

Методы машинного обучения

Лекция 2

Задачи анализа данных. Метод k ближайших соседей.

Эльвира Зиннурова

elvirazinnurova@gmail.com

НИУ ВШЭ, 2019

Напоминание

- \mathbb{X} — пространство объектов, \mathbb{Y} — пространство ответов
- $x = (x^1, \dots, x^d)$ — признаковое описание
- $X = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка
- $a(x)$ — алгоритм, модель
- $Q(a, X)$ — функционал ошибки алгоритма a на выборке X
- Обучение: $a(x) = \arg \min_{a \in \mathcal{A}} Q(a, X)$

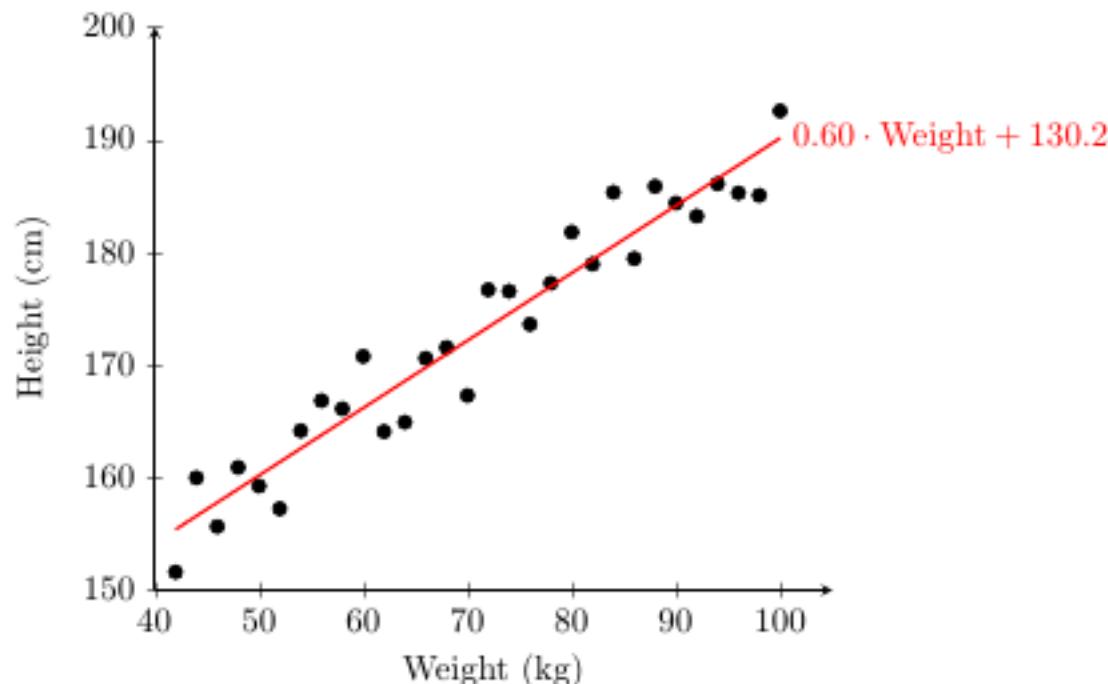
Вопросы на сегодня

- Какие бывают ответы?
- Какие бывают признаки?
- Какие задачи можно решать машинным обучением?
- Что такое вектор/матрица?
- Какие предположения лежат в основе модели kNN?

Типы ответов

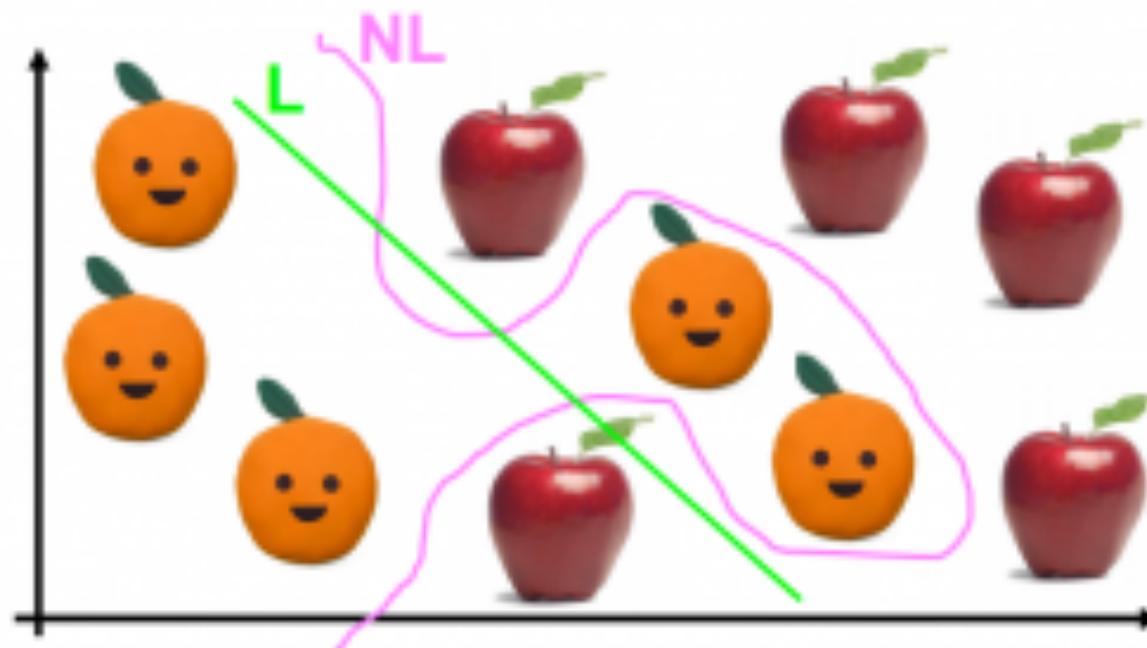
Регрессия

- Вещественные ответы: $\mathbb{Y} = \mathbb{R}$
- (вещественные числа — числа с любой дробной частью)
- Пример: предсказание роста по весу



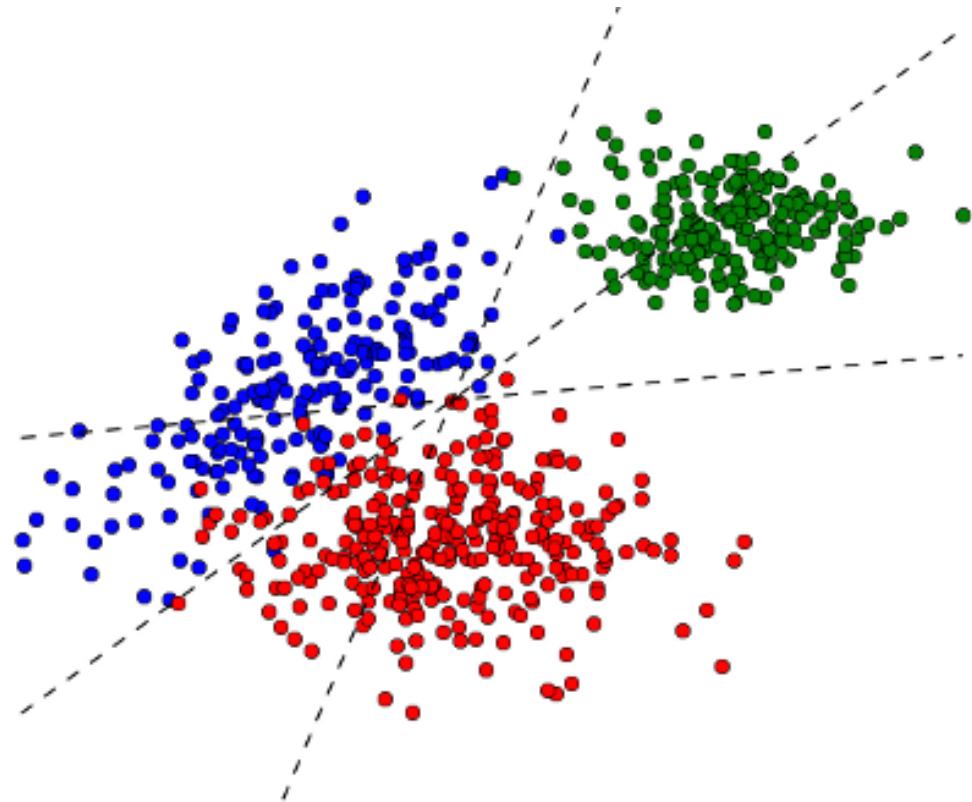
Классификация

- Конечное число ответов: $|\mathbb{Y}| < \infty$
- Бинарная классификация: $\mathbb{Y} = \{-1, +1\}$



Классификация

- Многоклассовая классификация: $\mathbb{Y} = \{1, 2, \dots, K\}$



Классификация

- Классификация с пересекающимися классами: $\mathbb{Y} = \{0, 1\}^K$
 - (multi-label classification)
- Ответ — набор из K нулей и единиц
- i -й элемент ответа — принадлежит ли объект i -му классу
- Какие темы присутствуют в статье?
- (математика, биология, экономика)

Ранжирование

- Набор документов d_1, \dots, d_n
- Запрос q
- Задача: отсортировать документы по *релевантности* запросу
- $a(q, d)$ — оценка релевантности

Ранжирование

Яндекс

картинки с котиками — 5 млн ответов



Найти

Поиск

[Картинки с кошками | Fun Cats — Забавные коты](#)

[funcats.by > pictures/](#) ▾

Картинки с кошками. Прикольные коты. 777 изображений. ... 32 изображения. Кошки

Стамбула. 41 изображение. Веселые котята.

Картинки

Видео

[Уморные котики \(57 фото\) » Бяки.нет | Картинки](#)

[byaki.net > Картинки > 14026-umornye-kotiki-57...](#) ▾

Бяки нет! . NET. Уморные котики (57 фото). 223. Коментариев:9Автор:4ertonok

Просмотров:161 395 Картинки28-10-2008, 00:03.

Карты

Маркет

Ещё

[Смешные картинки кошек с надписями | Лолкот.Ру](#)

[lolkot.ru](#) ▾

Смешные картинки для новых приколов! Сделать свой прикол очень просто. ... Котик

верит в чудеса. Он в носке подарок ищет...

[Красивые картинки и фото кошек, котят и котов](#)

[foto-zverey.ru > Кошки](#) ▾

Фото и картинки кошек и котят потрясающей красоты и нежности. Здесь мы собрали

такие изображения, которые всегда вызывают море положительных эмоций...

[Обои для рабочего стола Котята | картинки на стол Котята](#)

[7fon.ru > Чёрные обои и картинки > Обои котята](#) ▾

Картинки Котята с 1 по 15. Обои для рабочего стола Котята. ... Скачать Картинки Котята на рабочий стол бесплатно.

Прогнозирование временных рядов

- Позже — на примере

Построение рекомендательных систем

- Позже — на примере

Кластеризация

- \mathbb{Y} — отсутствует
- Нужно найти группы похожих объектов
- Сколько таких групп?
- Как измерить качество?
- Пример: сегментация пользователей мобильного оператора

Типы признаков

Типы признаков

- f_j — j -й признак
- D_j — множество значений признака

Бинарные признаки

- $D_j = \{0, 1\}$
- Доход клиента выше среднего по городу?
- Цвет фрукта — зеленый?

Вещественные признаки

- $D_j = \mathbb{R}$
- Возраст
- Площадь квартиры
- Количество звонков в колл-центр

Категориальные признаки

- D_j — неупорядоченное множество
- Цвет глаз
- Город
- Образование (может быть упорядоченным)

- Очень трудны в обращении

Порядковые признаки

- D_j — упорядоченное множество
- Воинское звание
- Роль в фильме (первого плана, второго плана, массовка)
- Тип населенного пункта

Множествозначные признаки

- (set-valued)
- D_j — множество всех подмножеств некоторого множества
- Какие фильмы посмотрел пользователь?
- Какие слова входят в текст?

Задачи анализа данных

Медицинская диагностика

- Объект — пациент в определенный момент времени
- Ответ — диагноз
- Классификация с пересекающимися классами

Медицинская диагностика — признаки

- Бинарные: пол, головная боль, слабость, и т.д.
- Порядковые: тяжесть состояния, желтушность, и т.д.
- Вещественные: возраст, пульс, артериальное давление, содержание гемоглобина в крови, доза препарата, и т.д.

Медицинская диагностика — особенности

- Много пропусков в данных (missing data)
- Недостаточный объем данных
- Алгоритм должен быть интерпретируемым
- Нужна оценка вероятности для каждого заболевания

Кредитный скоринг

- Объект — заявка на выдачу кредита банком
- Ответ — вернет ли клиент кредит
- Бинарная классификация

Кредитный скоринг — признаки

- Бинарные: пол, наличие телефона, и т.д.
- Категориальные: место жительства, профессия, семейный статус, работодатель, и т.д.
- Порядковые: образование, должность, и т.д.
- Вещественные: возраст, зарплата, стаж работы, доход семьи, сумма кредита, и т.д.

Кредитный скоринг — особенности

- Нужно оценивать вероятность (и связанные с этим риски) невозврата кредита

Предсказание оттока клиентов

- Объект — абонент в определенный момент времени
- Ответ — уйдет или не уйдет в следующем месяце
- Бинарная классификация

Предсказание оттока клиентов — признаки

- Бинарные: корпоративный клиент, подключенные услуги, и т.д.
- Категориальные: регион проживания, тарифный план, и т.д.
- Вещественные: длительность разговоров, количество СМС, частота оплаты, объем трафика, и т.д.

Предсказание оттока клиентов — особенности

- Нужно оценивать вероятность ухода
- Сверхбольшие выборки
- Исходные данные — сырье логи

Стоимость недвижимости

- Объект — квартира в Москве
- Ответ — стоимость в рублях
- Регрессия

Стоимость недвижимости — признаки

- Бинарные: наличие балкона, мусоропровода, лифта, охраны, парковки, и т.д.
- Категориальные: район города, тип дома (кирпичный/блочный/панельный/монолит), ближайшая станция метро и т.д.
- Вещественные: число комнат, жилая площадь, расстояние до центра, до метро, возраст дома, и т.д.

Стоимость недвижимости — особенности

- Выборка неоднородная, меняется со временем
- Разнотипные признаки
- Нужна интерпретируемая модель

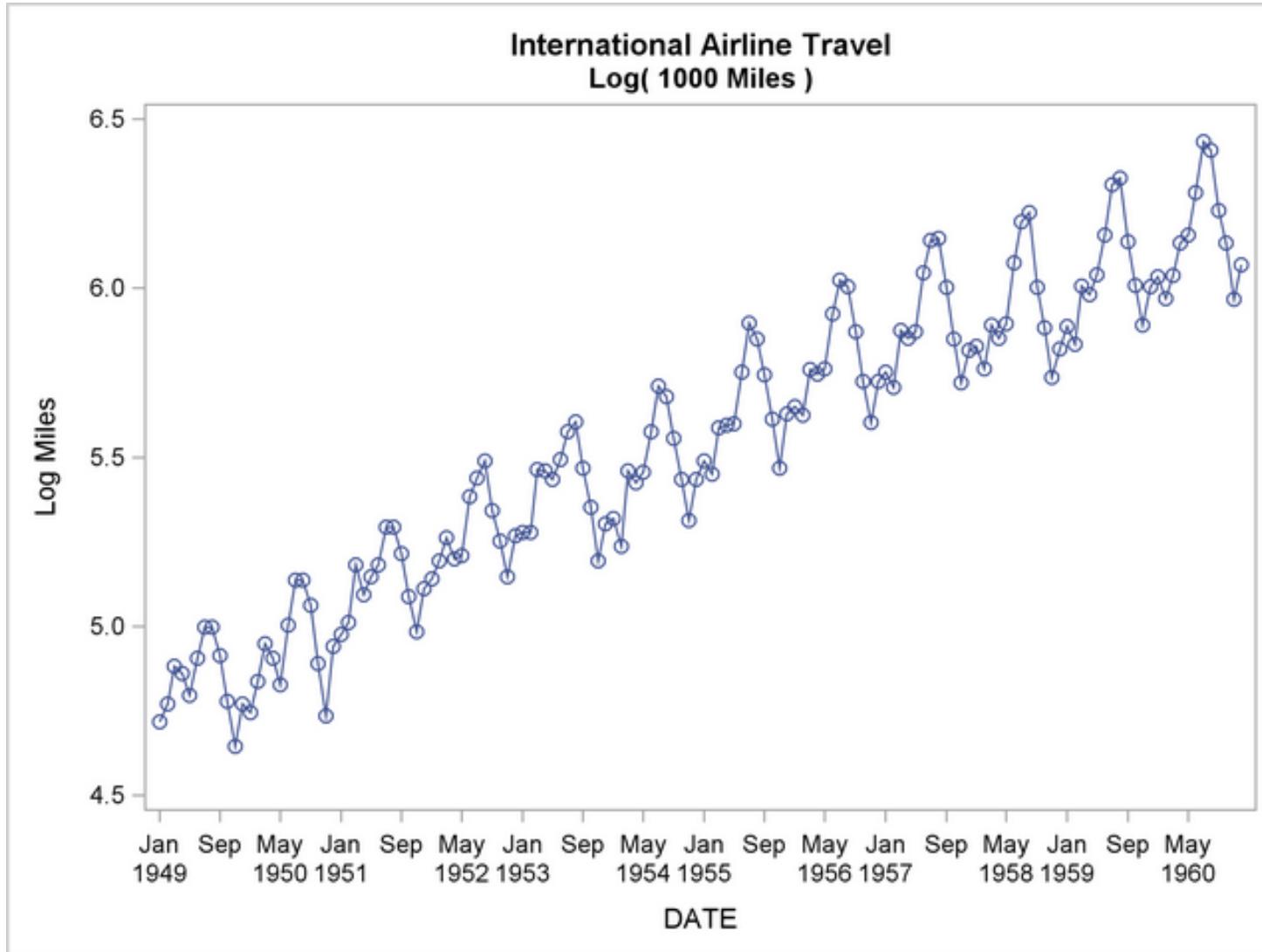
Прогнозирование продаж

- Объект — тройка (товар, магазин, день)
- Ответ — объем продаж
- Регрессия
- Прогнозирование временных рядов

Прогнозирование продаж — признаки

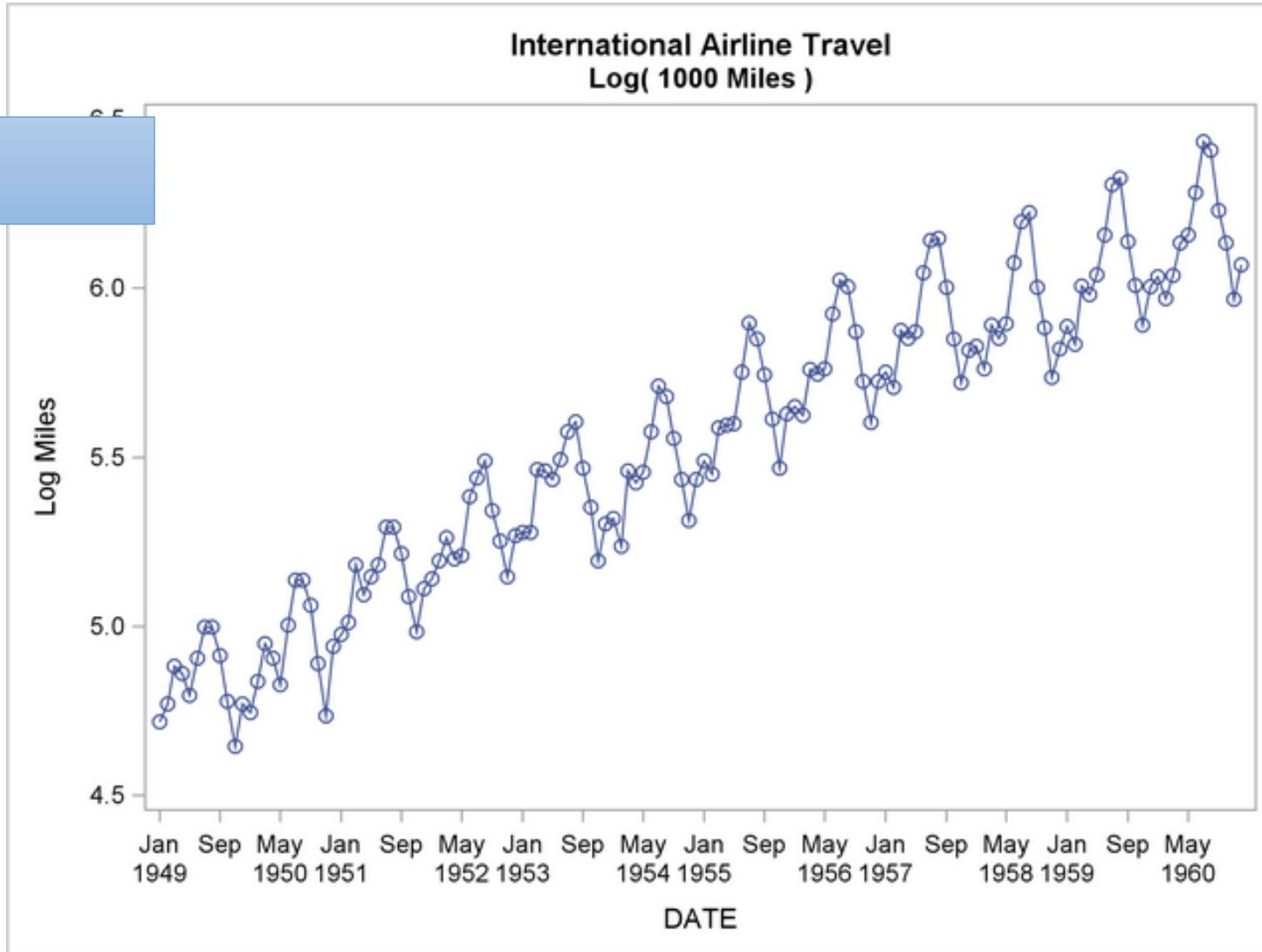
- Бинарные: выходной день, праздник, промоакция, и т.д.
- Вещественные: продажи в прошлые дни

Временные ряды



Временные ряды

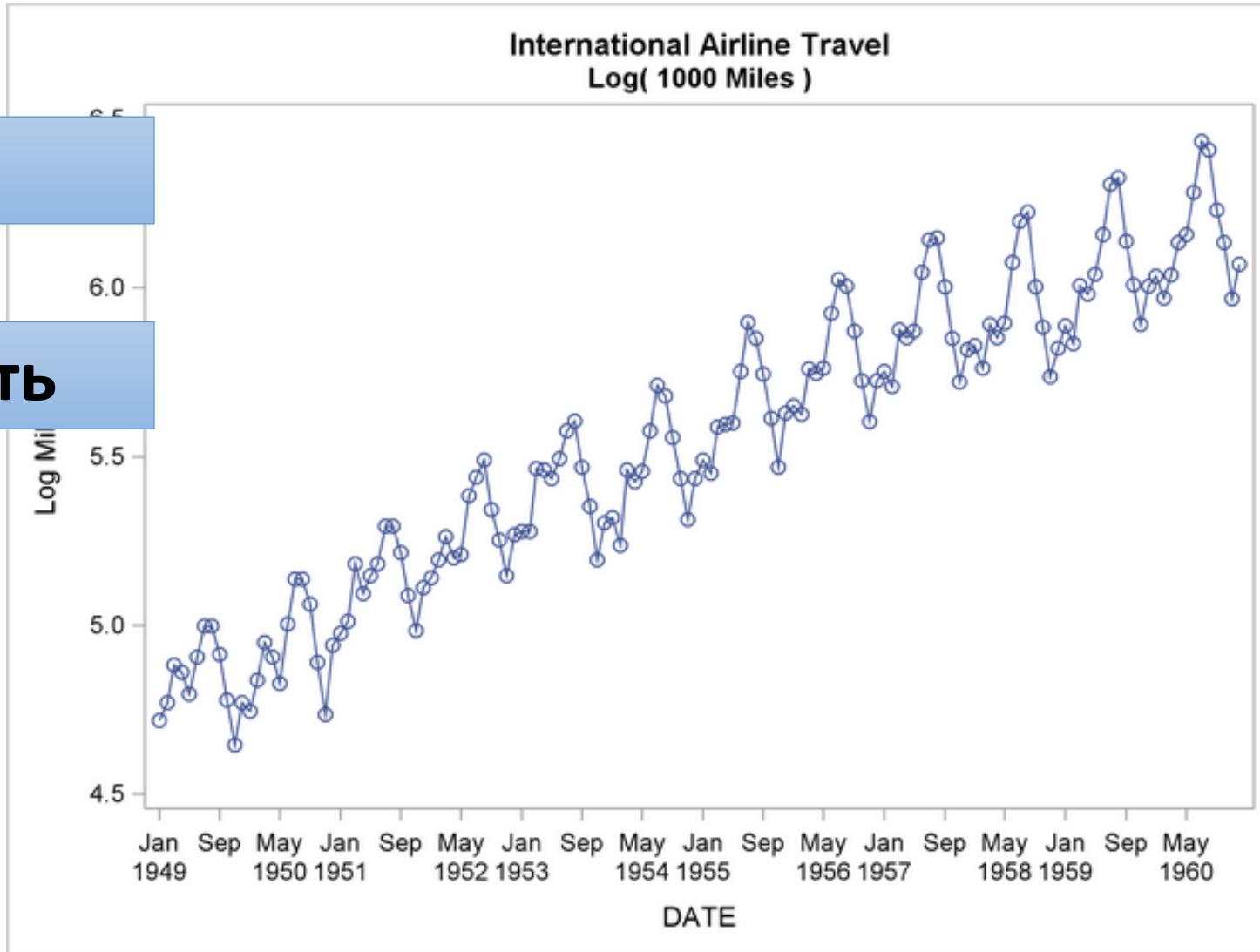
Тренд



Временные ряды

Тренд

Сезонность



Рекомендательная система фильмов

- Объект — пара (пользователь, фильм)
- Ответ — понравится ли пользователю фильм?
- Регрессия? Классификация?

Рекомендательная система — признаки

- Оценки фильмов от пользователей
- Возможно, профиль пользователя
- Возможно, информация о фильме

Рекомендательная система — Amazon

Frequently Bought Together



Price For All Three: \$86.01

Add all three to Cart Add all three to Wish List

Show availability and shipping details

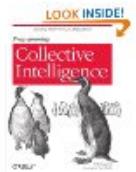
This item: Machine Learning for Hackers by Drew Conway Paperback \$33.87

Machine Learning in Action by Peter Harrington Paperback \$25.75

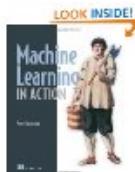
Programming Collective Intelligence: Building Smart Web 2.0 Applications by Toby Segaran Paperback \$26.39

Customers Who Bought This Item Also Bought

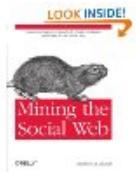
Page 1 of 17



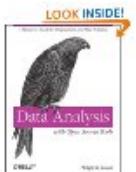
Programming Collective
Intelligence: Building ...
▶ Toby Segaran
 (84)
Paperback
\$26.39



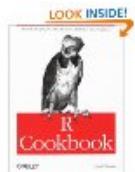
Machine Learning in Action
▶ Peter Harrington
 (10)
Paperback
\$25.75



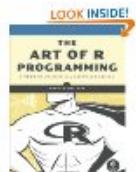
Mining the Social Web:
Analyzing Data from ...
▶ Matthew A. Russell
 (19)
Paperback
\$26.36



Data Analysis with Open
Source Tools
▶ Philipp K. Janert
 (29)
Paperback
\$24.05



R Cookbook (O'Reilly
Cookbooks)
▶ Paul Teator
 (18)
Paperback
\$32.43



The Art of R Programming: A
Tour of Statistical ...
Norman Matloff
 (29)
Paperback
\$25.06

Are any of these items inappropriate for this page? [Let us know](#)

Рекомендательная система — особенности

- Много метрик для оптимизации: число кликов по рекомендациям, число новых для пользователя товаров, разнообразие предлагаемых жанров, и т.д.
- Особый вид данных: (пользователь, фильм/товар, оценка)
- Получение оценки — явный и неявный отклик

Интересные предложения

РЕКОМЕНДУЕМ



Кулеры и системы охлаждения д...
[Deepcool NEPTWIN V2](#)

от 2 548 ₽

ХИТ ПРОДАЖ



Защитные пленки и стекла для т...
[Защитное стекло Xiaomi](#)

от 149 ₽

ТОВАР НЕДЕЛИ · **ДО -20%**



Электрообогреватели и тепловы...
[Ballu BOH/CL-07](#)

от 1 630 ₽

СКИДКИ **ДО -10%**



Цифровые бытовые метеостанции
[Oregon Scientific LW301](#)

от 8 990 ₽

РЕКОМЕНДУЕМ



Жесткие диски, SSD и сетевые н...
[Samsung MZ-7KE512BW](#)

Территория Grohe



ХИТ ПРОДАЖ



Дрели, шуруповерты, гайковерты
[Интерскол DA-12EP-01 ...](#)

РЕКОМЕНДУЕМ



Жесткие диски, SSD и сетевые н...
[Samsung MZ-7KE1T0BW](#)

Похожие товары



Canon EOS 200D Kit

от 32 390 ₽

1 отзыв 109 предложений

Любительская зеркальная
фотокамера

Байонет Canon EF/EF-S

Объектив в комплекте, модель
уточняйте у продавца

Матрица 25.8 МП (APS-C)

Цвет:

Цены 109



4.0

Canon EOS 750D Kit

от 32 650 ₽

5 отзывов 133 предложения

Любительская зеркальная
фотокамера

Байонет Canon EF/EF-S

Объектив в комплекте, модель
уточняйте у продавца

Матрица 24.7 МП (APS-C)

Цвет:

Цены 133



Canon EOS M10 Kit

от 21 250 ₽

4 отзыва 89 предложений

Фотокамера с поддержкой
сменных объективов

Байонет Canon EF-M

Объектив в комплекте, модель
уточняйте у продавца

Матрица 18.5 МП (APS-C)

Цвет:

Цены 89



до -30%



Canon EOS M6 Kit

от 38 300 ₽

до 84 940 ₽

2 отзыва 105 предложений

Фотокамера с поддержкой
сменных объективов

Байонет Canon EF-M

Объектив в комплекте, модель
уточняйте у продавца

Матрица 25.8 МП (APS-C)

Цвет:

Цены 105



С этим товаром часто покупают



4.0

HIPER MP15000

Универсальные внешние аккумулято...

от 1 453 ₽

30 отзывов 33 предложения

Цвет: ●

[Цены 33](#)[Цены 66](#)[Цены 16](#)[Цены 4](#)

2.5

Apple EarPods (Lightning)

Наушники и Bluetooth-гарнитуры

от 1 590 ₽

6 отзывов 66 предложений

Цвет: ○



Сетевая зарядка Apple

Зарядные устройства и адаптеры

от 642 ₽

16 предложений

Производитель: Apple

Тип: сетевая зарядка

Разъем подключения: for Apple
(Lightning)



Smart cover Apple

Чехлы для планшетов

от 3 080 ₽

4 предложения

Производитель: Apple

Совместимость: Apple

Тип: smart cover

Функция подставки: Да

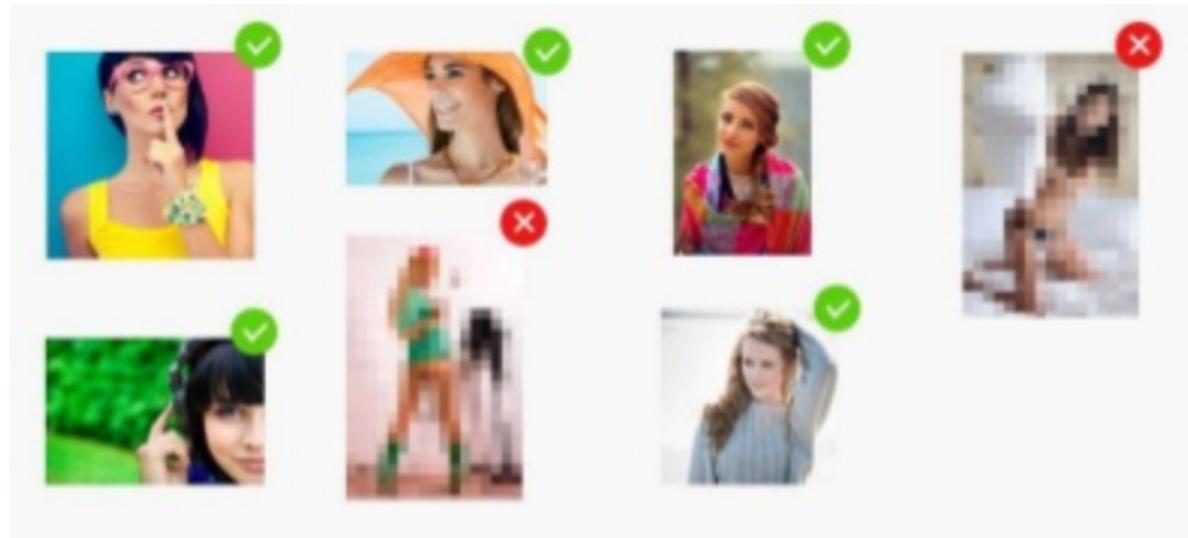
Цвет: ● ●

Рекомендательные системы

- Персональные предложения для каждого пользователя на основе большого числа факторов
- Учёт статистики по всем пользователям сервиса
- Новый взгляд на маркетинг
- Работа маркетолога — правильное использование рекомендательных механизмов, выбор каналов для взаимодействия ИИ и пользователей

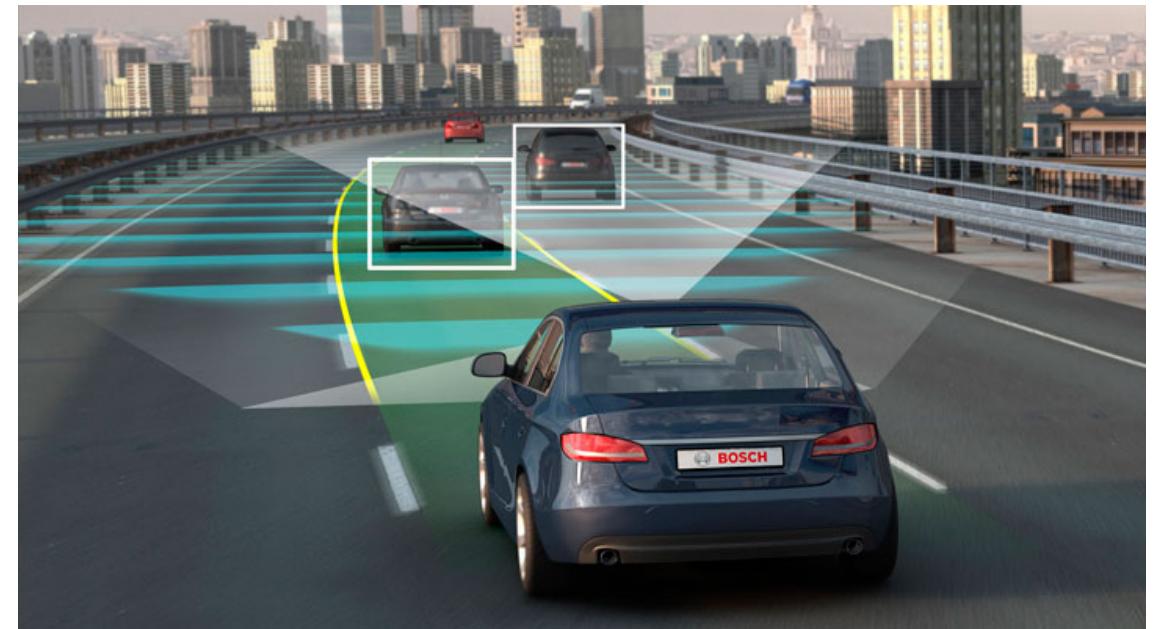
Автоматическая модерация

- Сервис онлайн-знакомств
- Фото в профиле не должно содержать эротику и не должно быть фотографией знаменитости
- Точность 88%
- Экономия порядка \$1,000,000 в год



Экономические вызовы

- Создание новых отраслей экономики
 - Личные помощники
 - Персонализированная медицина
 - Гибкая сфера услуг
- Отмирание ряда профессий
 - Дальнобойщики
 - Водители такси
 - Пилоты самолётов
 - Операторы колл-центров
- Замена человека в творческих видах деятельности



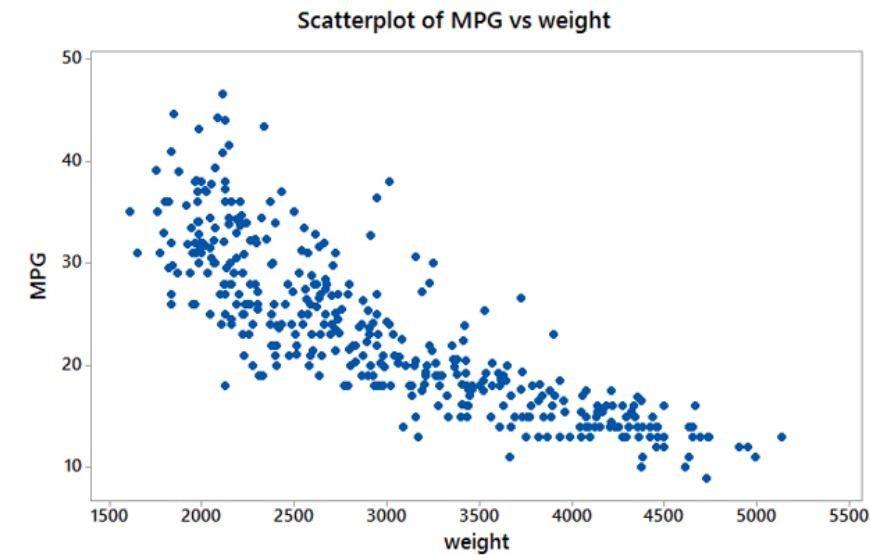
Векторы и матрицы

Вектор

- $x = (x^1, \dots, x^d)$ — признаковое описание
- x^1, \dots, x^d — вещественные числа
- x — набор из d чисел — **вектор**

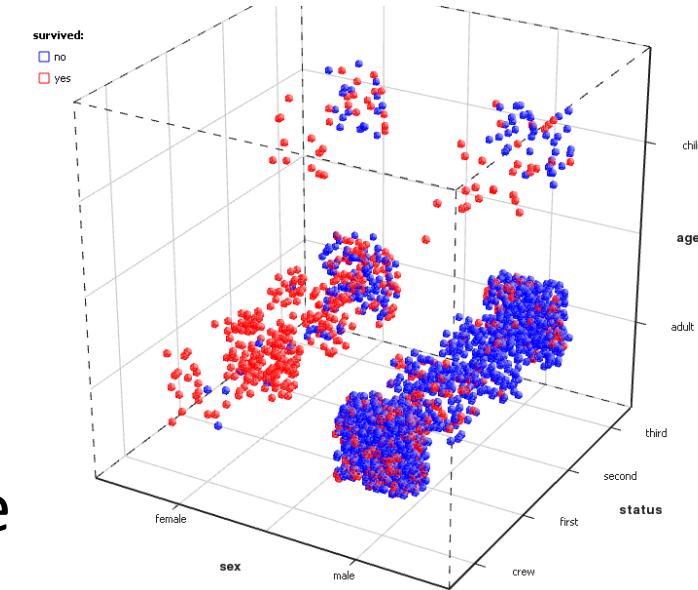
Вектор

- 5 — число
- $(5, 3)$ — точка на плоскости



Вектор

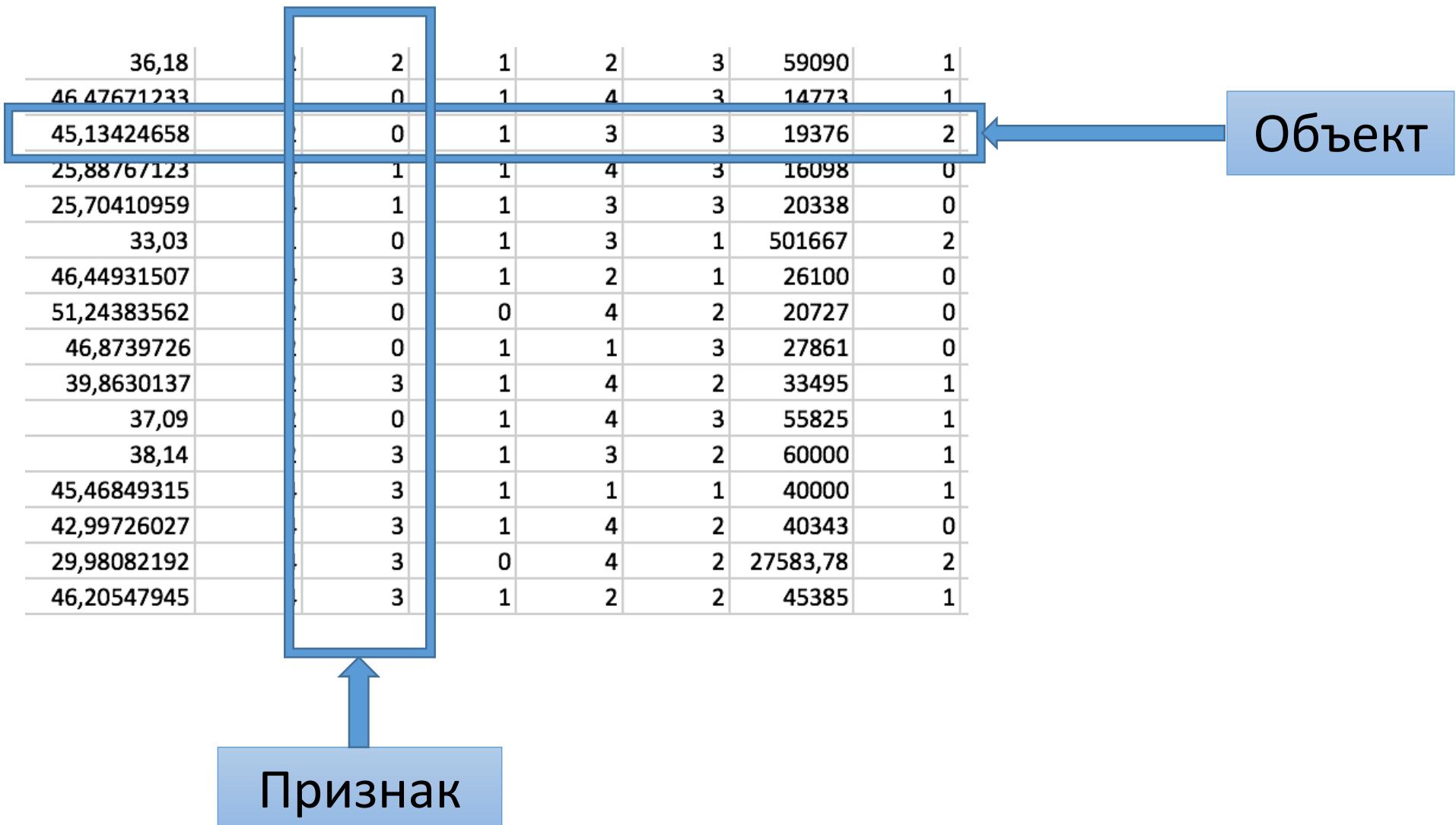
- 5 — число
- $(5, 3)$ — точка на плоскости
- $(5, 3, 9)$ — точка в пространстве
- $(5, 3, 9, 1)$ — точка в четырехмерном пространстве
- ...
- Пространство наборов из d вещественных чисел — евклидово пространство \mathbb{R}^d



Матрицы

- Вектор описывает один объект
- А если объектов несколько?

Матрицы



Матрицы

- Матрица — таблица с числами
- Пример:

$$A = \begin{pmatrix} 1 & 2 & 5 & 1 \\ 5 & 3 & 9 & 0 \\ 0 & 7 & 1 & 4 \end{pmatrix}$$

- Два индекса: строка и столбец
- $a_{11} = 1$
- $a_{23} = 9$

Матрицы

- Матрица — таблица с числами
- Пример:

$$A = \begin{pmatrix} 1 & 2 & 5 & 1 \\ 5 & 3 & 9 & 0 \\ 0 & 7 & 1 & 4 \end{pmatrix}$$

- Два индекса: строка и столбец
- $a_{11} = 1$
- $a_{23} = 9$
- Пространство матриц 3 на 4: $\mathbb{R}^{3 \times 4}$

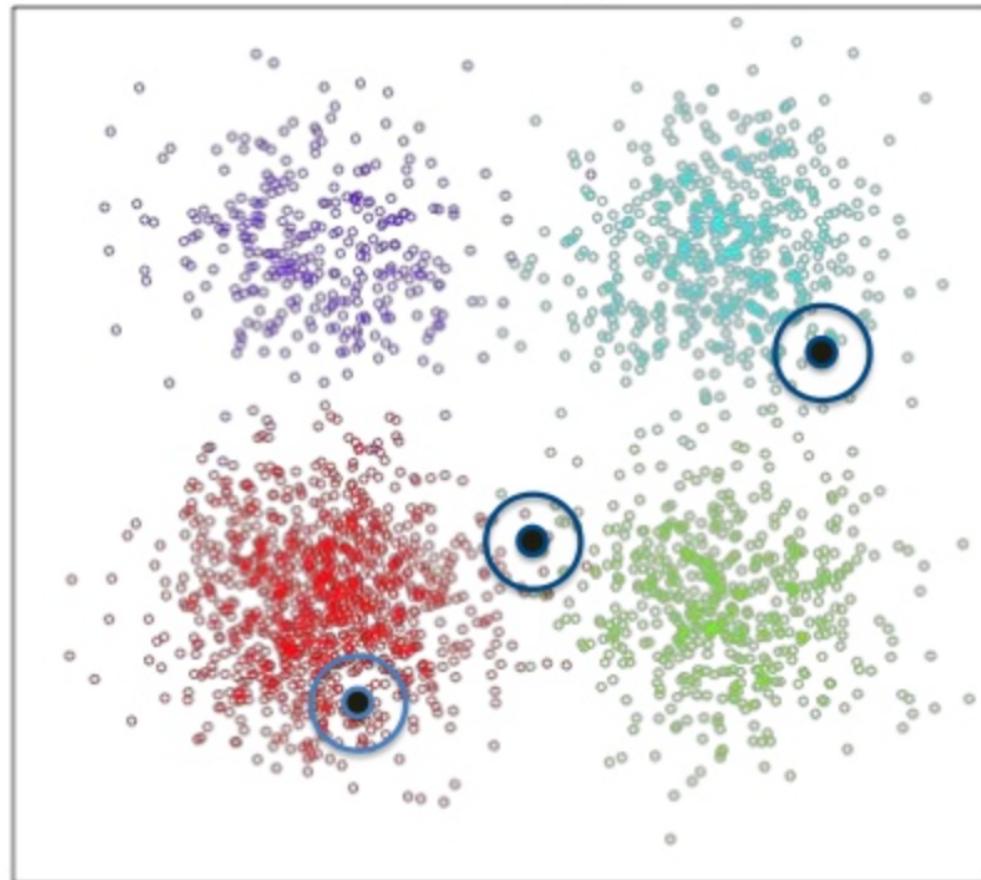
Матрицы

- Выборка объектов описывается матрицей «объекты-признаки»
- По строкам — объекты
- По столбцам — признаки

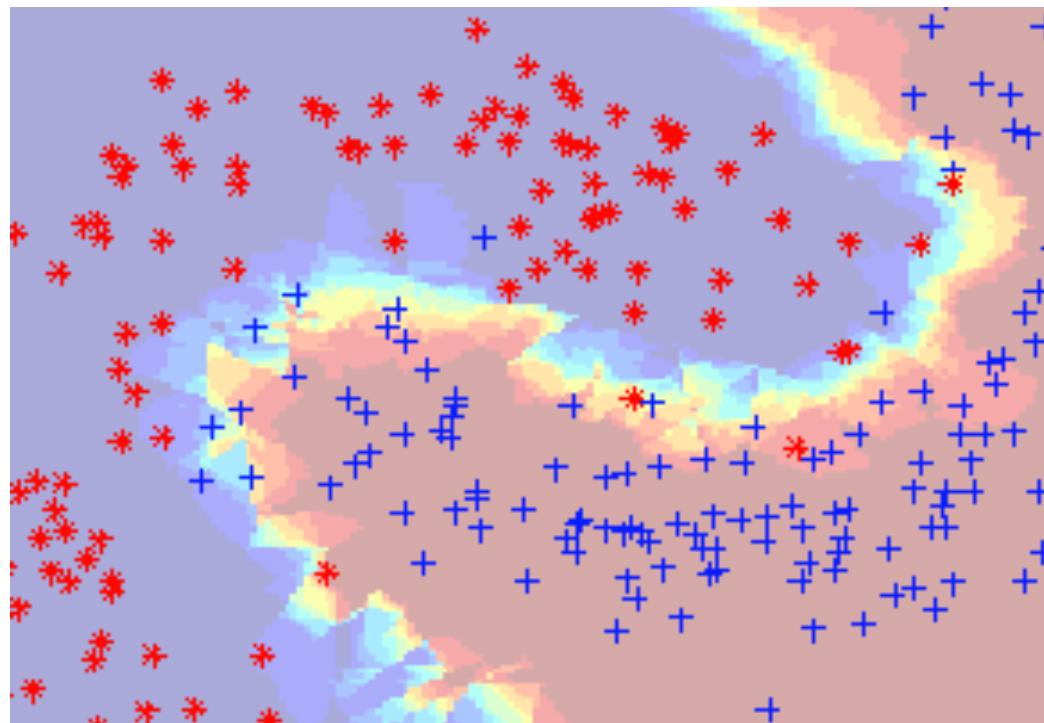
$$X = \begin{pmatrix} 1 & 1000 & 5 & 3 & 4 \\ 9 & 9000 & 10 & 5 & 7.5 \\ 5 & 5000 & 1 & 3 & 2 \end{pmatrix}$$

Метрические методы

Гипотеза компактности



Гипотеза компактности



Гипотеза компактности



Гипотеза компактности

- Для классификации: близкие объекты, как правило, лежат в одном классе
- Для регрессии: близким объектам соответствуют близкие ответы
- Что такое «близкие объекты»?

Измерение сходства

- Необходимо ввести расстояние между объектами
- В математике для этого используют понятие метрики
- $\rho(x, z)$ — функция расстояния (не обязательно метрика)
- Типичный пример: евклидова метрика

$$\rho(x, z) = \sqrt{\sum_{j=1}^d (x_j - z_j)^2}$$

Метод k ближайших соседей

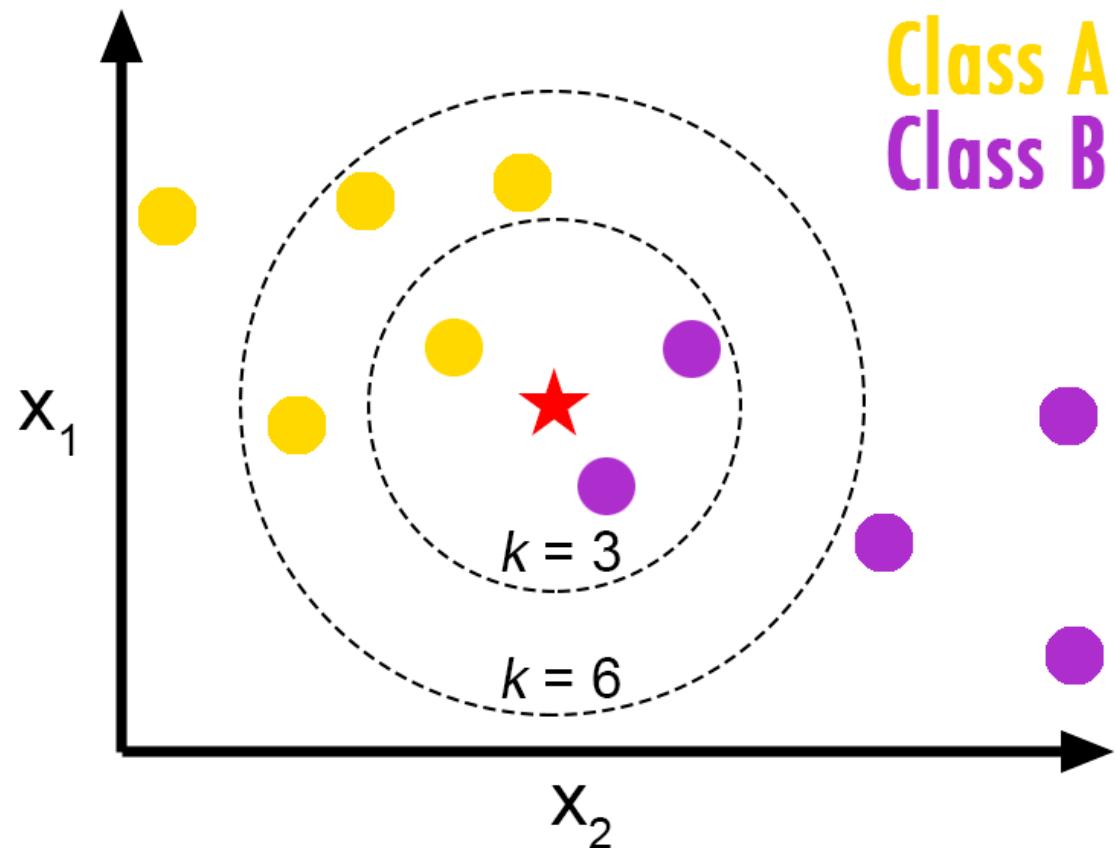
- k nearest neighbors (kNN)
- Задача классификации
- Дано: выборка $X = (x_i, y_i)_{i=1}^\ell$
- Этап обучения: запоминаем выборку X

Метод k ближайших соседей

- Новый объект x
- Сортируем объекты обучающей выборки по расстоянию до x :
$$\rho(x, x_{(1)}) \leq \dots \leq \rho(x, x_{(\ell)})$$
- Выбираем самый популярный класс среди k ближайших соседей:

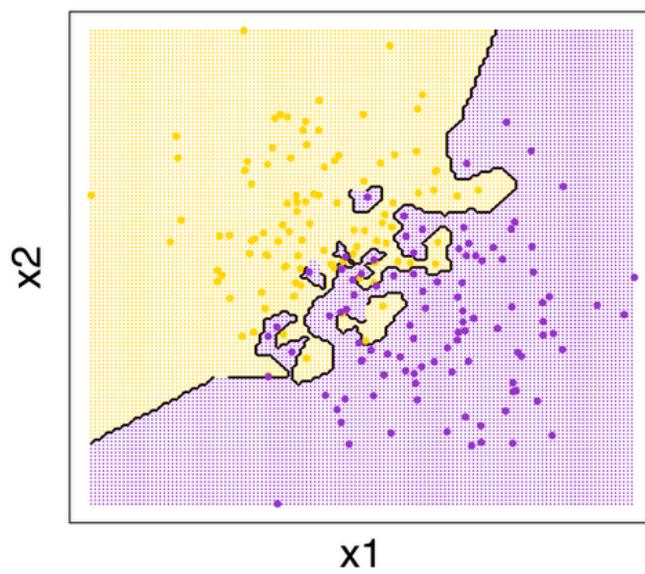
$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^{\textcolor{red}{k}} [y_{(i)} = y]$$

Метод k ближайших соседей

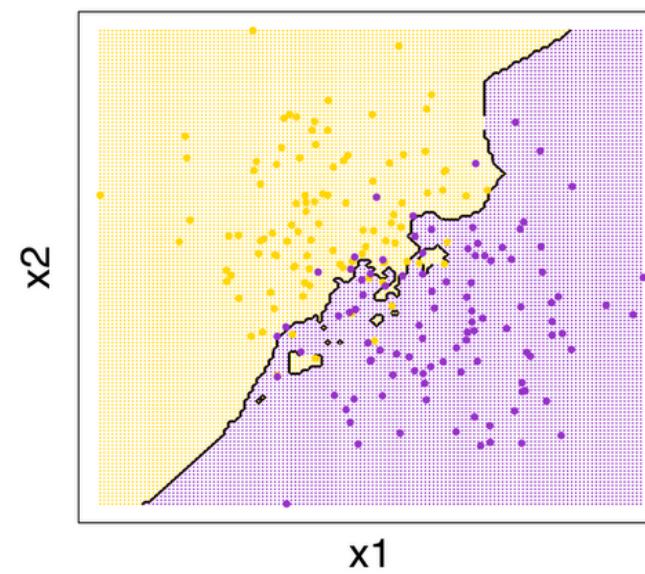


Выбор числа соседей

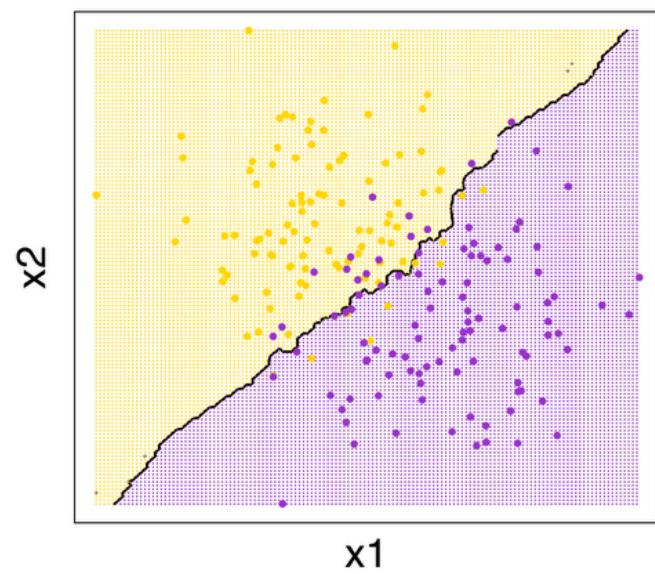
Binary kNN Classification (k=1)



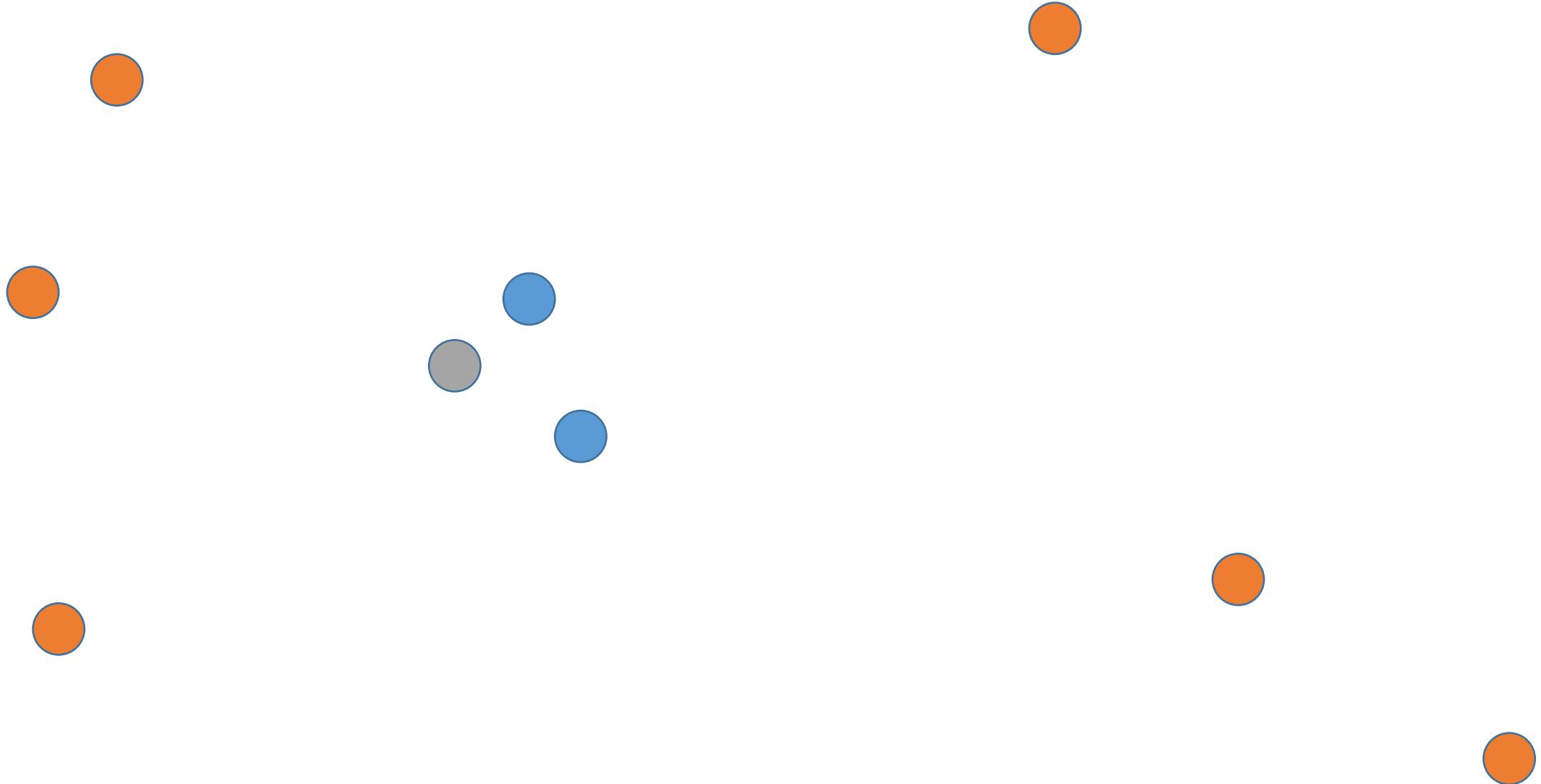
Binary kNN Classification (k=5)



Binary kNN Classification (k=25)



Проблема kNN



Проблема kNN

- Никак не учитываются расстояния до k ближайших соседей
- Более близкие соседи должны быть важнее

kNN с весами

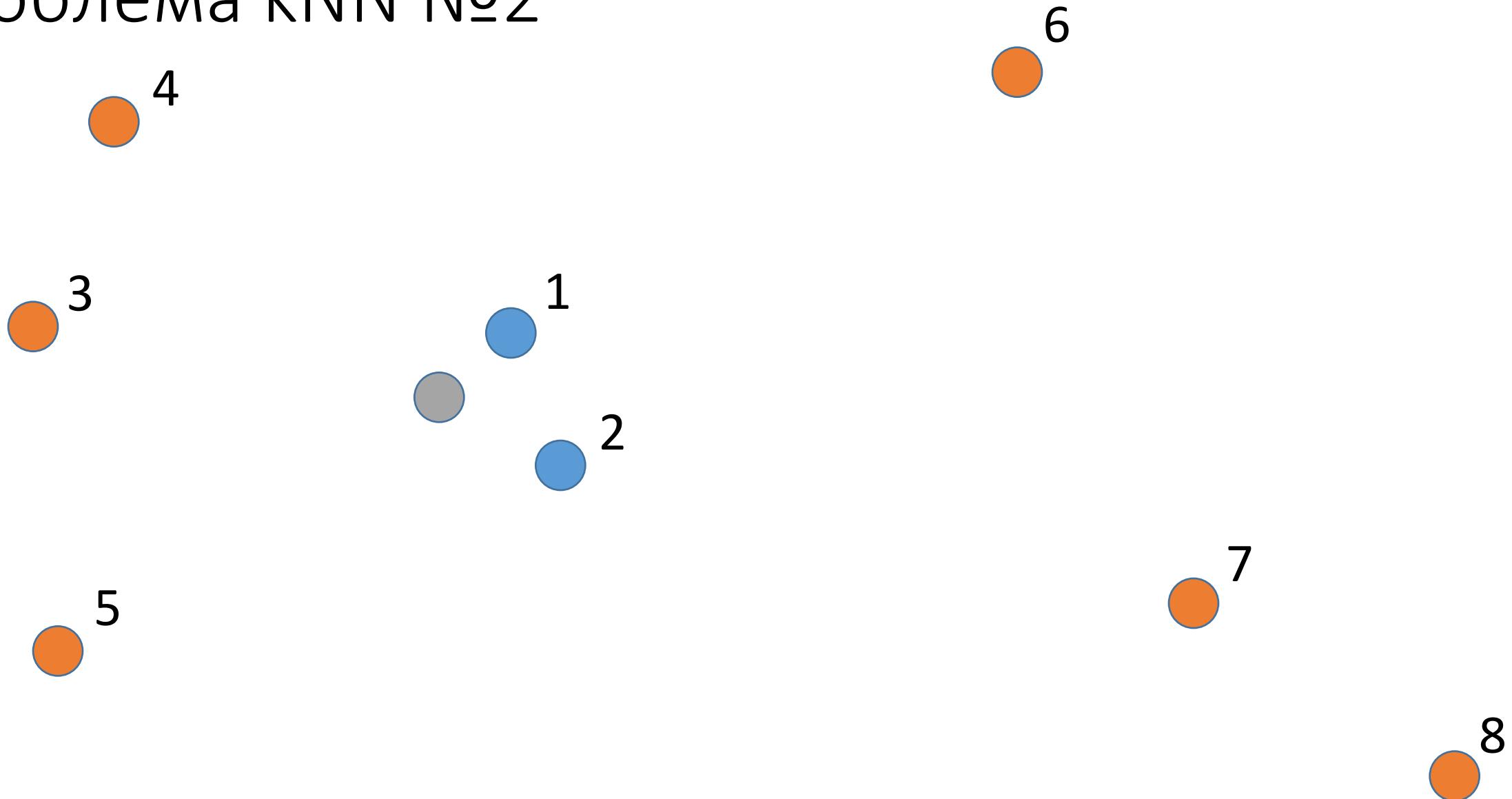
$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k w_i [y_{(i)} = y]$$

Варианты:

- $w_i = \frac{k+1-i}{k}$

- $w_i = q^i$

Проблема kNN №2



Проблема kNN №2

- Никак не учитываются сами расстояния
- w_i может учитывать расстояние до объекта, а не только его номер
- Подробнее — позже

Резюме

- Много типов признаков — у всех свои особенности
- Много постановок задач — у всех свои особенности
- Много особенностей у конкретных прикладных задач
- Данные обычно представляют из себя матрицу
- kNN
 - Компактность данных
 - Не нужно обучать
 - Как выбирать k?