

Введение в анализ данных

Лекция 3

Метод k ближайших соседей и линейная регрессия

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2020

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

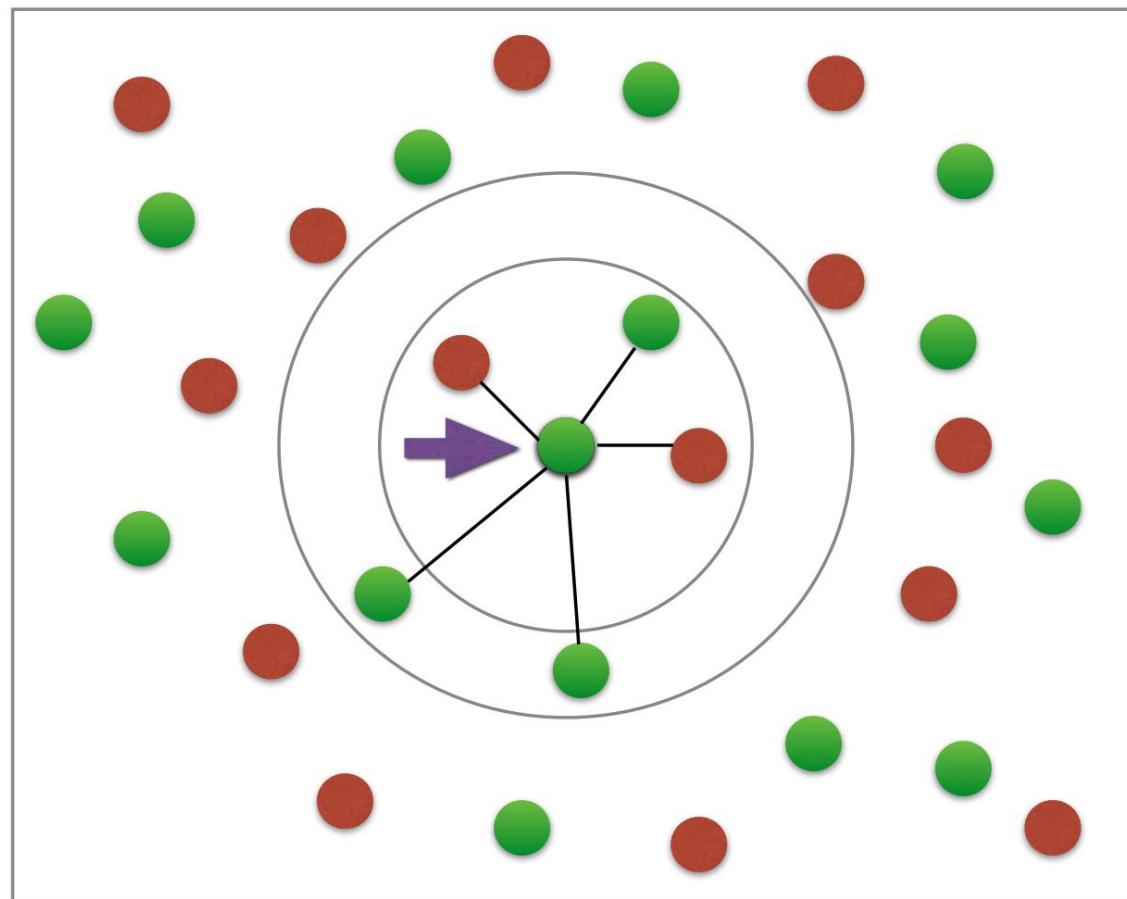
Дано: новый объект x

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:

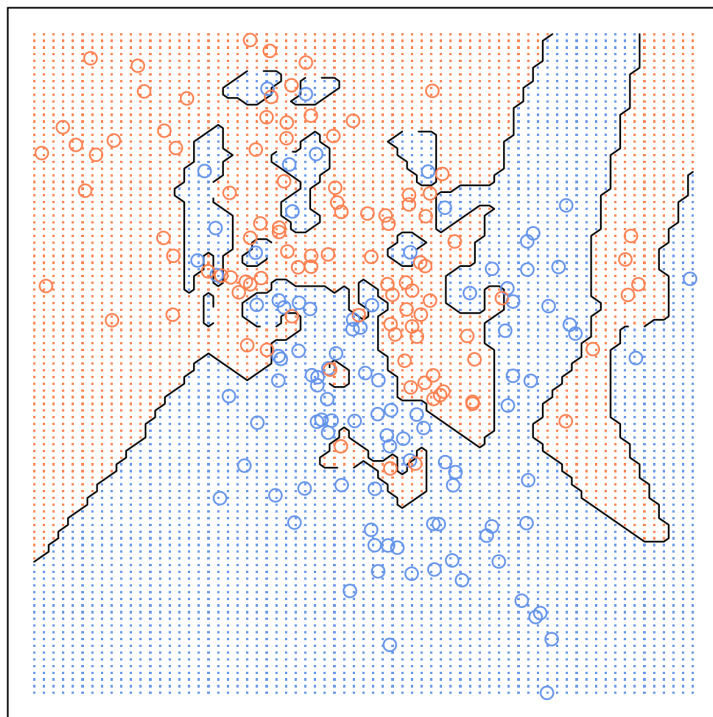
$$a(x) = \arg \max_{y \in \mathbb{Y}} \sum_{i=1}^k [y_{(i)} = y]$$

kNN: применение

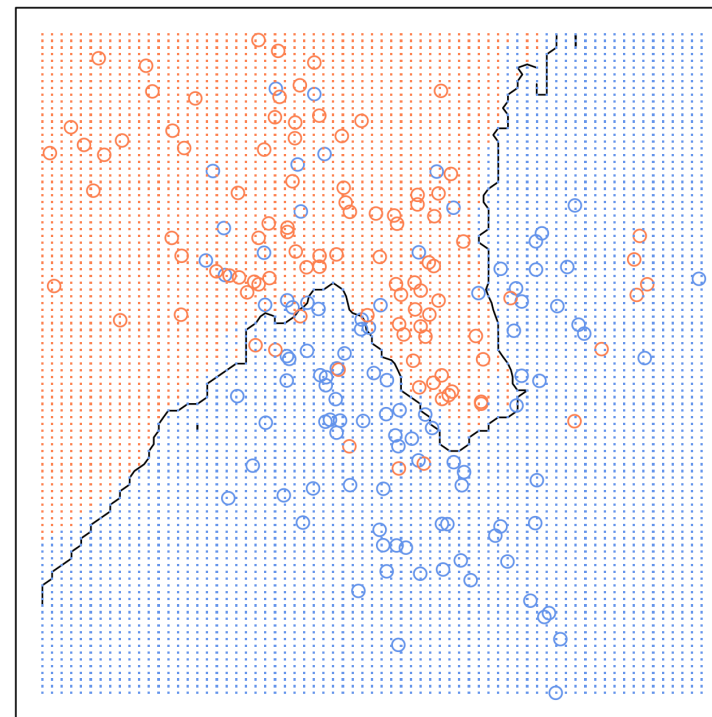


Как выбрать k ?

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

Разобраться в предмете и
усвоить алгоритмы решения
задач

Обобщающая способность

Как готовиться к экзамену?

Заучить все примеры с
занятий

Разобраться в предмете и
усвоить алгоритмы решения
задач

Переобучение (overfitting)

Обобщение (generalization)

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

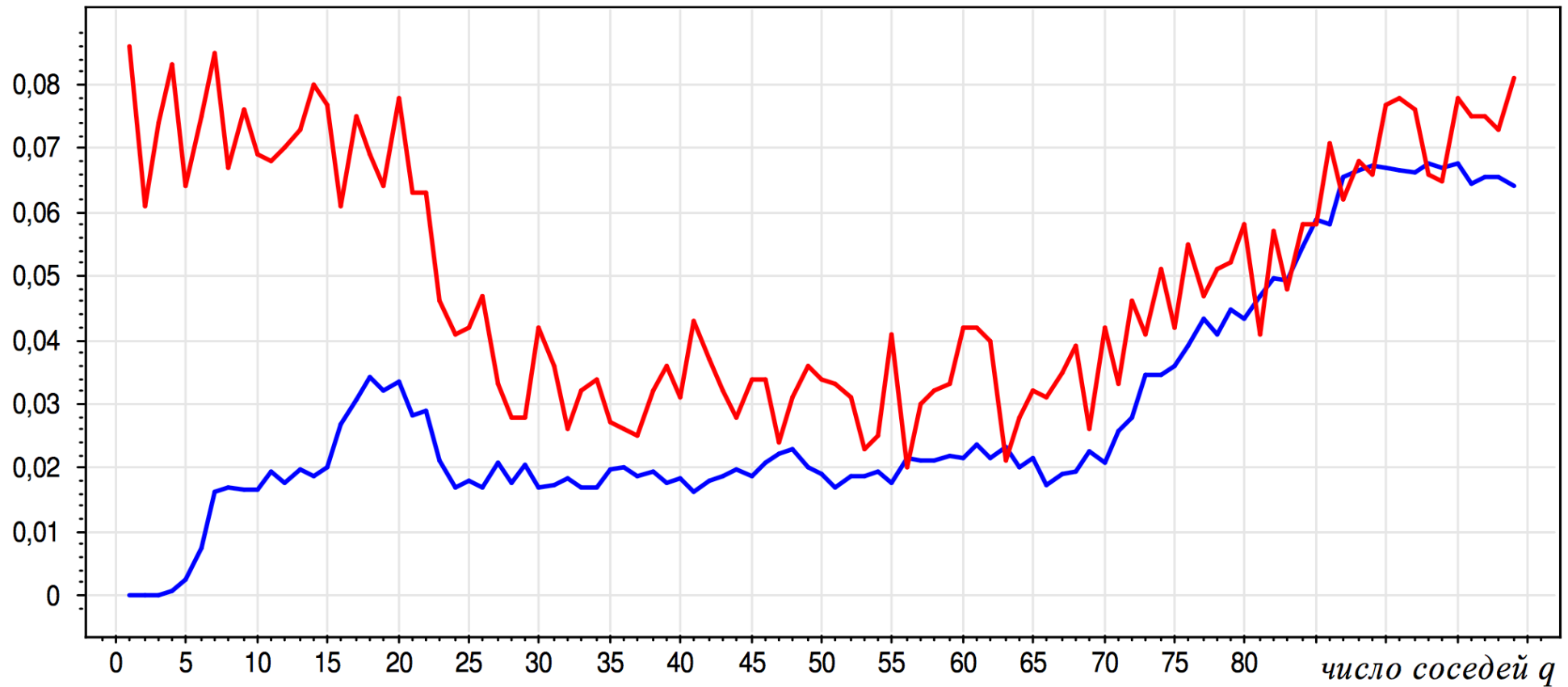


Кросс-валидация

- 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

частота ошибок



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

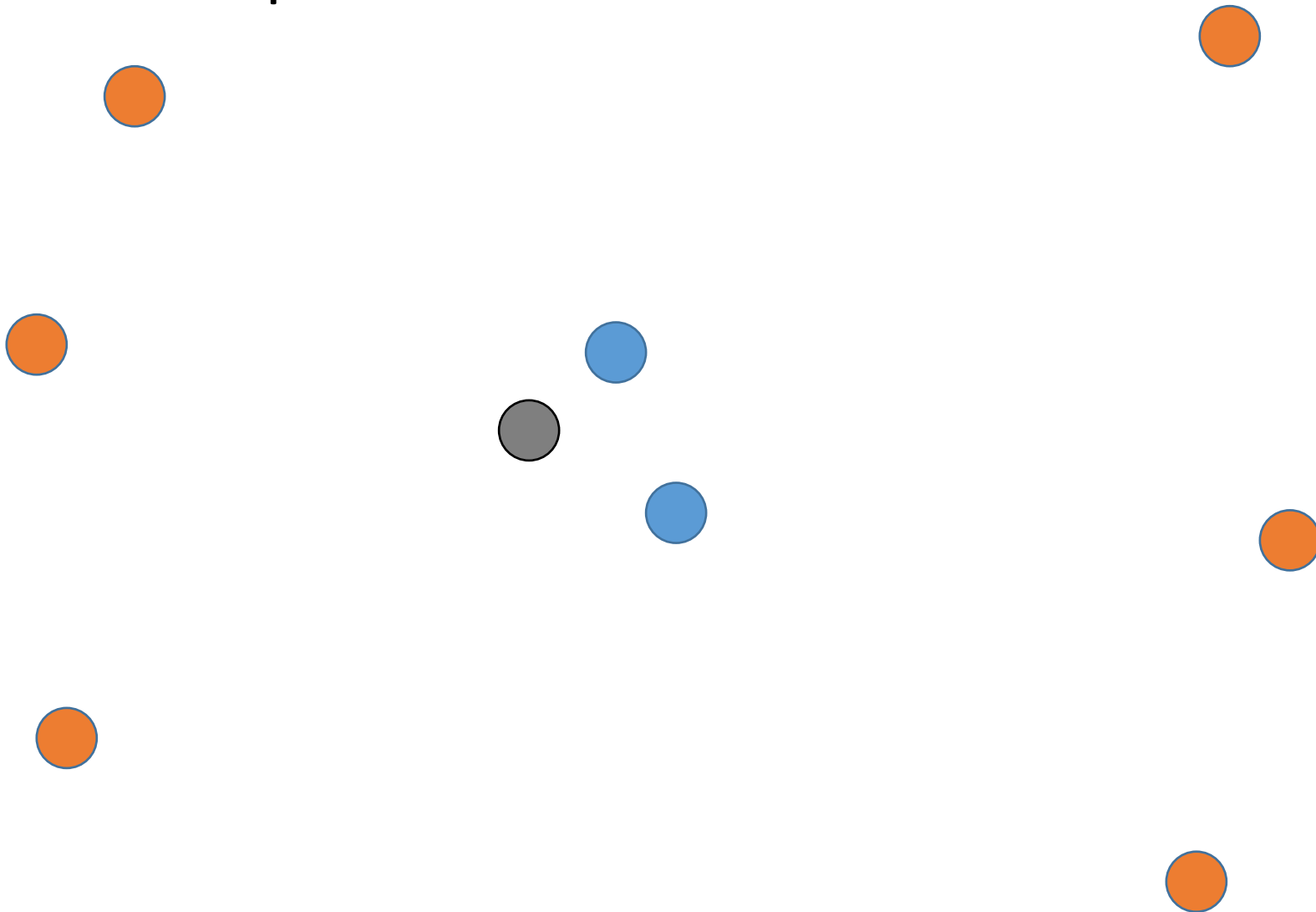
Дано: новый объект x

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 knn

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

Варианты:

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

Взвешенный knn

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

Парзеновское окно:

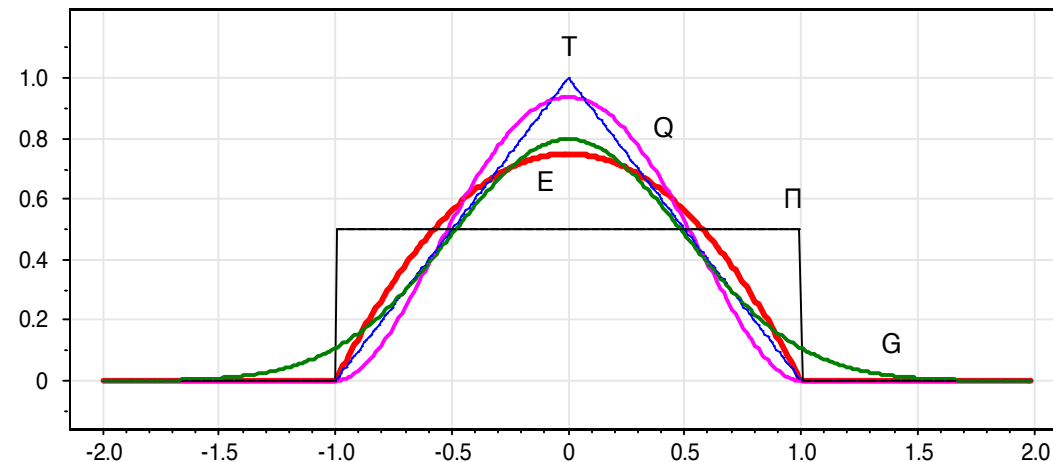
- $w_i = K \left(\frac{\rho(x, x_{(i)})}{h} \right)$
- K — ядро
- h — ширина окна

Ядра для весов

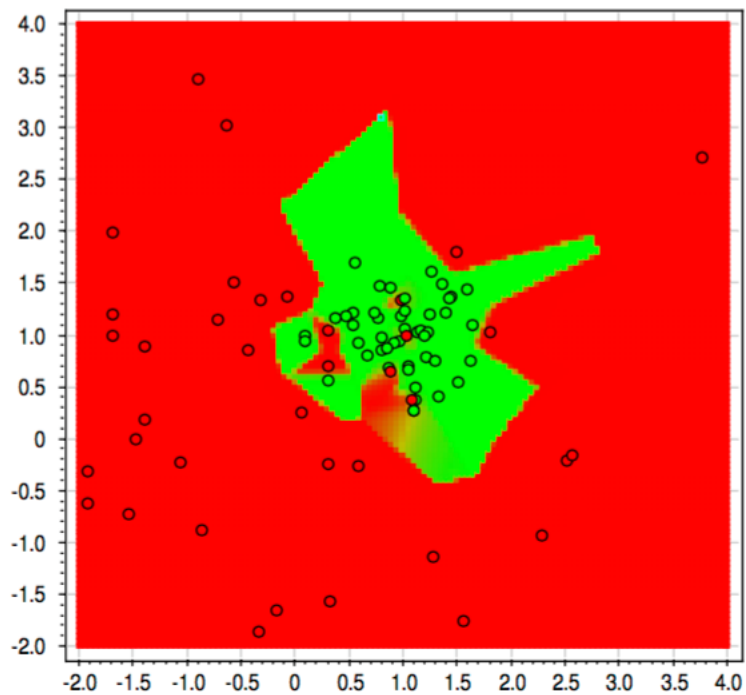
- Гауссовское ядро:

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

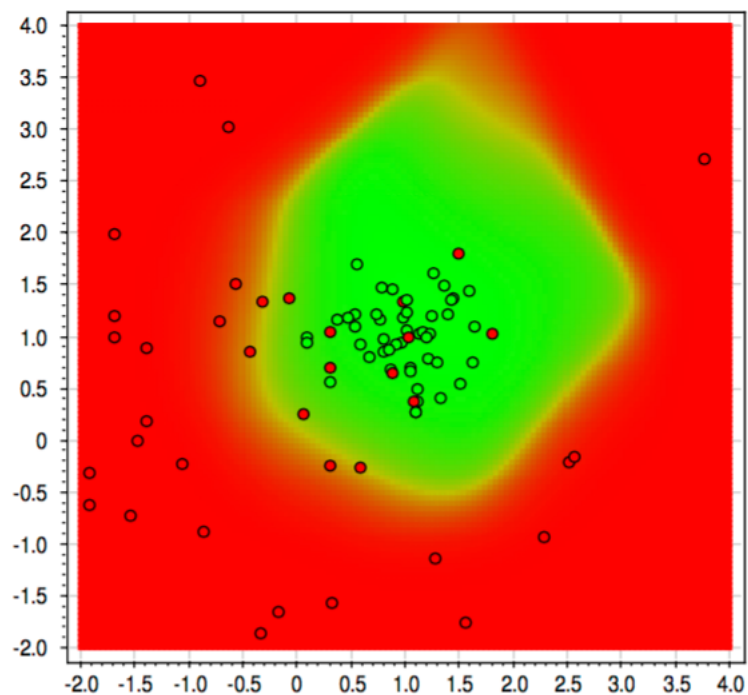
- И много других:



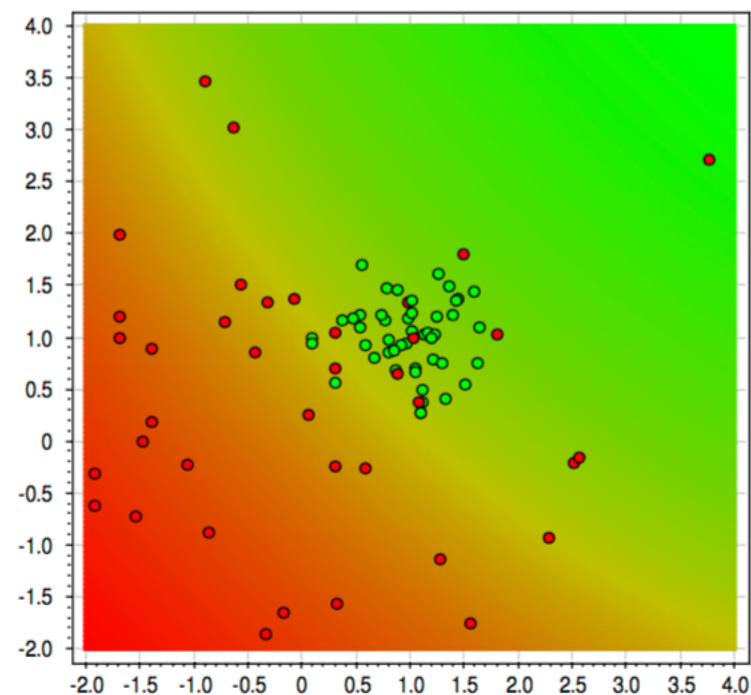
Ядра для весов



$h = 0.05$



$h = 0.5$



$h = 5$

kNN для регрессии

kNN: обучение

- Дано: обучающая выборка $X = (x_i, y_i)_{i=1}^{\ell}$
- Задача регрессии (ответы из множества $\mathbb{Y} = \mathbb{R}$)
- 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: $x_{(1)}, \dots, x_{(k)}$
- Усредняем ответы:

$$a(x) = \frac{1}{k} \sum_{i=1}^k y_{(i)}$$

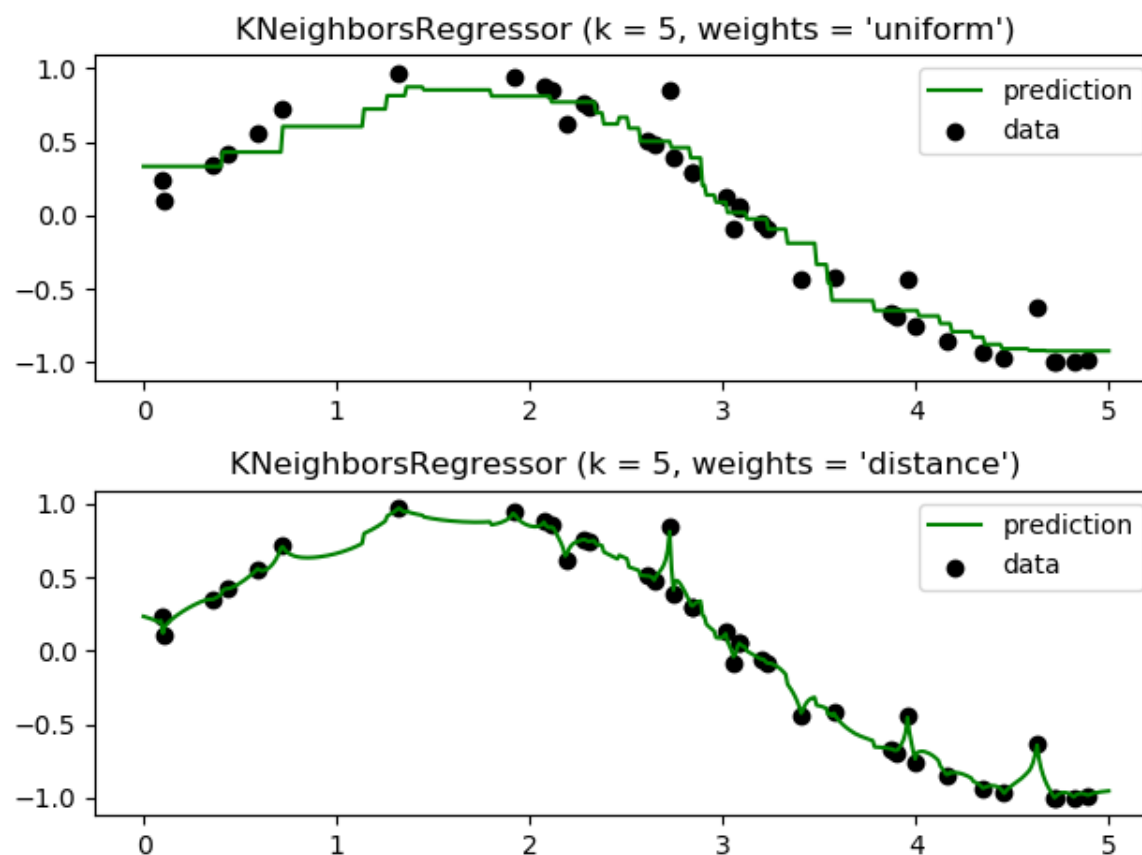
kNN: применение

- Можно добавить веса:

$$a(x) = \frac{\sum_{i=1}^k w_i y_{(i)}}{\sum_{i=1}^k w_i}$$

- $w_i = K \left(\frac{\rho(x, x_{(i)})}{h} \right)$
- Формула Надарая-Ватсона

kNN: применение



Функция потерь для регрессии

- Частый выбор — квадратичная функция потерь

$$L(y, a) = (a - y)^2$$

- Функционал ошибки — среднеквадратичная ошибка (mean squared error, MSE)

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i) - y_i)^2$$

Функция потерь для регрессии

- Ещё один вариант — средняя абсолютная ошибка (mean absolute error, MAE)

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

- Слабее штрафует за серьёзные отклонения от правильного ответа

Резюме

Плюсы kNN

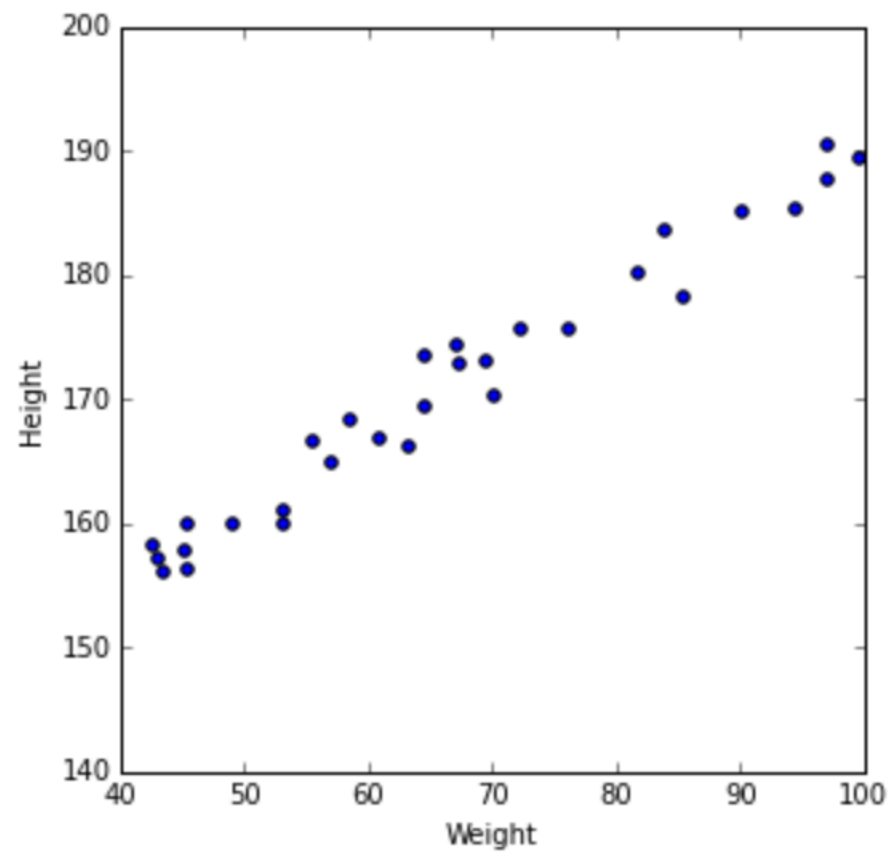
- Если данных много и для любого объекта найдётся похожий в обучающей выборке, то это лучшая модель
- Очень простое обучение
- Мало гиперпараметров
- Бывают задачи, где гипотеза компактности уместна
 - Классификация изображений
 - Классификация текстов на много классов

Минусы kNN

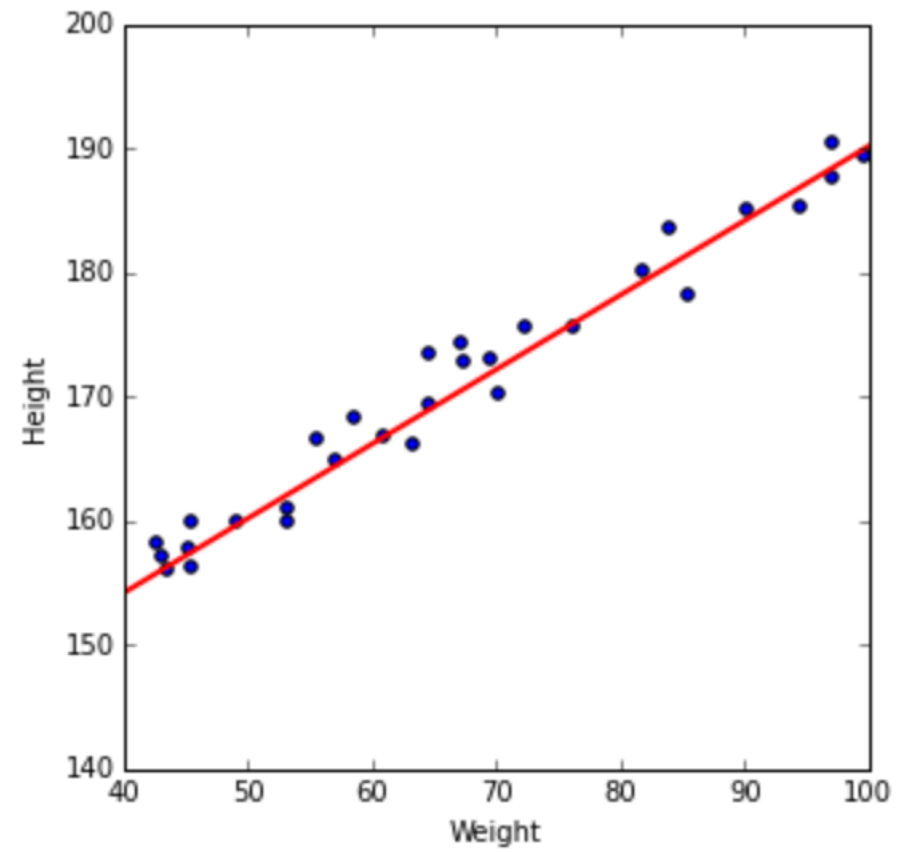
- Часто другие модели оказываются лучше
- Надо хранить в памяти всю обучающую выборку
- Искать k ближайших соседей довольно долго
- Мало способов настроить модель

Линейная регрессия

Парная регрессия



Парная регрессия



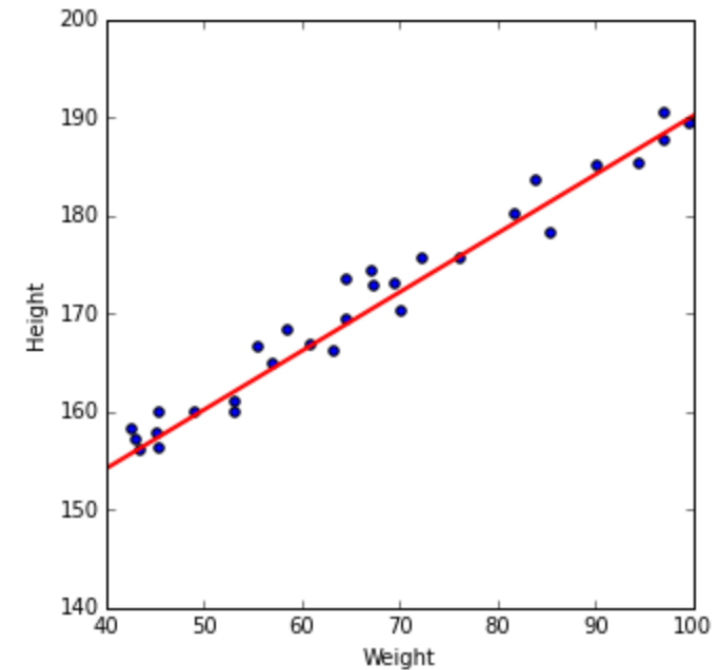
Парная регрессия

- Простейший случай: один признак
- Модель: $a(x) = w_1 x + w_0$
- Два параметра: w_1 и w_0
- w_1 — тангенс угла наклона
- w_0 — где прямая пересекает ось ординат

Почему модель линейная?

$$a(x) = 2x + 1$$

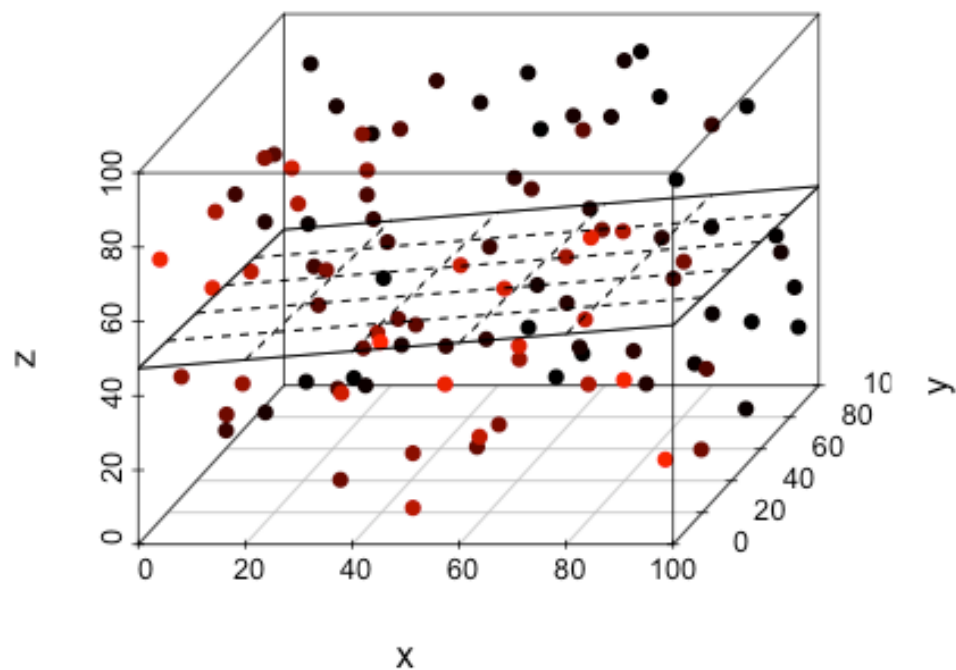
- $x = 1, a(x) = 3$
- $x = 2, a(x) = 5$
- $x = 10, a(x) = 21$
- $x = 20, a(x) = 41$



Два признака

- Чуть более сложный случай: два признака
- Модель: $a(x) = w_0 + w_1 x_1 + w_2 x_2$
- Три параметра

Два признака



Много признаков

- Общий случай: d признаков
- Модель

$$a(x) = w_0 + w_1x_1 + \dots + w_dx_d$$

- Количество параметров: $d + 1$

Много признаков

- Общий случай: d признаков
- Модель

$$a(x) = w_0 + w_1x_1 + \dots + w_dx_d$$

Свободный коэффициент/сдвиг/bias

Веса/коэффициенты

- Количество параметров: $d + 1$

Много признаков

$$\begin{aligned} a(x) &= w_0 + w_1x_1 + \dots + w_dx_d = \\ &= w_0 + \langle w, x \rangle \end{aligned}$$

- Будем считать, что есть признак, всегда равный единице:

$$\begin{aligned} a(x) &= w_1x_1 + \dots + w_dx_d = \\ &= w_1 * 1 + w_2x_2 + \dots + w_dx_d = \\ &= \langle w, x \rangle \end{aligned}$$