

Машинное обучение

Лекция 2
Введение в машинное обучение
Метод k ближайших соседей

Андрей Нарцев
andrei.nartsev@gmail.com
anartsev@hse.ru

НИУ ВШЭ, 2025

План лекции

Введение в машинное обучение:

- Основные понятия (remind)
- Виды задач
- Типы признаков

Метод k ближайших соседей:

- Гипотеза компактности
- Метрики на пространстве признаков
- Оценка обобщающей способности и подбор гиперпараметров
- Взвешенный kNN

Напоминание

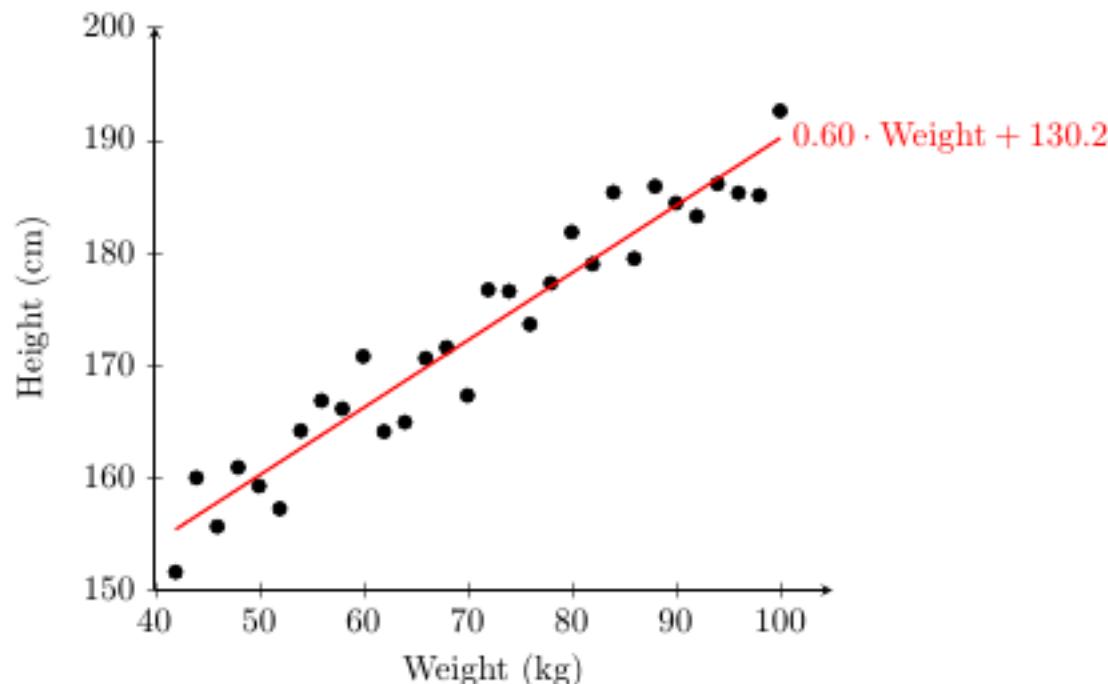
- \mathbb{X} — пространство объектов, \mathbb{Y} — пространство ответов
- $x = (x_1, \dots, x_d)$ — признаковое описание
- $X = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка
- $a(x)$ — алгоритм, модель
- $Q(a, X)$ — функционал ошибки алгоритма a на выборке X
- Обучение: $a(x) = \arg \min_{a \in \mathcal{A}} Q(a, X)$

Виды задач

Типы ответов

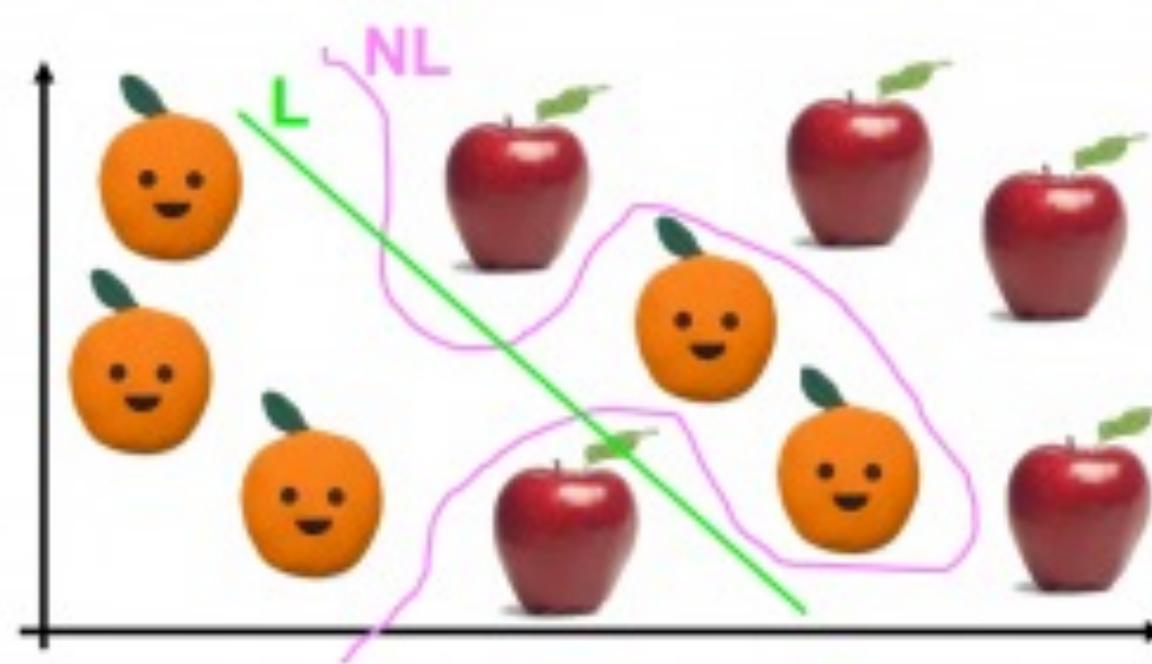
Регрессия

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



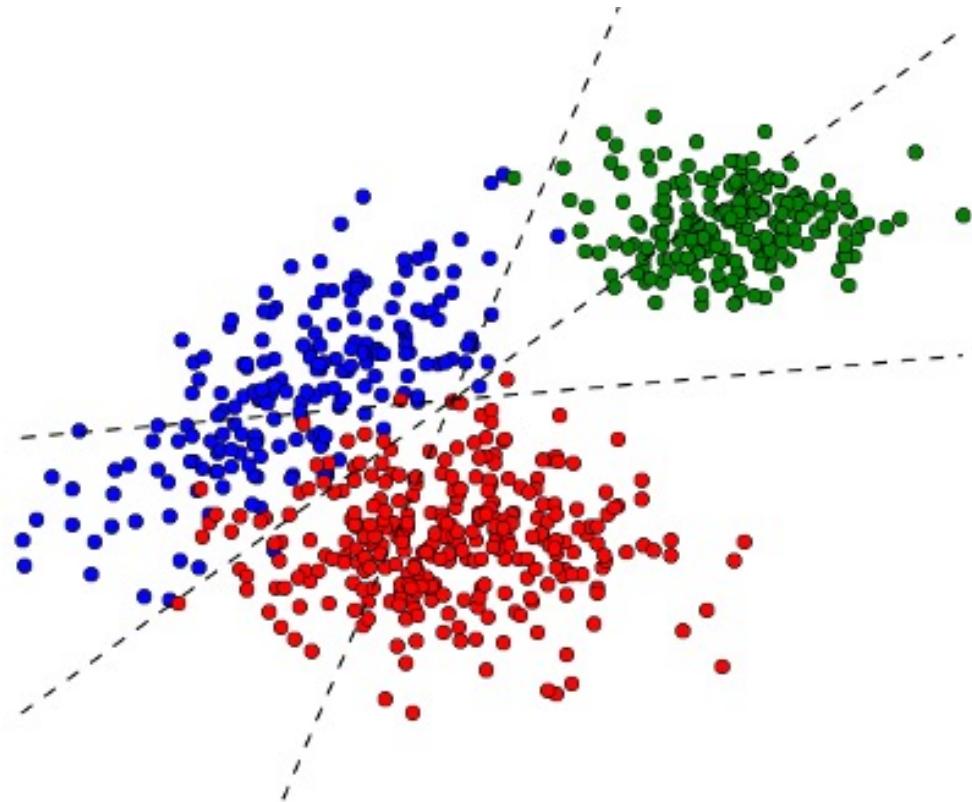
Классификация

- 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.

Кластеризация

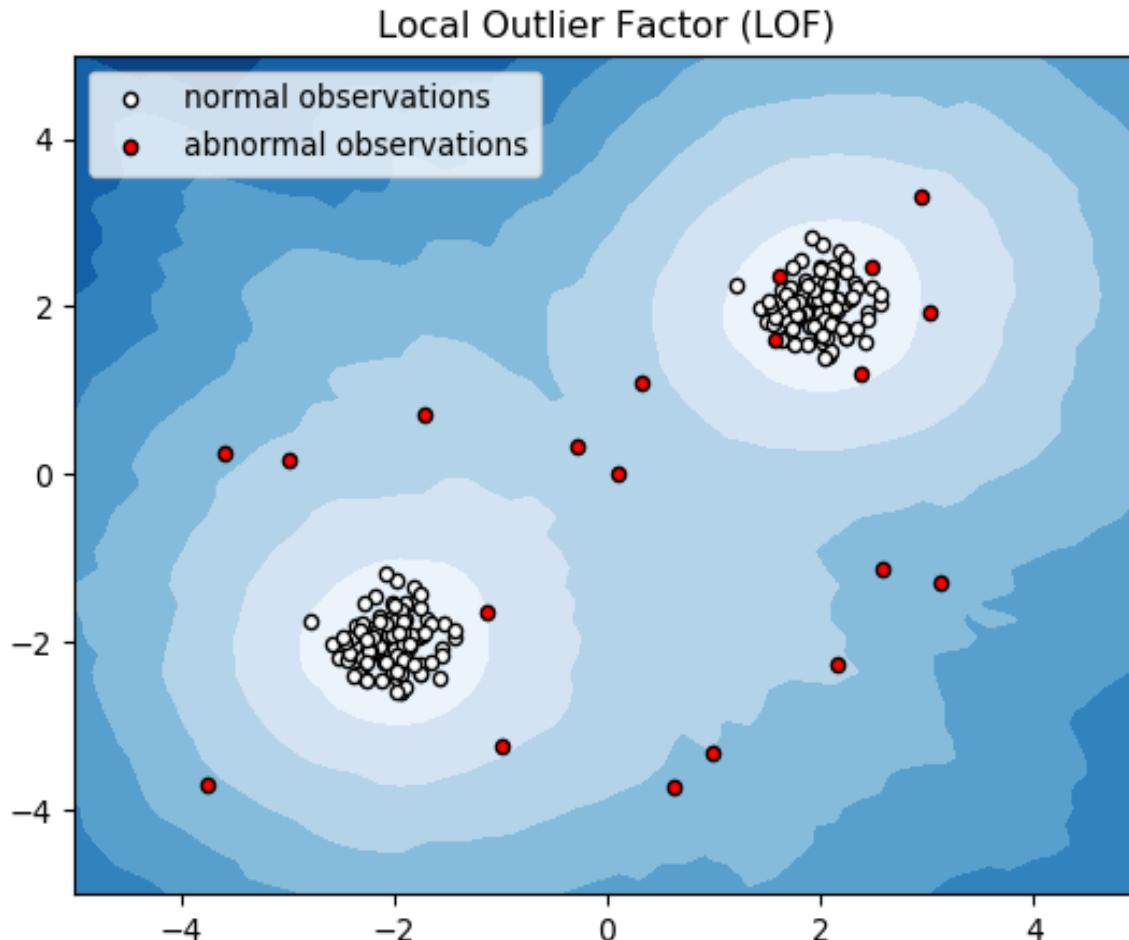
- \mathbb{Y} — отсутствует
- Нужно найти группы похожих объектов
- Сколько таких групп?
- Как измерить качество?
- Пример: сегментация пользователей мобильного оператора

ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация** – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности** – задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.
- **Оценивание плотности** – задача приближения распределения объектов.

ПРИМЕР ОЦЕНИВАНИЯ ПЛОТНОСТИ

- Поиск аномалий с помощью оценивания плотностей



ЗАДАЧИ, В КОТОРЫХ НЕТ ЦЕЛЕВОЙ ПЕРЕМЕННОЙ

- **Кластеризация** – задача разделения объектов на группы, при этом где целевые переменные для объектов неизвестны (или не существуют). Разделение происходит только на основе признаковых описаний объектов.
- **Понижение размерности** – задача генерации новых признаков (их число меньше, чем число старых), так, что с их помощью задача решается не хуже, чем с исходными.
- **Оценивание плотности** – задача приближения распределения объектов.
- **Визуализация** – задача изображения многомерных объектов в 2х или 3хмерном пространстве с сохранением зависимостей между ними.

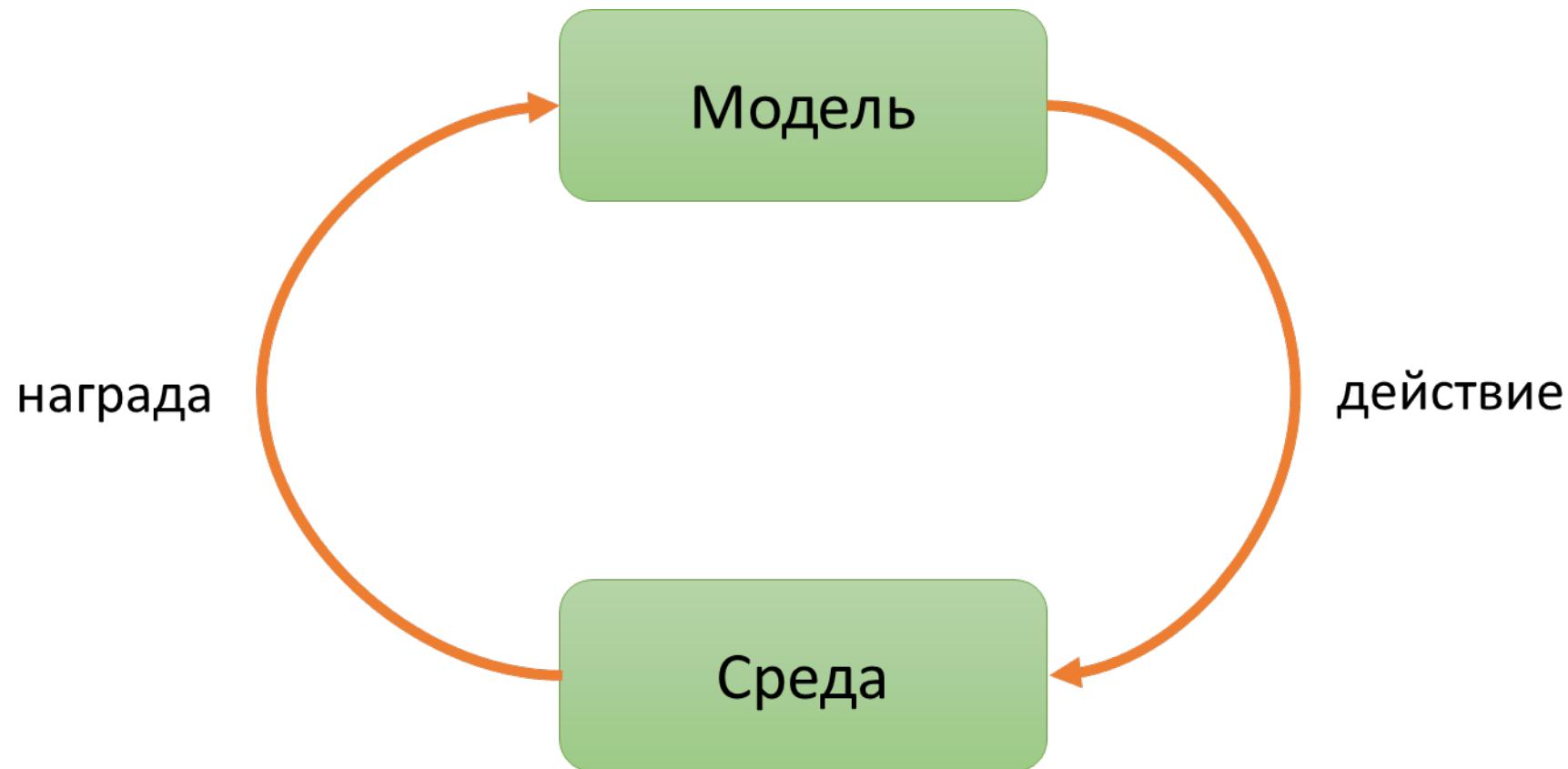
ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.

ТИПЫ ЗАДАЧ МАШИННОГО ОБУЧЕНИЯ

- Если нам известны значения целевой переменной, то есть алгоритм обучается так, чтобы правильно предсказывать целевую переменную – это **обучение с учителем**. Сюда относят классификацию, регрессию и ранжирование.
- Если нам неизвестны значения целевой переменной или целевая переменная вообще отсутствует, то есть алгоритм обучается только по признакам объектов, то это **обучение без учителя**. Примерами обучения с учителем являются кластеризация, понижение размерности и др.

Обучение с подкреплением



Типы признаков

Типы признаков

- 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

Типы признаков

- Бинарные
- Числовые
- Категориальные и порядковые
- Есть и более сложные: тексты, изображения, звук и т.д.

Гипотеза компактности и knn

Как отличить ель от сосны?



Как отличить ель от сосны?



Как отличить ель от сосны?



Ель:

- Ветки смотрят вверх
- Ствол не видно
- Густые иголки
- Цвет ближе к зелёному



Сосна:

- Ветки параллельны земле
- Ствол видно
- Иголки более редкие
- Цвет ближе к жёлтому

Как отличить ель от сосны?



Ветки вверх
Ствол не видно
Густые иголки
Цвет ближе к синему

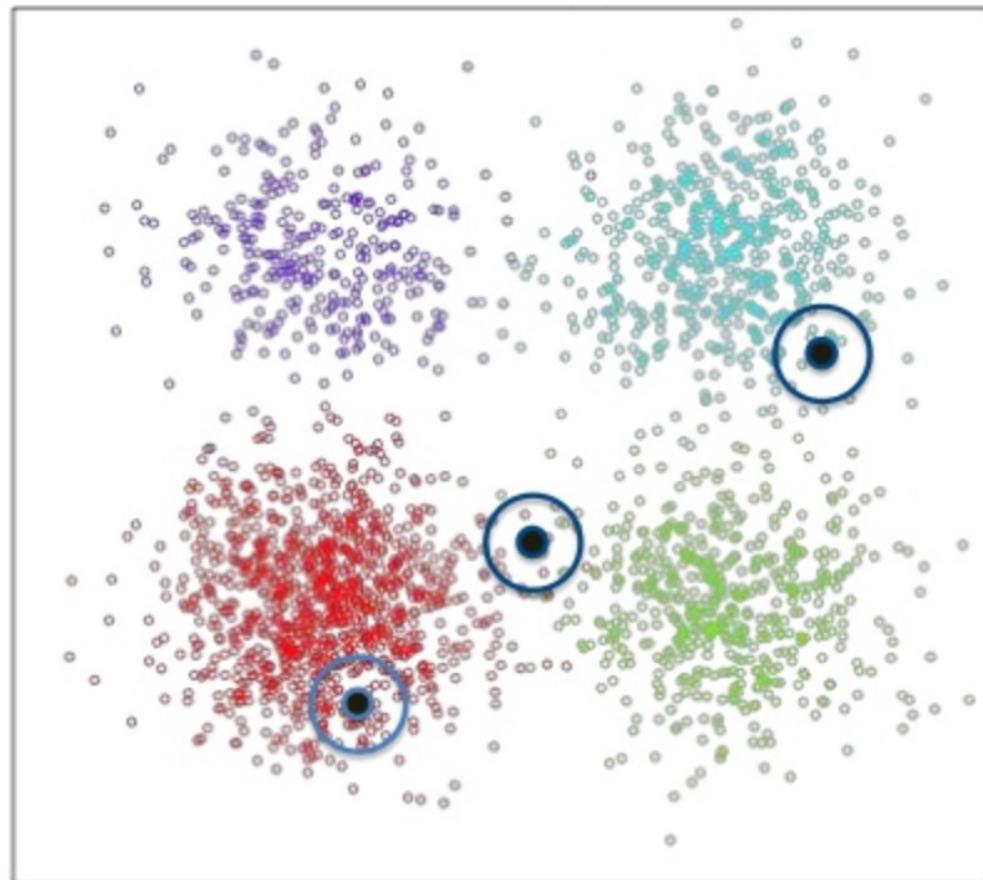


Скорее всего ель

Что такое обучение?

- Запоминаем примеры (объекты и ответы)
- Когда приходит новый объект, сравниваем с запомненными примерами
- Выдаём ответ от наиболее похожего примера

Гипотеза компактности



Гипотеза компактности

Если два объекта похожи друг на друга, то ответы на них
тоже похожи

kNN: обучение

- Дано: обучающая выборка $X = (x_i, y_i)_{i=1}^{\ell}$
- Задача классификация (ответы из множества $\mathbb{Y} = \{1, \dots, K\}$)
- Обучение модели:
 - Запоминаем обучающую выборку X

kNN: применение

Дано: новый объект x

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем k ближайших объектов: $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

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

kNN: применение

Дано: новый объект x

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем k ближайших объектов: $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

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

kNN: применение

Дано: новый объект x

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем k ближайших объектов: $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

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

kNN: применение

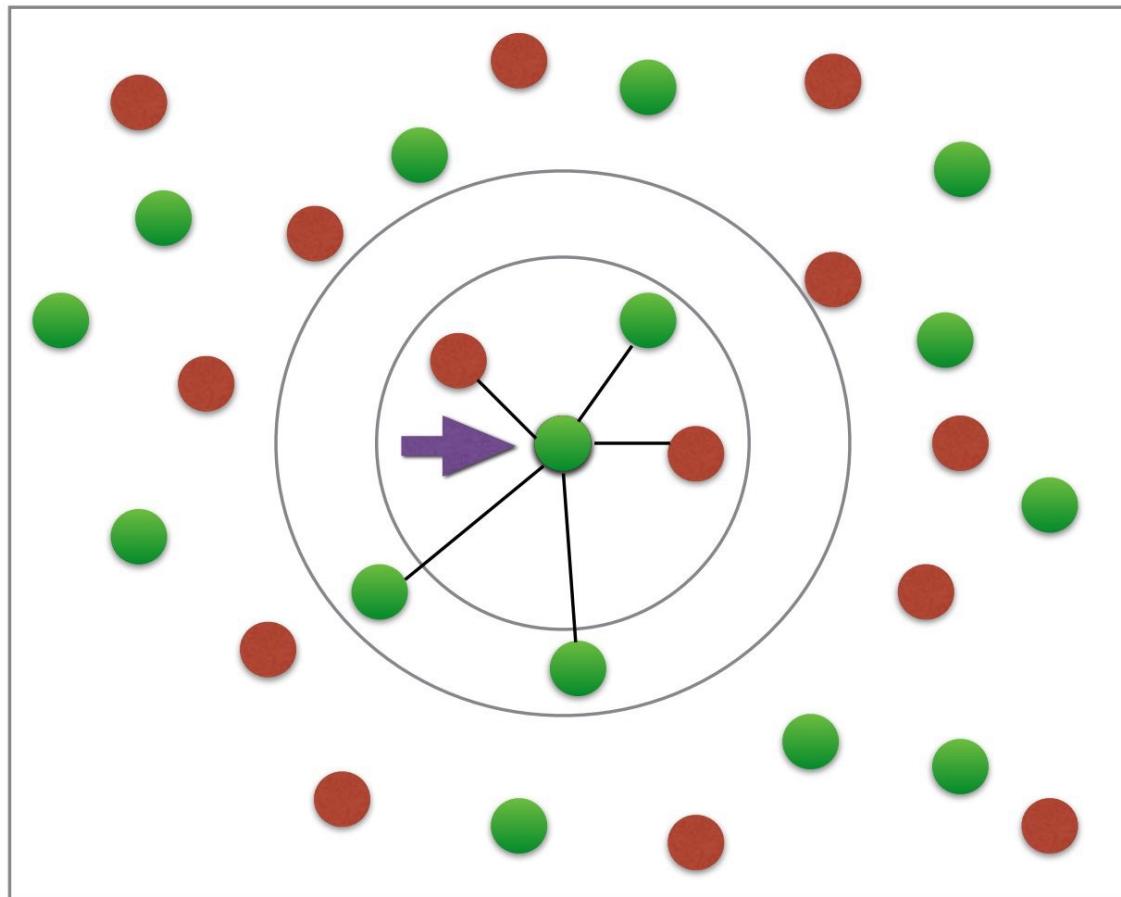
Дано: новый объект x

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем k ближайших объектов: $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

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

kNN: применение



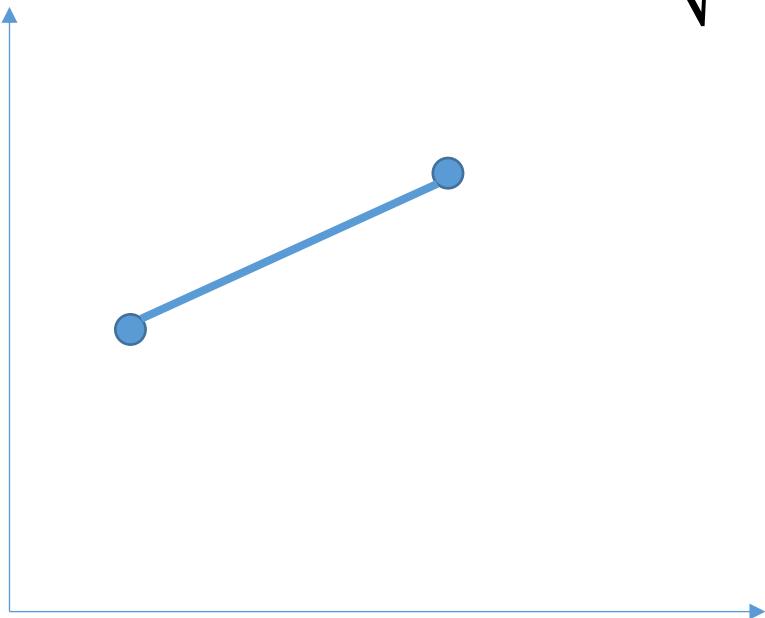
Сравнение объектов и метрики

Числовые данные

Сколько раз в день вызывает такси	Средние расходы на такси в день	Как часто вызывал комфорт	Возраст	Согласился повысить категорию?
2	400	0.3	29	да
0.3	80	0	28	нет
...

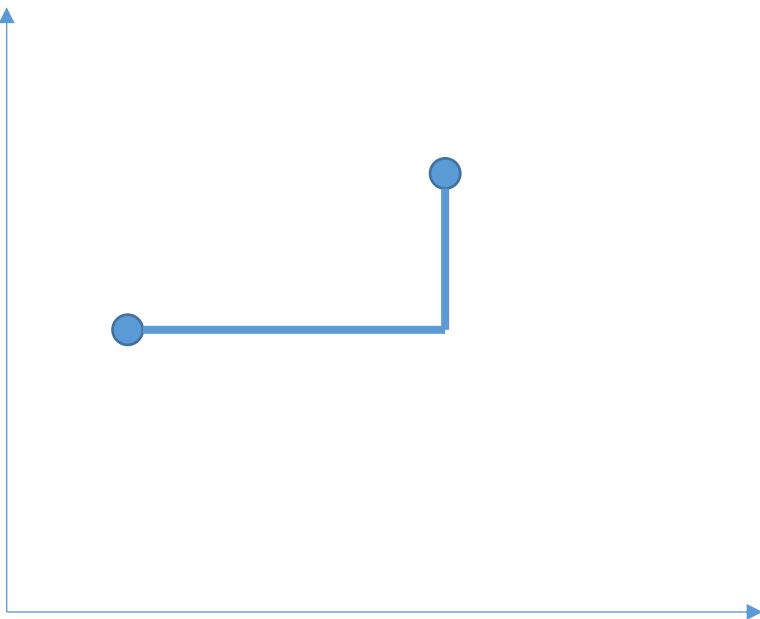
Евклидова метрика

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



Манхэттенская метрика

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



Обобщение

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

- Метрика Минковского
- Можно подбирать p под конкретную задачу

Категориальные данные

На каком классе чаще всего ездит	Ближайшее к дому метро	Способ оплаты	Согласился повысить категорию?
Эконом	Таганская	Карта	да
Комфорт	Юго-Западная	Наличные	нет
...

Считывающая метрика

- Простейшая метрика: подсчёт различий

$$\rho(x, z) = \sum_{j=1}^d [x_j \neq z_j]$$

Измерение ошибки модели

Вопросы

- Как сравнить две модели?
- Как подобрать k и метрику?

ФУНКЦИЯ ПОТЕРЬ ДЛЯ КЛАССИФИКАЦИИ

- Частый выбор — бинарная функция потерь

$$L(y, a) = [a \neq y]$$

- Функционал ошибки — доля ошибок (error rate)

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i]$$

- Нередко измеряют долю верных ответов (accuracy):

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

Accuracy

$a(x)$	y
-1	-1
+1	+1
-1	-1
+1	-1
+1	+1

Доля ошибок: 0.2

Доля верных ответов: 0.8

Accuracy

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i]$$

Решаем задачу выявления редкого заболевания

- 950 здоровых ($y = +1$)
- 50 больных ($y = -1$)

Модель: $a(x) = +1$

Доля ошибок: 0.05

Accuracy

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} [a(x_i) \neq y_i]$$

- Всегда смотрите на баланс классов!
- Доля верных ответов не обязательно меняется от 0.5 до 1 для разумных моделей

Как выбрать k?

Обучающая выборка

На каком классе чаще всего ездит	Ближайшее к дому метро	Способ оплаты	Согласился повысить категорию?
Эконом	Таганская	Карта	да
Комфорт	Юго-Западная	Наличные	нет
Комфорт	Строгино	Карта	да

Применяем модель:

Эконом	Таганская	Карта	?
--------	-----------	-------	---

Как выбрать k?

Обучающая выборка

На каком классе чаще всего ездит	Ближайшее к дому метро	Способ оплаты	Согласился повысить категорию?
Эконом	Таганская	Карта	да
Комфорт	Юго-Западная	Наличные	нет
Комфорт	Строгино	Карта	да

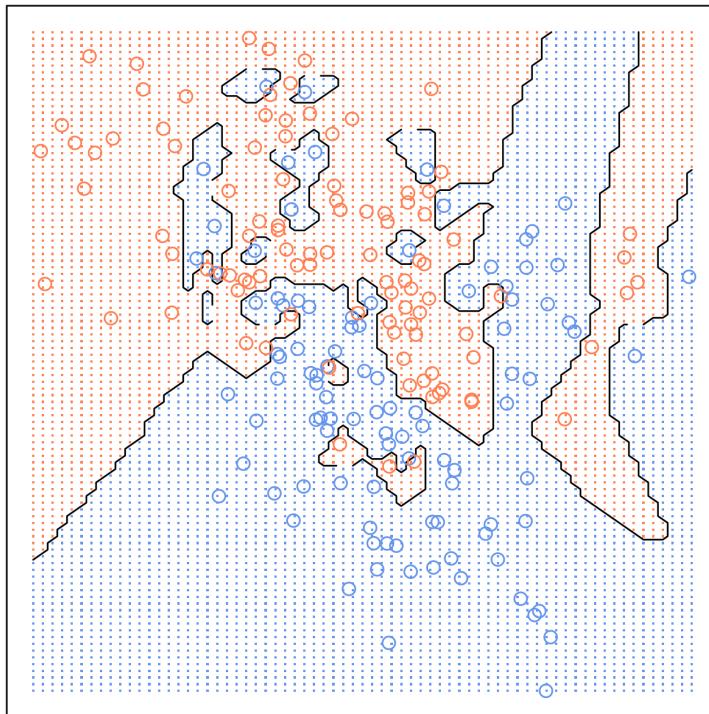
Применяем модель:

Эконом	Таганская	Карта	да
--------	-----------	-------	----

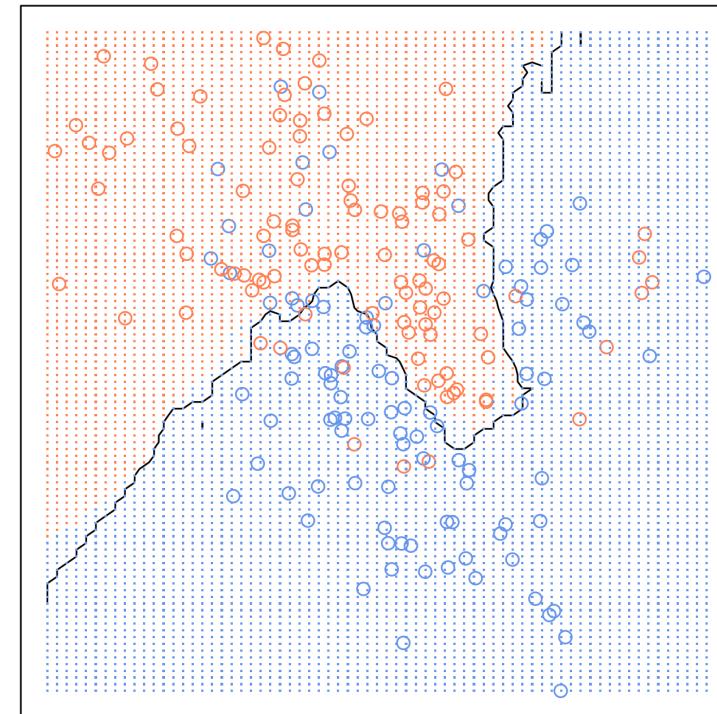
С точки зрения качества на обучающей выборке лучший выбор $k = 1$

Как выбрать k?

1-nearest neighbours



20-nearest neighbours



<https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/>

Гиперпараметры

- Нельзя подбирать k по обучающей выборке — **гиперпараметр**
- Нужно использовать дополнительные данные

Обобщающая способность

Обобщающая способность

Как готовиться к экзамену?

Заучить все примеры с занятий

Переобучение (overfitting)

Хорошее качество на обучении
Низкое качество на новых данных

Разобраться в предмете и усвоить алгоритмы решения задач

Обобщение (generalization)

Хорошее качество на обучении
Хорошее качество на новых данных

Отложенная выборка



Обучение



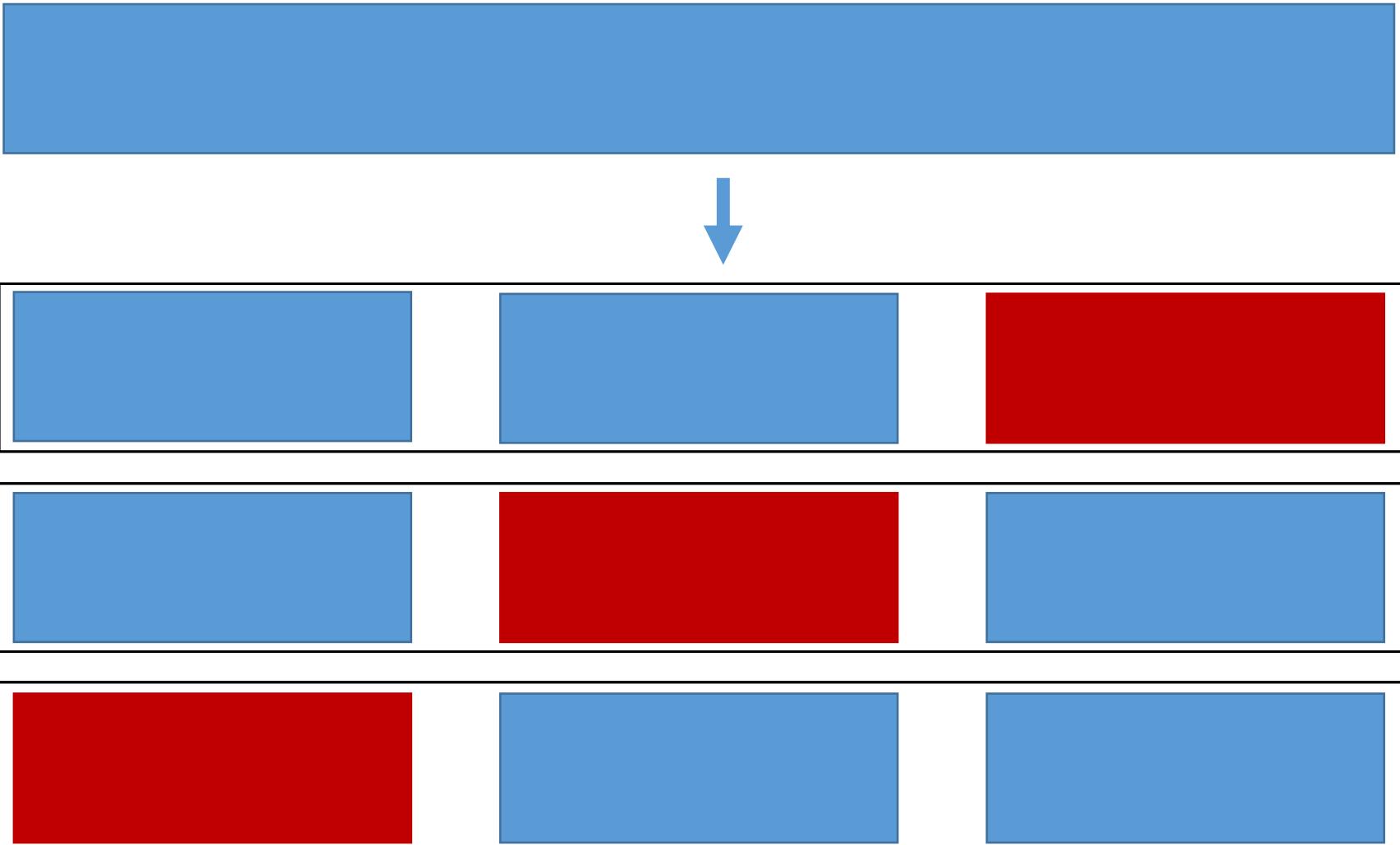
Тест

Отложенная выборка



- Слишком большое обучение — тестовая выборка нерепрезентативна
- Слишком большой тест — модель не сможет обучиться
- Обычно: 70/30, 80/20

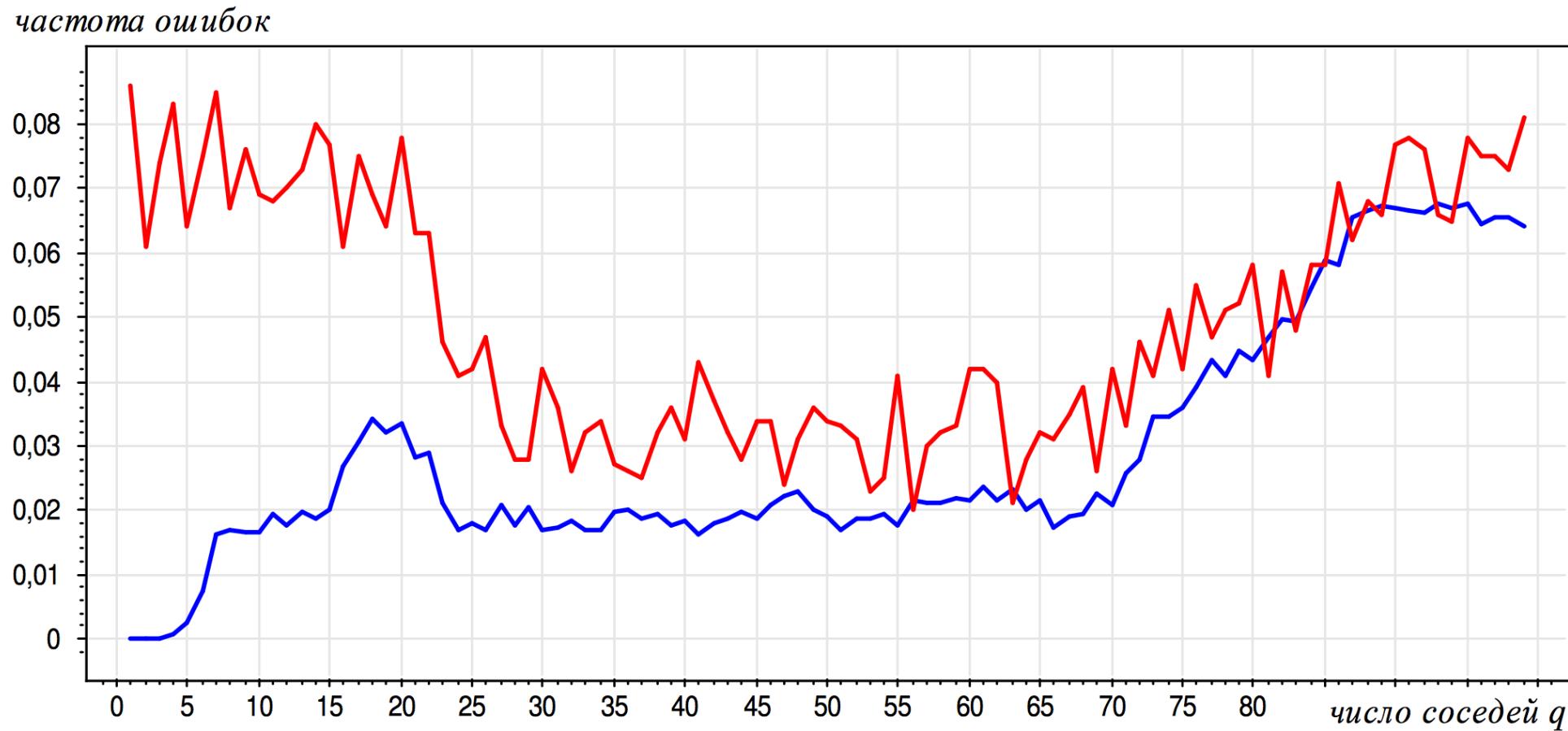
Кросс-валидация



Кросс-валидация

- Надёжнее отложенной выборки, но медленнее
- Параметр — количество разбиений n (фолдов, folds)
- Хороший, но медленный вариант — $n = \ell$ (leave-one-out)
- Обычно: $n = 3$ или $n = 5$ или $n = 10$

Подбор числа соседей



Чуть больше терминов

- После подбора всех гиперпараметров стоит проверить на совсем новых данных, что модель работает
- Обучающая выборка — построение модели
- Валидационная выборка — подбор гиперпараметров модели
- Тестовая выборка — финальная оценка качества модели

Метод k ближайших соседей с
весами

kNN: применение

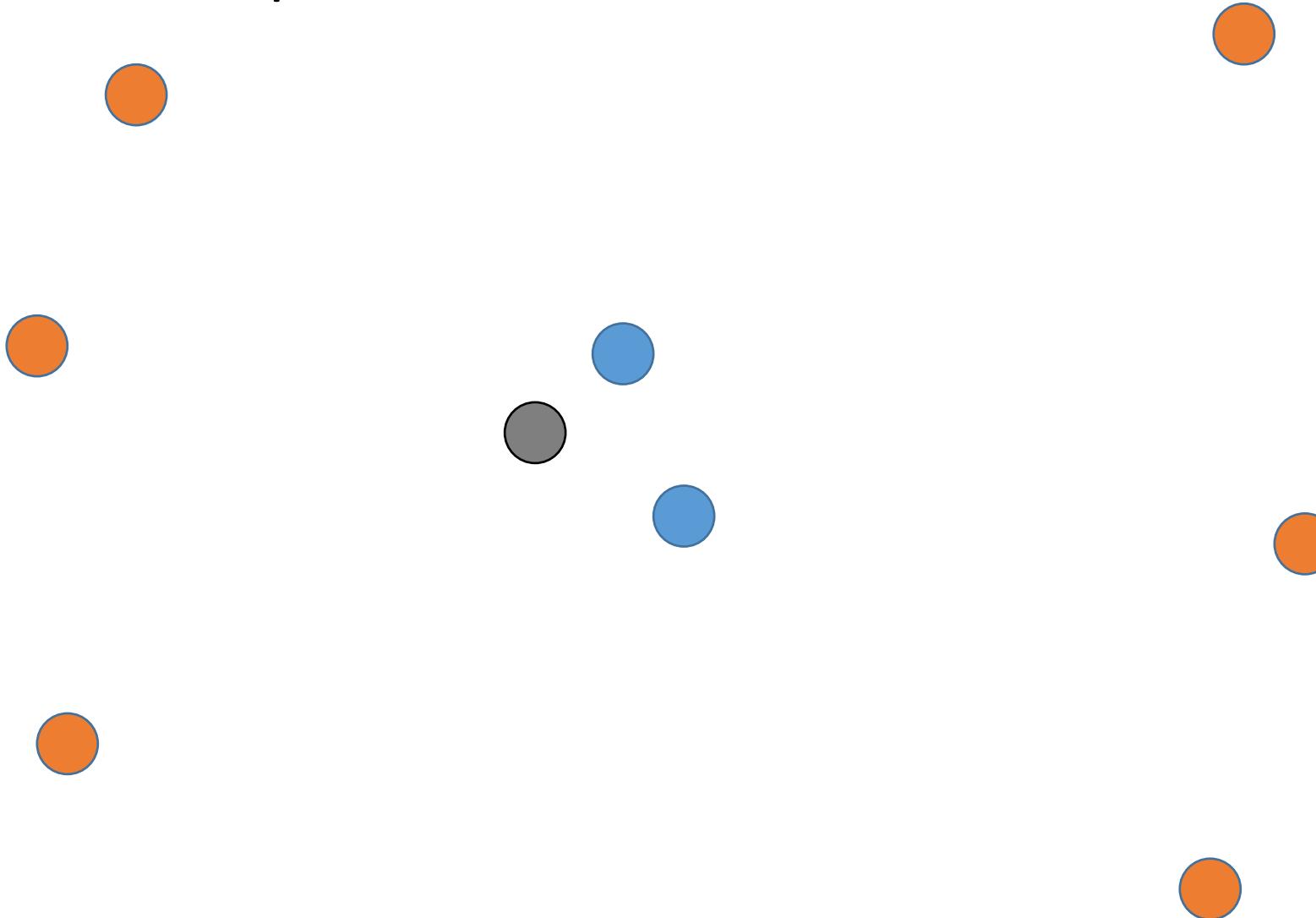
Дано: новый объект x

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта:
 $\rho(x, x_{(1)}) \leq \rho(x, x_{(2)}) \leq \dots \leq \rho(x, x_{(\ell)})$
- Выбираем k ближайших объектов: $x_{(1)}, \dots, x_{(k)}$
- Выдаём наиболее популярный среди них класс:

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

Проблема с расстояниями



Взвешенный kпп

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

Варианты:

- $w_i = \frac{k+1-i}{k}$
- $w_i = q^i$
- Не учитывают сами расстояния

Взвешенный kпп

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

Парзеновское окно:

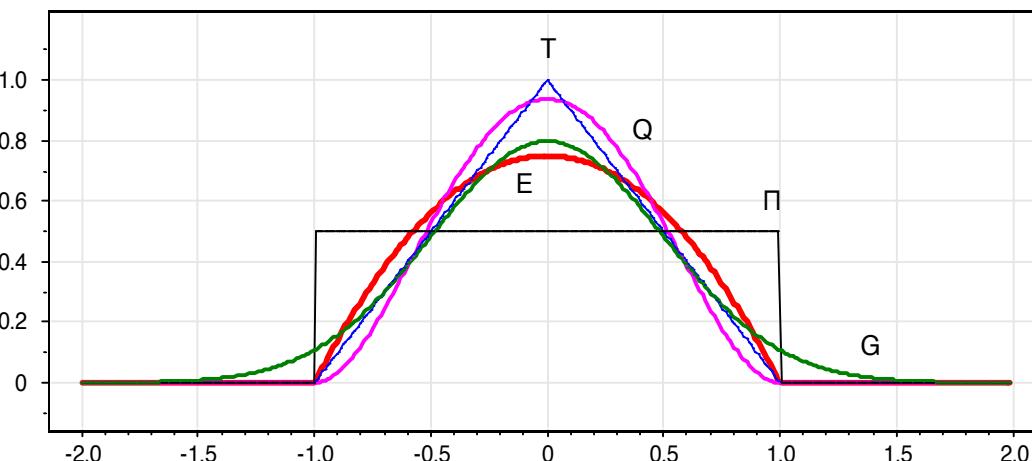
- $w_i = K \left(\frac{\rho(x, x_{(i)})}{h} \right)$
- K — ядро
- h — ширина окна

Ядра для весов

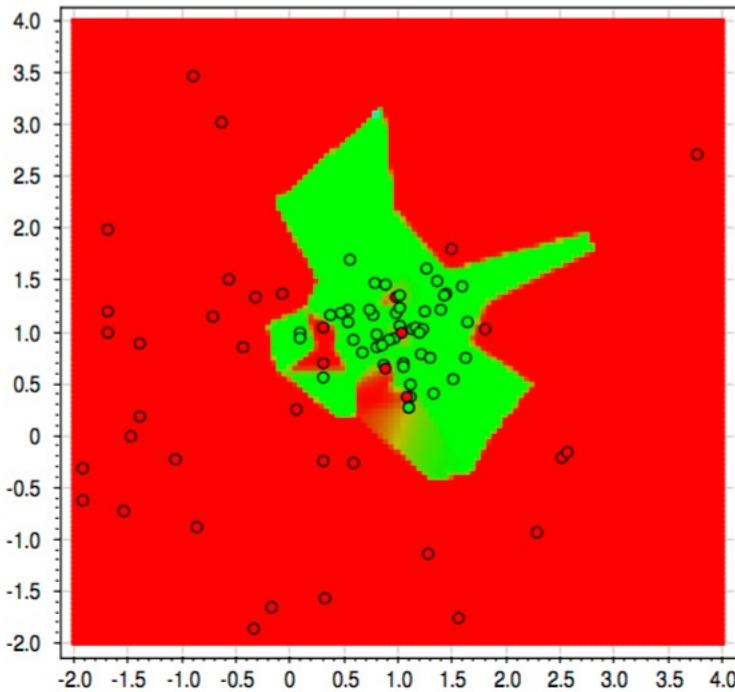
- Гауссовское ядро:

$$K(z) = (2\pi)^{-0.5} \exp\left(-\frac{1}{2}z^2\right)$$

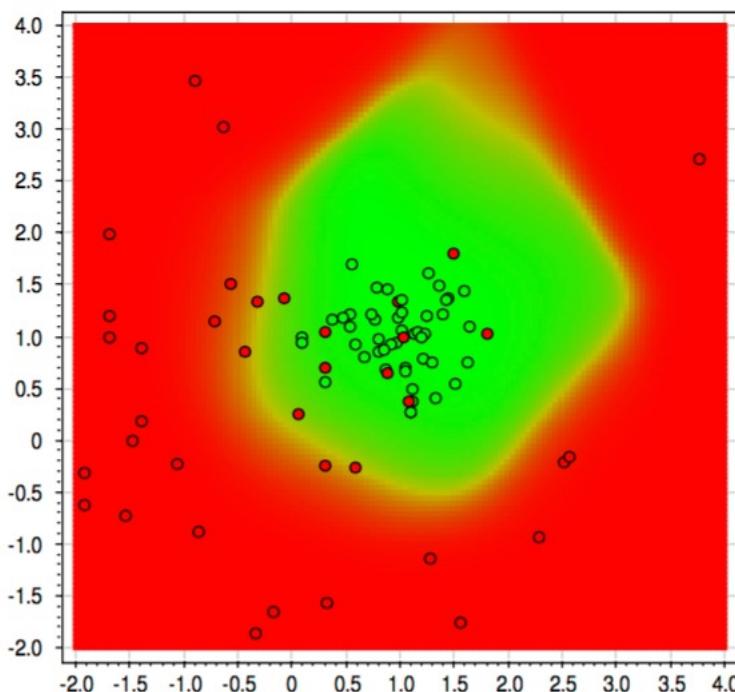
- И много других:



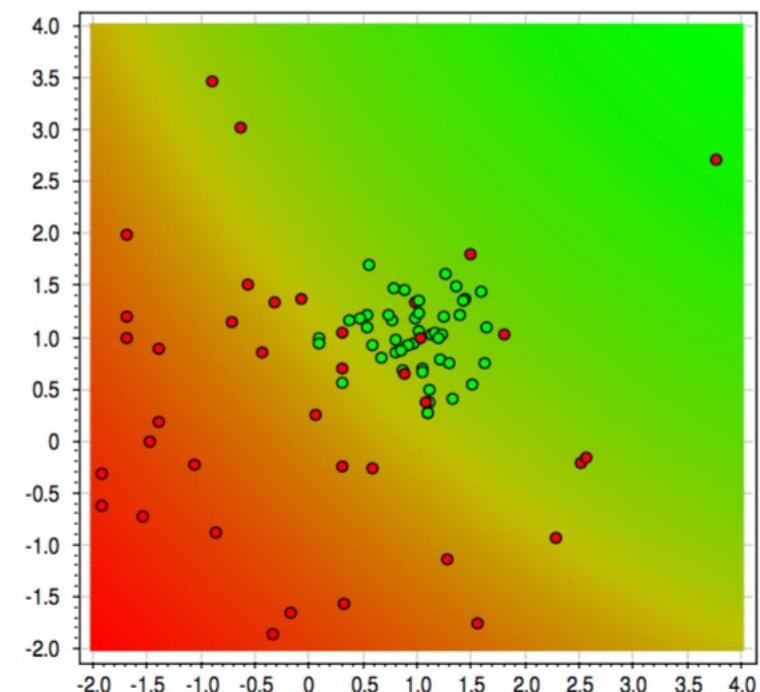
Ядра для весов



$$h = 0.05$$



$$h = 0.5$$



$$h = 5$$