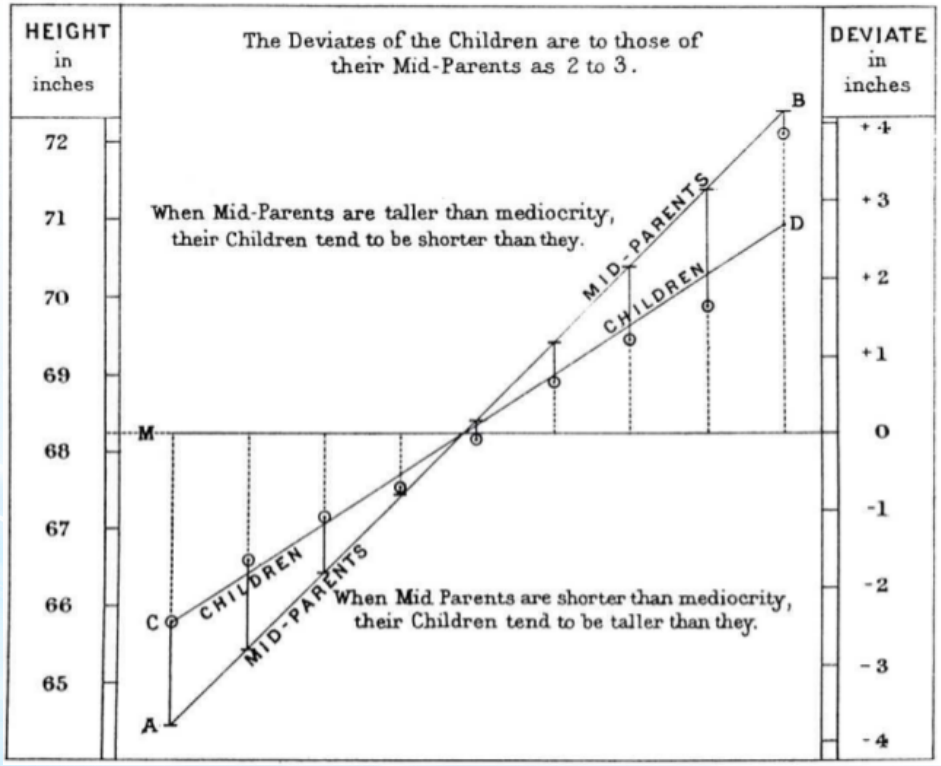
Первое упоминание регрессии

Впервые регрессия упоминается в работе Гальтона

"*Регрессия к середине в наследственности роста*", 1885 г.

x — рост родителей, y — рост детей

Установлена зависимость $y - \overline{y} \approx \frac{2}{3}(x - \overline{x})$, т.е. регрессия к середине.



Задача регрессии: интуиция

Есть объект, обладающий признаками x.

Примеры признаков: рост песика, экспрессия белка, энергия частицы.

Мы предполагаем, что есть зависимость какой-то численной характеристики объекта y от его признаков:

$$y \approx f(x)$$

Пример: зависимость уровня когнитивных способностей от параметров поражения мозга при рассеянном склерозе.

Однако мы не знаем, какова эта зависимость на самом деле.

На основании данных – набора объектов, для которых известны x и y, мы пытаемся "восстановить" зависимость:

$$y \approx \widehat{f}(x)$$

Пример

Пусть x — рост песика, а y — его вес.

Что мы знаем?

- чем крупнее песик, тем больший вес он имеет;
- песики одинакового роста могут иметь разный вес.

Выводы:

- для фиксированного роста песика x

  его вес $y = f(x)$ является случайной величиной;
- в среднем вес $f(x)$ возрастает при увеличении роста песика x.

Пример

Простая зависимость:

$$y = \theta_0 + \theta_1 x + \varepsilon,$$

x — рост песика,

y — вес песика,

θ_0, θ_1 — неизвестные параметры,

ε — случайная составляющая с нулевым средним (погрешность).

Зависимость **линейна по параметрам**, линейна по аргументу.

Пример

Более сложная зависимость:

$$y = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_2^2 + \varepsilon,$$

x_1 — рост песика,

x_2 — обхват туловища песика,

y — вес песика,

$\theta_0, \theta_1, \theta_2, \theta_3$ — неизвестные параметры,

ε — случайная составляющая с нулевым средним.

Зависимость **линейна по параметрам**, квадратична по аргументам.

Модель линейной регрессии

Рассматриваем функциональную зависимость вида

$$y = y(x) = \theta_1 x_1 + ... + \theta_d x_d$$

$x_1, ..., x_d$ — признаки ,

$\theta = (\theta_1, ..., \theta_d)^T$ — вектор параметров.

Для оценки θ производится n испытаний вида

$$Y_i = \theta_1 x_{i1} + ... + \theta_d x_{id} + \varepsilon_i, \quad i = 1, ..., n,$$

$x_i = (x_{i1}, ..., x_{id})$ — признаковые описания объекта i

(обычно неслучайные),

ε_i — случайная ошибка измерений.

Модель линейной регрессии

Введем обозначения

$$Y = \begin{pmatrix} Y_1 \\ \dots \\ Y_n \end{pmatrix}, \quad X = \begin{pmatrix} x_{11} & \dots & x_{1d} \\ \dots & & \\ x_{n1} & \dots & x_{nd} \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \dots \\ \varepsilon_n \end{pmatrix}.$$

Матричная форма записи проведенных испытаний

$$Y = X\theta + \varepsilon.$$

$X \in \mathbb{R}^{n \times d}$ — регрессоры (или матрица плана эксперимента),

$Y \in \mathbb{R}^n$ — отклик.

Матричный вид зависимости: $y(x) = x^T\theta$.

Замечание

Зависимость $y = y(x)$ должна быть **линейна по параметрам**, но не обязана быть линейной по признакам.

Пусть $z_1, ..., z_k$ — набор "независимых" переменных.

Можно рассматривать модель

$$y(x) = \theta_1 x_1(z_1, ..., z_k) + ... + \theta_d x_d(z_1, ..., z_k),$$

где $x_j(z_1, ..., z_k)$ — некоторые функции (м.б. нелинейные).

Примеры:

- $x(z_1, ..., z_k) = 1$;
- $x(z_1, ..., z_k) = z_1$;
- $x(z_1, ..., z_k) = \ln z_1$;
- $x(z_1, ..., z_k) = z_1^2 z_2$.

Определение моментов инерции твёрдых тел с помощью трифилярного подвеса

- На платформу помещается тело — диск, разрезанный по диаметру;
  - I — момент инерции тела;
  - m — масса тела;
  - h — расстояние от половинок до оси вращения;
  - I_0 — момент инерции нераздвинутого диска.
- Половинки диска постепенно раздвигаются;
- Снимается зависимость момента инерции системы I от h.

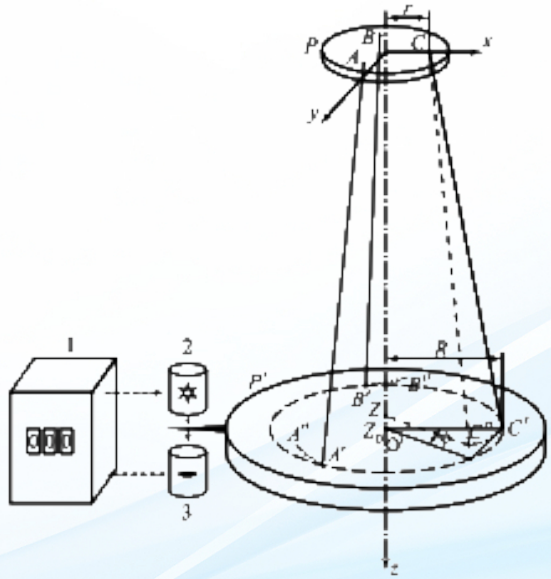


Рис. 2. Трифилярный подвес

По материалам "Модели и концепции физики: механика. Лабораторный практикум"

Пример: Момент инерции

Согласно теореме Гюйгенса-Штейнера должно выполняться:

$$I = I_0 + mh^2$$

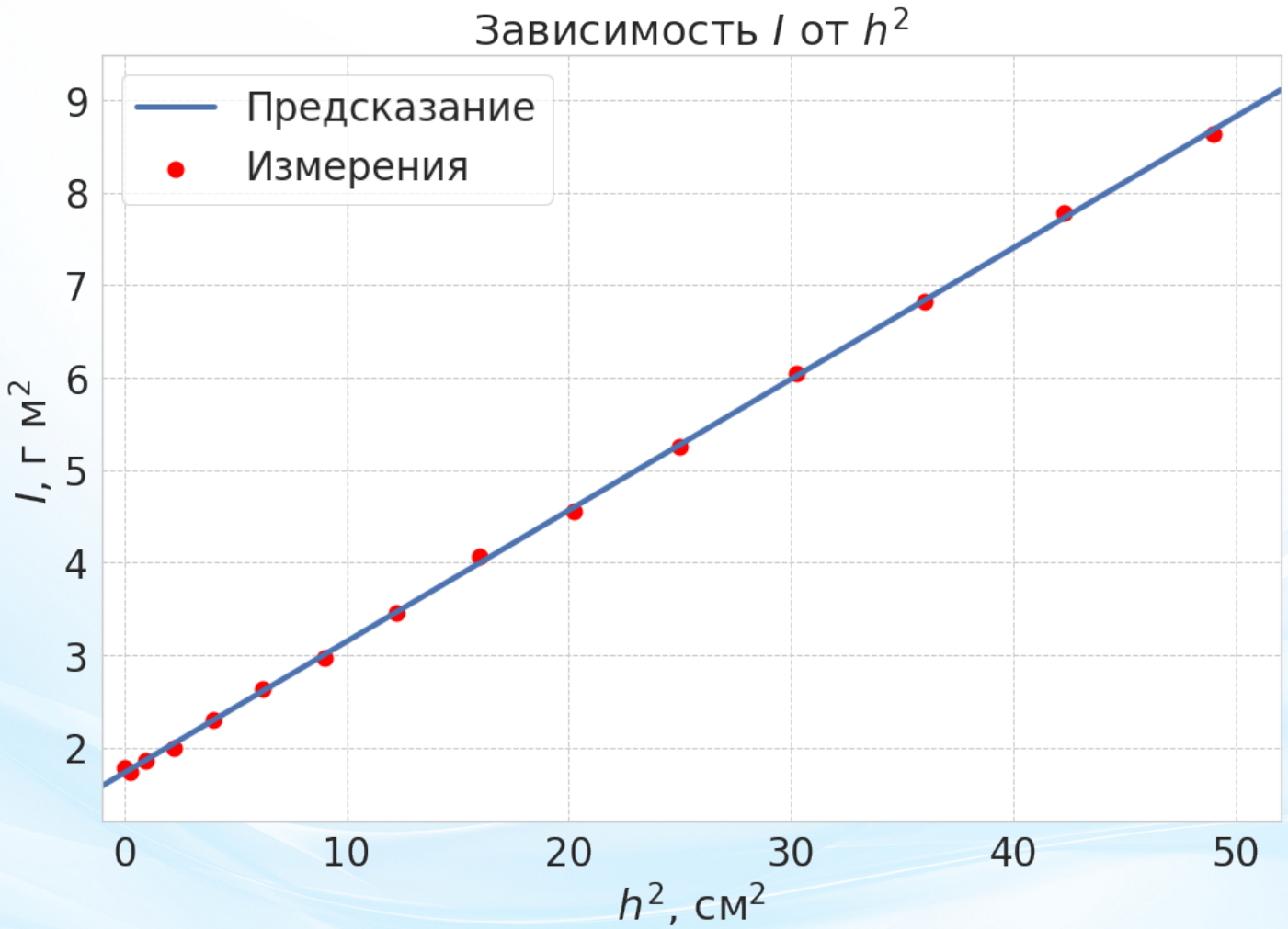
Итого, предполагается линейная зависимость момента инерции I от квадрата расстояния h^2. Мы хотим найти неизвестные m и I_0.

Наблюдения: $I_i = I_0 + mh_i^2 + \varepsilon_i$, где ε_i — погрешность.

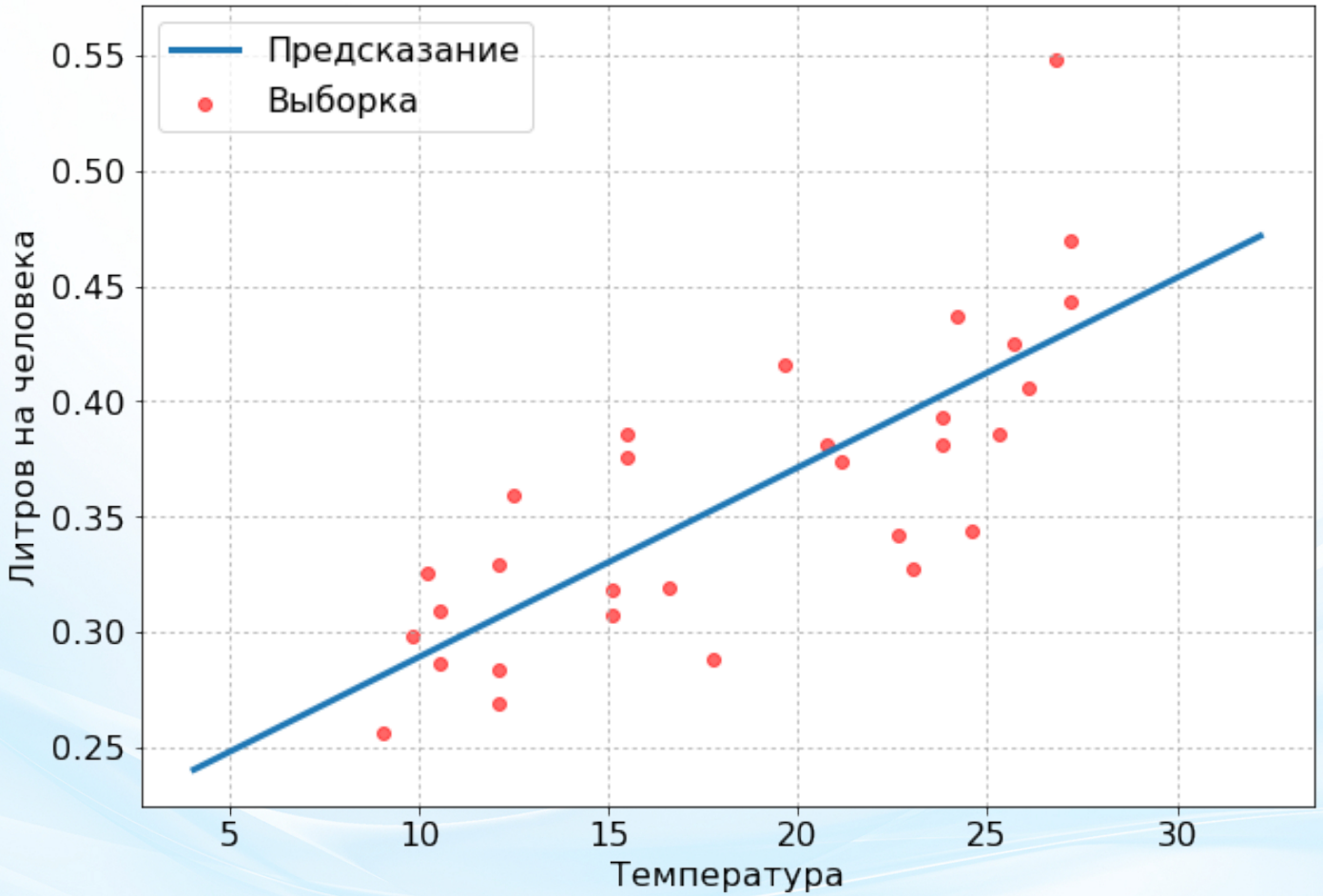
В данном примере $x_1(t) = 1, x_2(h) = h^2$,

$$X = \begin{pmatrix} 1 & h_1^2 \\ \dots & \\ 1 & h_n^2 \end{pmatrix}, Y = \begin{pmatrix} I_1 \\ \dots \\ I_n \end{pmatrix}, \theta = \begin{pmatrix} I_0 \\ m \end{pmatrix}.$$

Пример: Момент инерции



Пример: Потребление мороженого



Метод наименьших квадратов

Зависимость: $y(x) = x^T\theta$, $\theta \in \mathbb{R}^d$.

Испытания: $Y = X\theta + \varepsilon$, $X \in \mathbb{R}^{n\times d}$, $Y \in \mathbb{R}^n$.

Хотим как-то оценить параметр θ на основании полученных данных.

Пусть $\widehat{\theta} = \widehat{\theta}(X, Y)$ — наша оценка θ.

Как понять, что она хорошая?

Метрика MSE:

$$MSE(\widehat{\theta}) = \left|\left|Y - X\widehat{\theta}\right|\right|^2$$

Оценка $\widehat{\theta} = \arg\min_{\theta} MSE(\widehat{\theta})$ называется оценкой по методу наименьших квадратов параметра θ.

Метод наименьших квадратов

Теорема. Если матрица $X^T X$ невырождена, то $\widehat{\theta} = (X^T X)^{-1} X^T Y$.

$$MSE(\theta) = ||Y - X\theta||^2 = (Y - X\theta)^T (Y - X\theta) = Y^T Y - 2Y^T X\theta + \theta^T X^T X\theta$$

Берем производную по θ и приравниваем ее к нулю.

$$\frac{\partial MSE(\theta)}{\partial \theta} = -2Y^T X + 2\theta^T X^T X = 0$$

Отсюда получается утверждение теоремы. $\square$

Предсказанием отклика на новом объекте x будет величина $\widehat{y}(x) = x^T \widehat{\theta}$.

Реализация в sklearn

```
m = sklearn.linear_model.LinearRegression(fit_intercept=True)
```

Обучение модели:

```
m.fit(X, Y)
```

Вектор коэффициентов:

```
m.coef_
```

Свободный коэффициент:

```
m.intercept_
```

Предсказания:

```
m.predict(X)
```

Метрики качества

в задаче регрессии

Обозначения

Пусть $x_1, \dots x_n$ — признаковые описания объектов;

$Y = (Y_1, \dots Y_n)^T$ — наблюдения.

Пусть $\widehat{f}(x)$ — оцененная нами зависимость.

В случае линейной регрессии $\widehat{f}(x) = x^T\widehat{\theta}$.

Пусть $\widehat{Y}_i = \widehat{f}(x_i)$ — предсказание нашей модели на i-м объекте;

$\widehat{Y} = (\widehat{Y}_1, \dots \widehat{Y}_n)^T$.

Метрики качества в задаче регрессии

Y — реальные наблюдения, $\widehat{Y}$ — предсказания.

- MSE (Mean Squared Error):

$$MSE(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

- MAE (Mean Absolute Error):

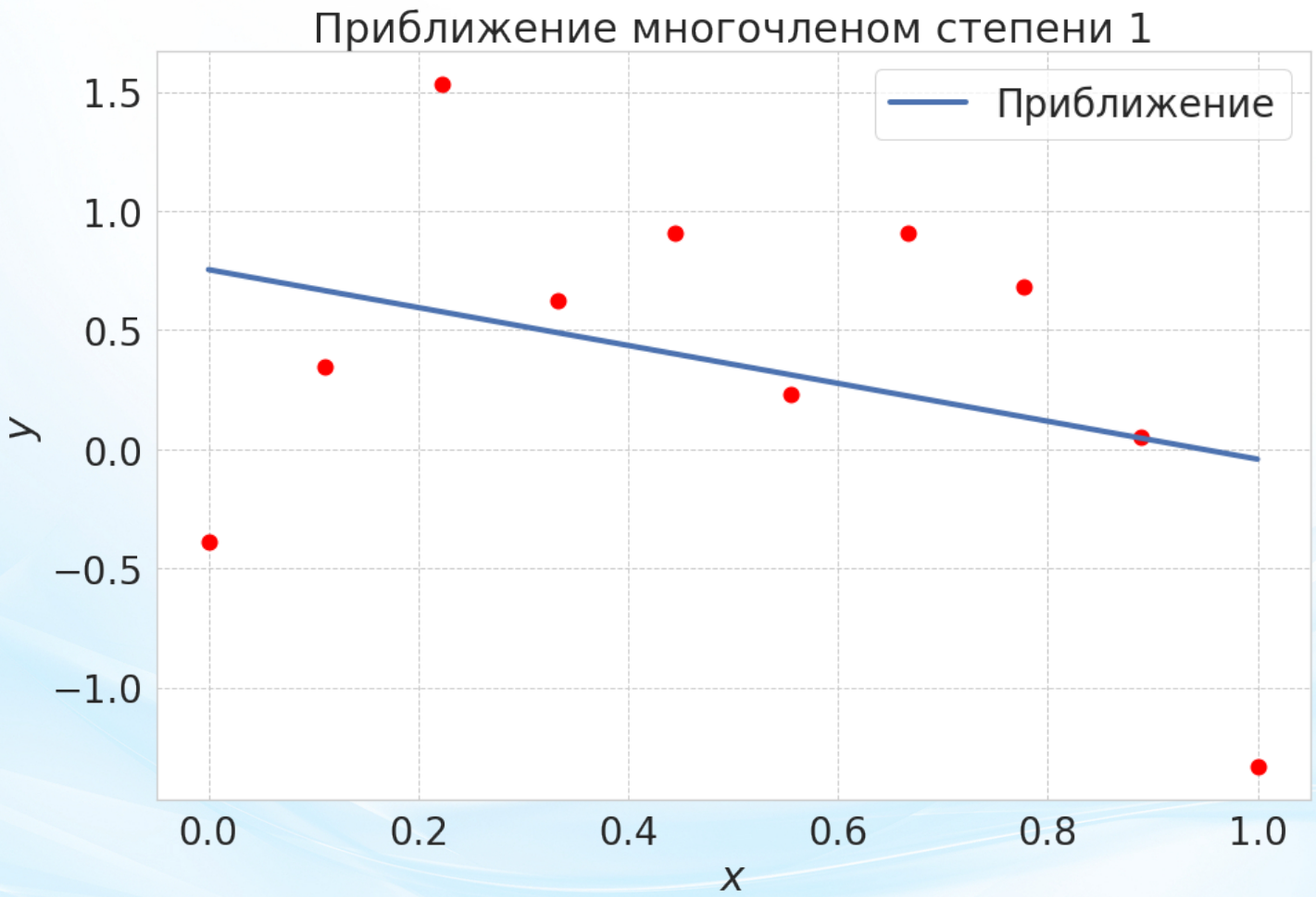
$$MAE(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \widehat{Y}_i|$$

- MAPE (Mean Absolute Percentage Error):

$$MAPE(Y, \widehat{Y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{Y_i - \widehat{Y}_i}{Y_i}\right| * 100\%$$

Недообучение vs Переобучение

Зависимость: $y = 5x - 6x^2$, имеется погрешность



Недообучение

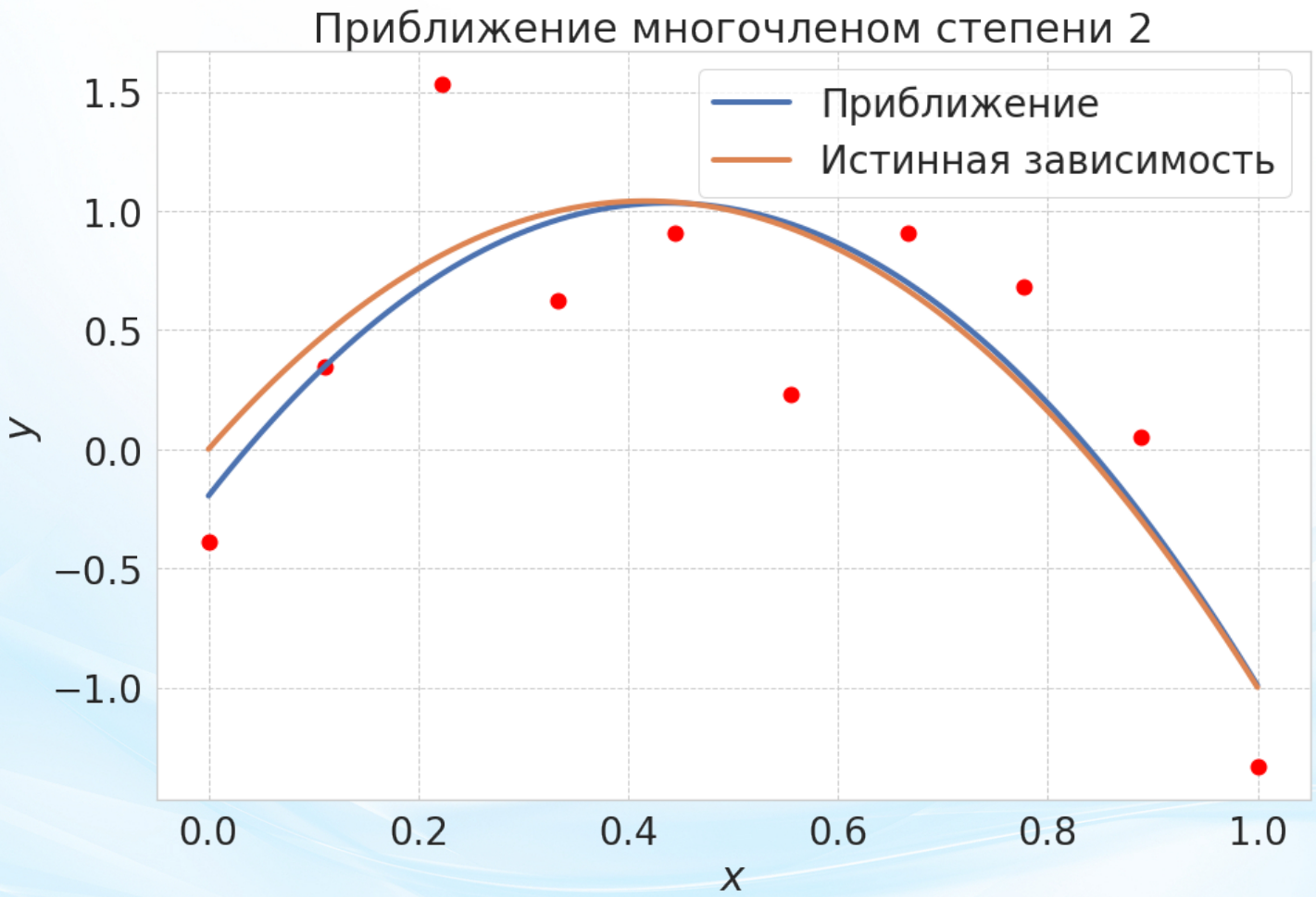
Недообучение vs Переобучение

Зависимость: $y = 5x - 6x^2$, имеется погрешность

Переобучение

Недообучение vs Переобучение

Зависимость: $y = 5x - 6x^2$, имеется погрешность



Нормально!

Тренировочная и тестовая выборки

Если все время работать с одной и той же выборкой
(это жаргон, корректно понимать "реализацией выборки")
и все больше улучшать модель, "подгонять" ее под выборку,

может возникнуть переобучение.

Предсказание на новом объекте может быть неадекватным.

Поэтому перед началом работы имеющиеся данные делят на две части:

тренировочную (обучающую) и тестовую выборки.

На тренировочной выборке происходит обучение моделей (например, оценка коэффициентов в линейной регрессии).

На тестовой выборке происходит оценка качества итоговой модели с использованием метрик качества.

Задача классификации

Классификатор KNN

Задача классификации: интуиция

У нас есть объекты, заданные своими признаковыми описаниями x.

Пример: рост, вес, уровни экспрессии генов, фотография куска неба.

У каждого объекта есть класс y — значение из конечного множества классов C. *Пример: порода песика, тип клетки, индикатор наличия звезды.*

На основании данных — объектов, для которых известны x и y, хотим построить зависимость $y \approx \widehat{f}(x)$.

Тогда мы сможем предсказывать класс на новом объекте по его признаковому описанию x: $\widehat{y} = \widehat{f}(x)$

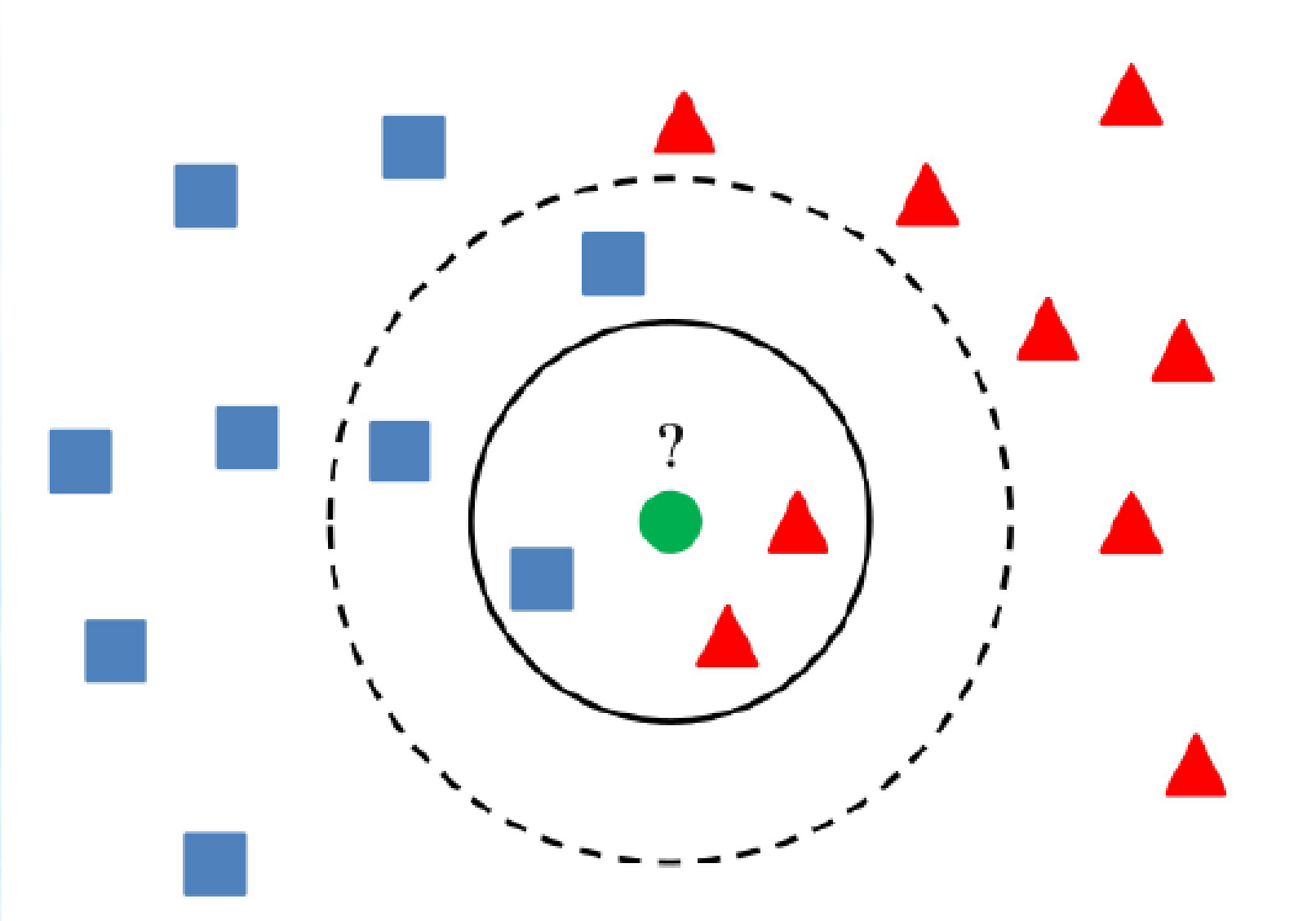
Бинарная классификация: классов всего два.

Пример: является ли клетка раковой.

Многоклассовая классификация: несколько классов.

Пример: тип клетки (нейрон, глия, эпителий, etc).

Метод ближайших соседей (kNN)



Картинка: https://www.researchgate.net

Метод ближайших соседей (kNN)

Пусть $\mathscr{X}$ — метрическое пространство.

$x_1, ..., x_n \in \mathscr{X}$ — обучающая выборка (признаки объектов).

$Y_1, ..., Y_n$ — соответствующая целевая переменная (например, класс).

Предположение:

Свойства объекта меняются не сильно в его окрестности.

Тогда давайте смотреть на свойства k ближайших соседей, где k — параметр.

Пусть $x \in \mathscr{X}$ — исследуемый объект.

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

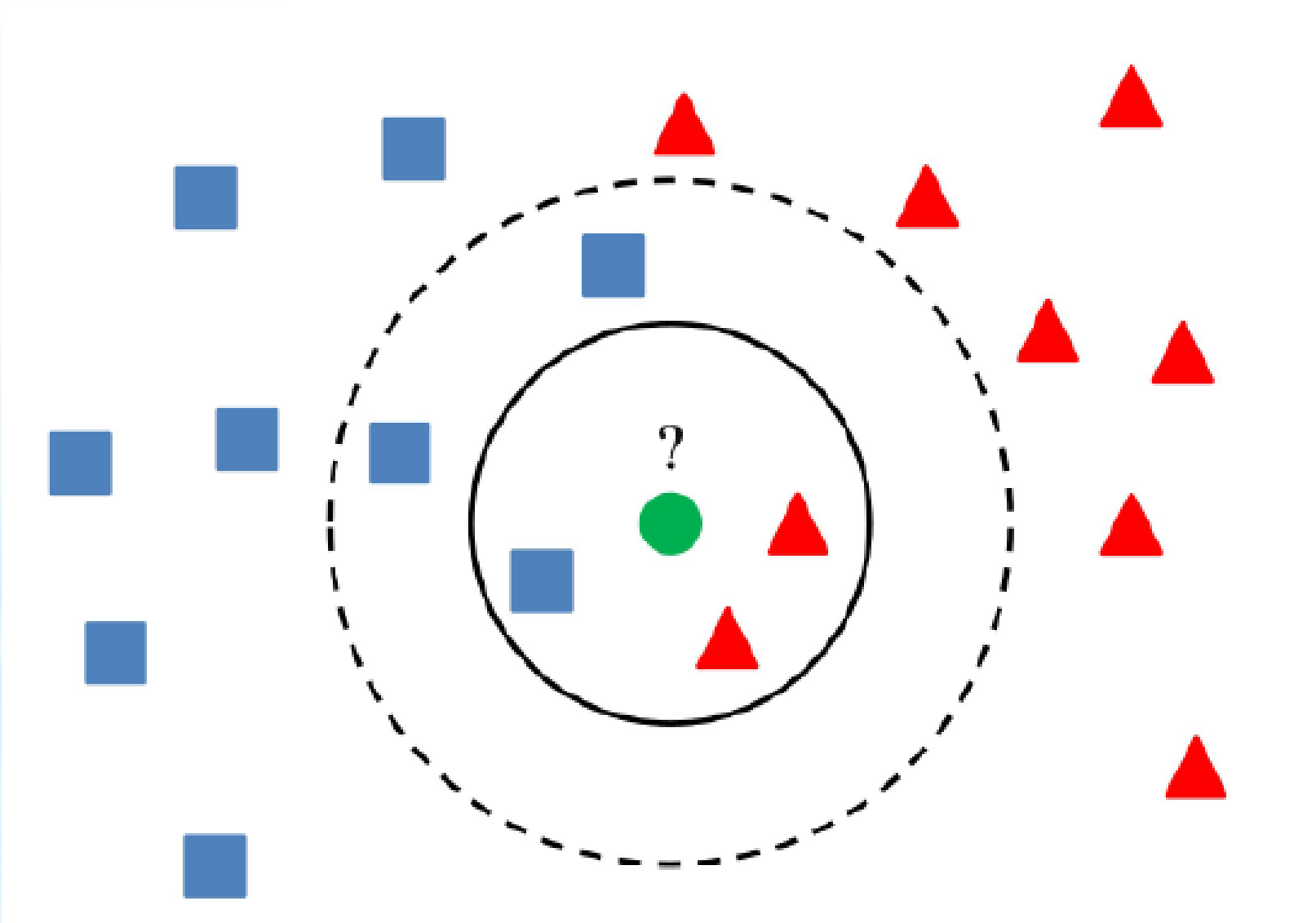
1. *Классификация.*

   Предсказание — наиболее часто встречаемый класс.

2. *Регрессия.*

   Предсказание — усреднение отклика по соседям.

А одинаково ли важны соседи?



Картинка : https://www.researchgate.net

Взвешенный kNN для задачи классификации

Пусть $x \in \mathscr{X}$ — исследуемый объект.

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

$Y_{(1)}, \ldots, Y_{(k)}$ — соответствующие классы, C — множество классов.

$w_1, \ldots, w_k$ — вес/вклад j-го соседа, определяемый пользователем.

Способы определения веса:

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

Тогда "оптимальный" класс в задаче классификации можно определить так:

$$\widehat{y}(x) = \arg\max_{c \in C} \sum_{j=1}^{k} w_j I\{Y_{(j)} = c\}$$

Смысл формулы:

- Складываем веса соседей для каждого класса;
- Берем класс с наибольшим суммарным весом.

Метрика accuracy в задаче классификации

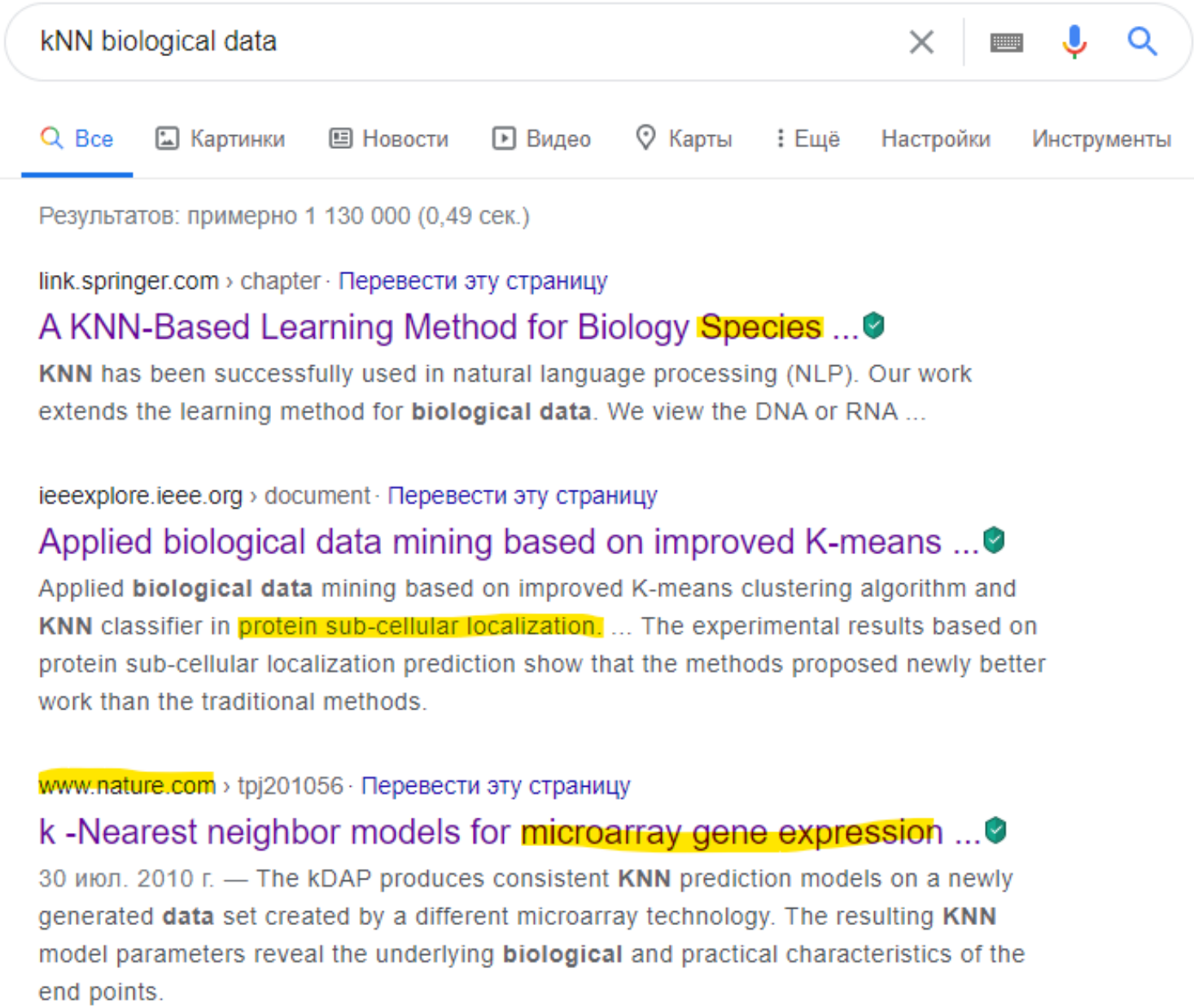
Данная метрика является самой тривиальной метрикой для задачи классификации и представляет собой долю правильных ответов.

$Y_1, \ldots, Y_n$ — истинные классы объектов.

$\widehat{Y}_1, \ldots, \widehat{Y}_n$ — предсказанные классы объектов, т.е. $\widehat{Y}_i = \widehat{f}(x_i)$.

$$accuracy(Y, \widehat{Y}) = \frac{1}{n} \sum_{i=1}^{n} I\{Y_i = \widehat{Y}_i\}$$

Применения kNN на практике



Применения kNN на практике

A KNN-Based Learning Method for Biology Species Categorization

Dang et al., https://link.springer.com/chapter/10.1007/11539087_127

- Категоризация видов;
- Работают с ДНК и РНК разных видов;
- Строят признаки по последовательности ДНК/РНК методами анализа текстов (след. семестр);
- Для полученных признаков строят классификатор kNN;
- Категоризовали 43 бактерии.

Применения kNN на практике

Applied biological data mining based on improved K-means clustering algorithm and KNN classifier in protein sub-cellular localization

Zhenfeng Lei, Shunfang Wang, https://ieeexplore.ieee.org/document/7883060

- Локализация белков в клетках;
- Строят признаковое описание белков;
- Придумали различные улучшения kNN;

k-Nearest neighbor models for microarray gene expression analysis and clinical outcome prediction

Parry et al., https://www.nature.com/articles/tpj201056

- Датасеты с тремя типами рака;

  breast cancer, neuroblastoma, multiple myeloma
- Экспериментировали с признаками, метриками, etc;
- Построили 463320 моделей KNN...

Применения kNN на практике

A Regression-based K nearest neighbor algorithm for gene function prediction from heterogeneous data

Zizhen Yao, Walter L Ruzzo,

https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-7-S1-S11

- Предсказывают функцию генов;
- Оптимизируют метрику с помощью регрессии;
- Метрика — "взвешенная комбинация базовых метрик" (напоминает линейную регрессию!);
- Предсказывают функции генов *Escherichia coli* и сравнивают с известными.


ВСЁ!