Основы машинного обучения

Лекция 4

Метод k ближайших соседей. Линейная регрессия.

Евгений Соколов

esokolov@hse.ru

НИУ ВШЭ, 2025

Метод k ближайших соседей с весами

kNN: применение

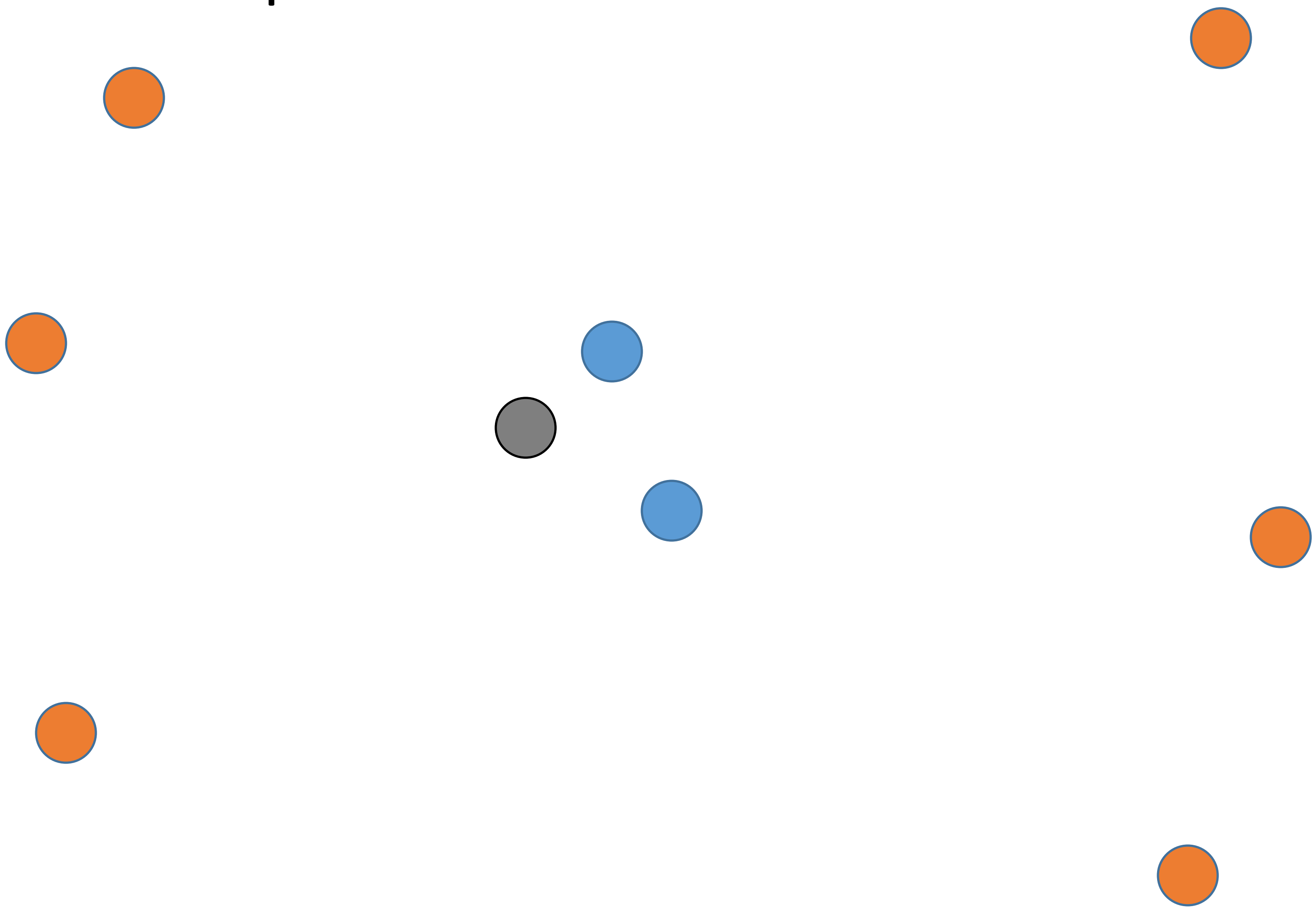
Дано: новый объект x

Применение модели:

- Сортируем объекты обучающей выборки по расстоянию до нового объекта: $\rho\left(x, x_{(1)}\right) \le \rho\left(x, x_{(2)}\right) \le \dots \le \rho\left(x, x_{(\ell)}\right)$
- 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:

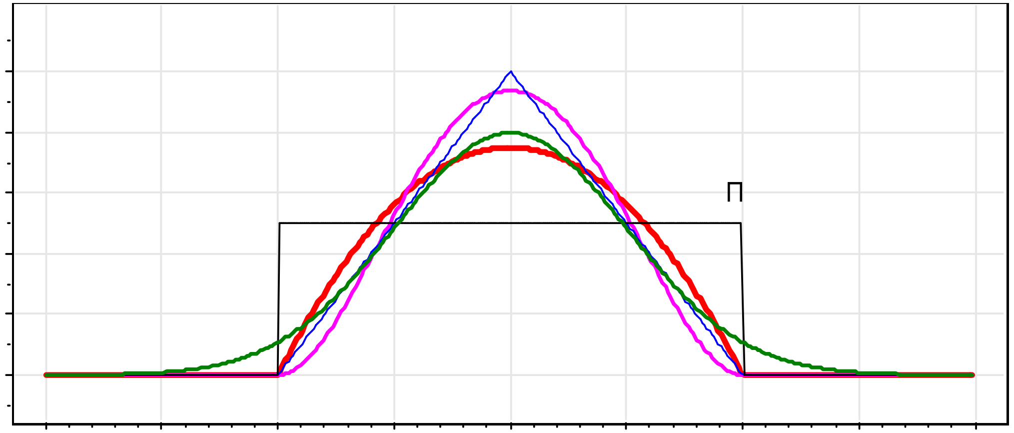
- $w_i = K\left(\frac{\rho(x, x_{(i)})}{h}\right)$
- K — ядро
- h — ширина окна

Ядра для весов

- Гауссовское ядро:

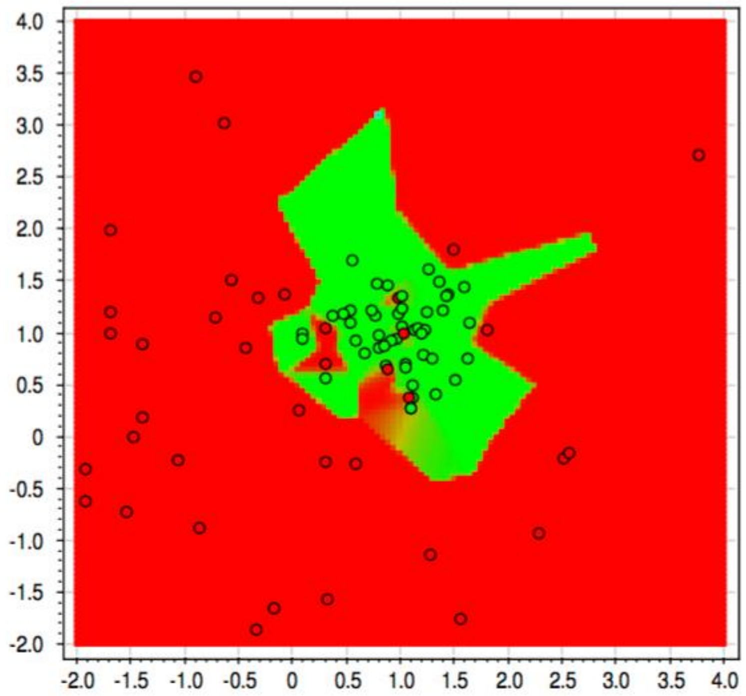
$$K(z) = (2\pi)^{-0.5} \exp\left(-\frac{1}{2} z^2\right)$$

- И много других:

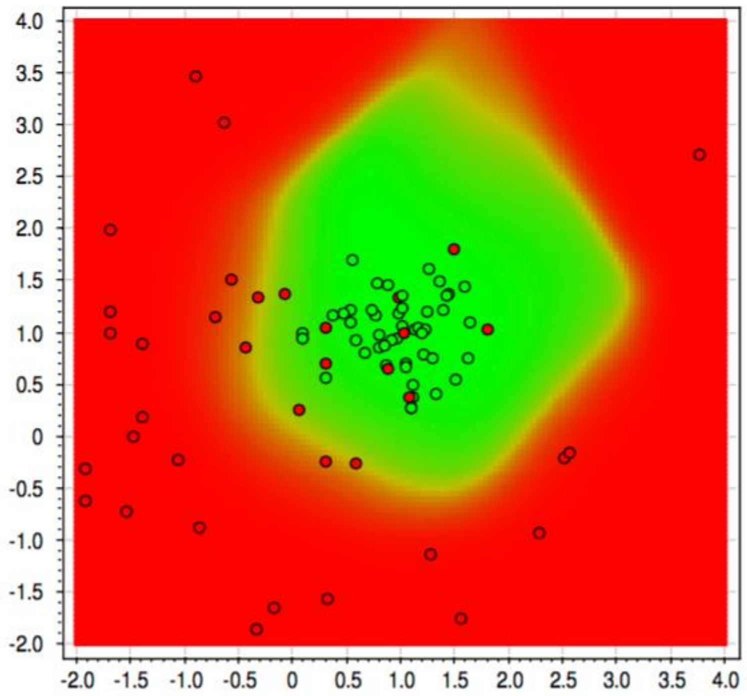


http://www.machinelearning.ru/wiki/index.php?title=MO

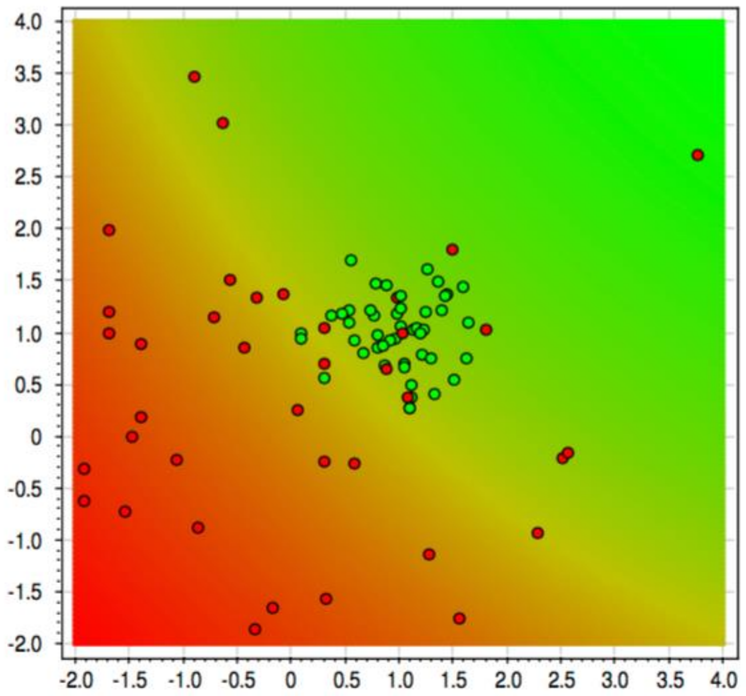
Ядра для весов



$h = 0.05$



$h = 0.5$



$h = 5$

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: $\rho(x, x_{(1)}) \le \rho(x, x_{(2)}) \le \dots \le \rho(x, x_{(\ell)})$

- Выбираем k ближайших объектов: $x_{(1)}, \dots, x_{(k)}$

- Усредняем ответы:

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

kNN: применение

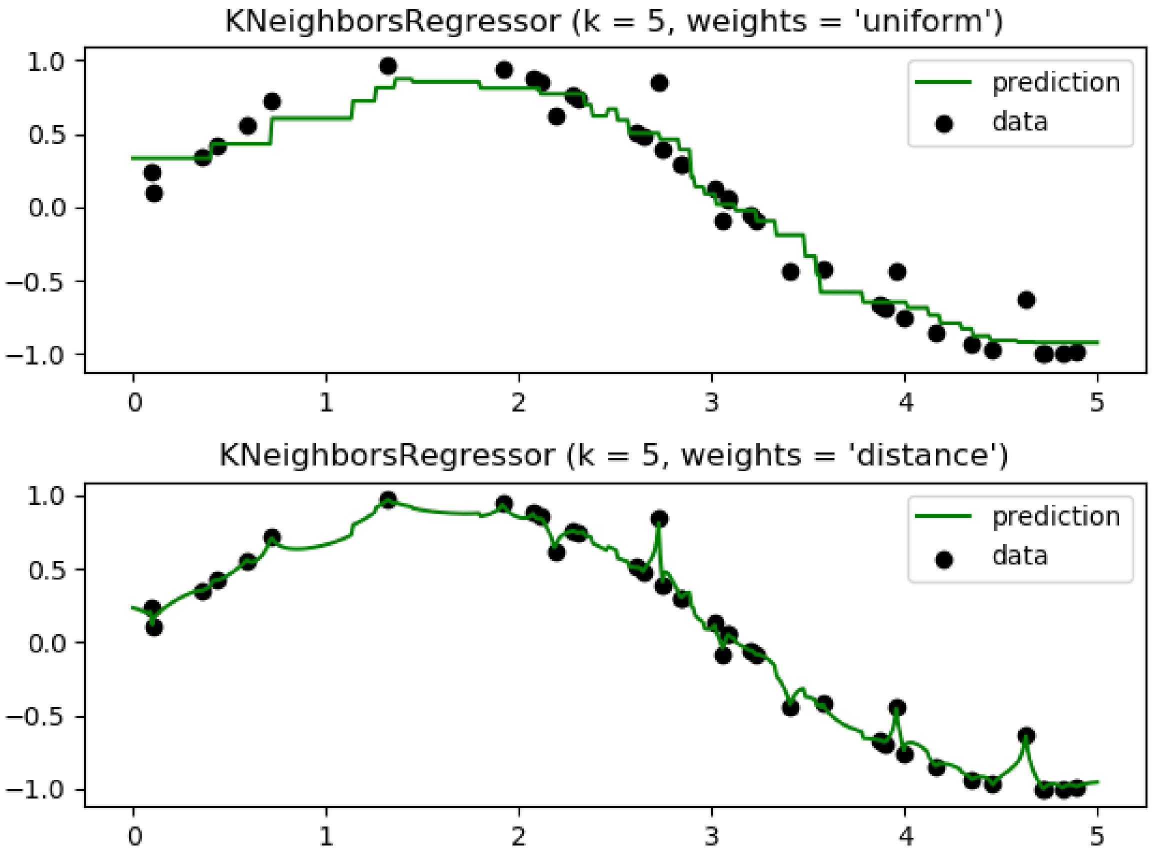
- Можно добавить веса:

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

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

- 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

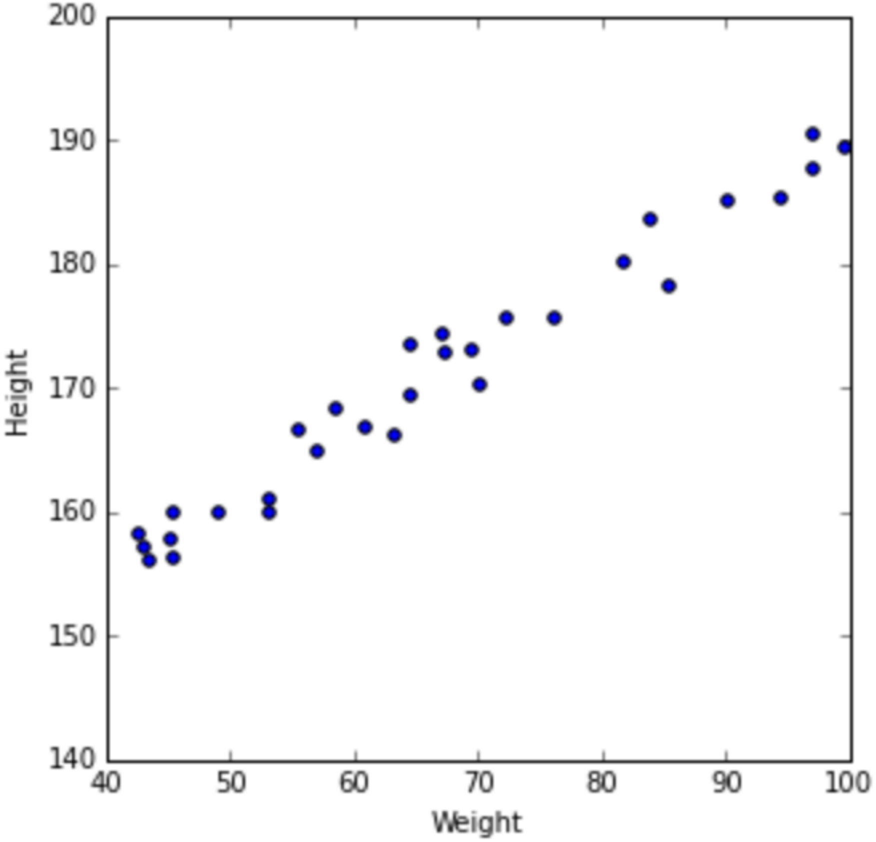
- Если данных много и для любого объекта найдётся похожий в обучающей выборке, то это лучшая модель
- Очень простое обучение
- Мало гиперпараметров
- Бывают задачи, где гипотеза компактности уместна
    - Классификация изображений
    - Классификация текстов на много классов

Минусы kNN

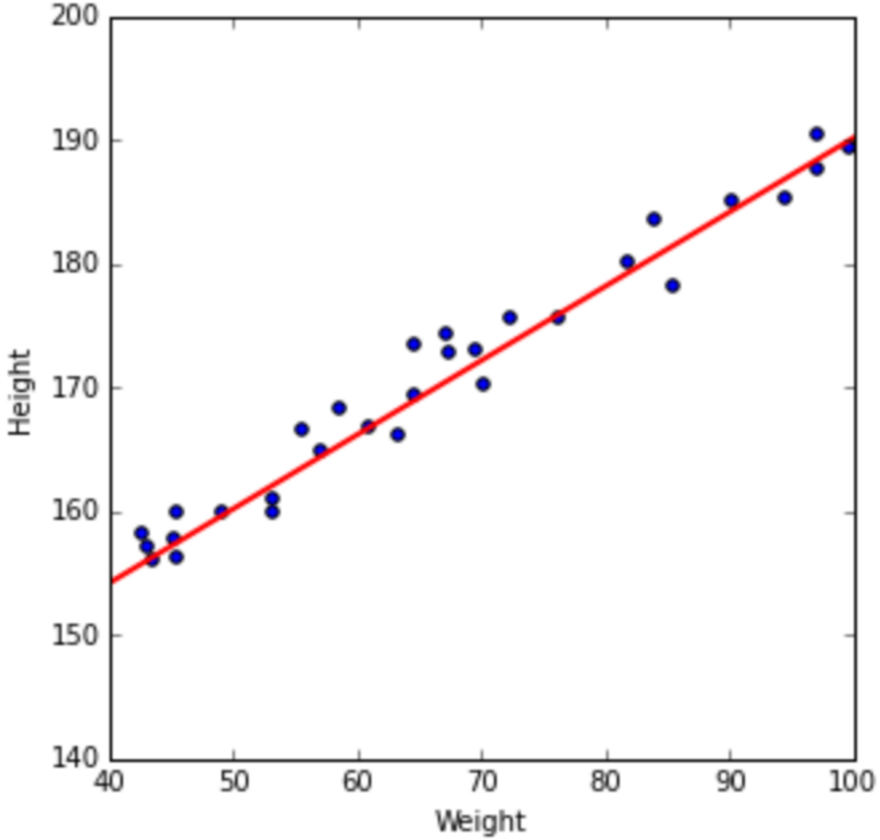
- Часто другие модели оказываются лучше
- Надо хранить в памяти всю обучающую выборку
- Искать k ближайших соседей довольно долго
- Мало способов настроить модель

Линейная регрессия

Парная регрессия



Парная регрессия



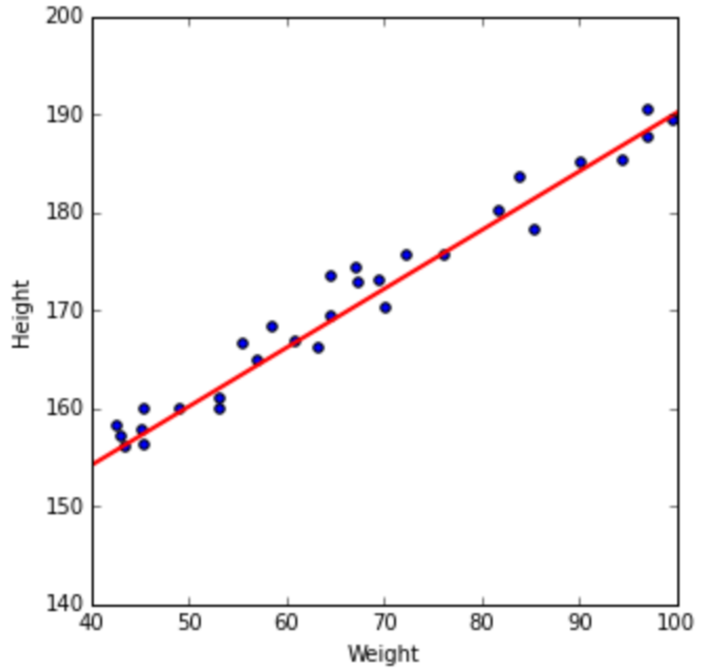
Парная регрессия

- 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*?

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

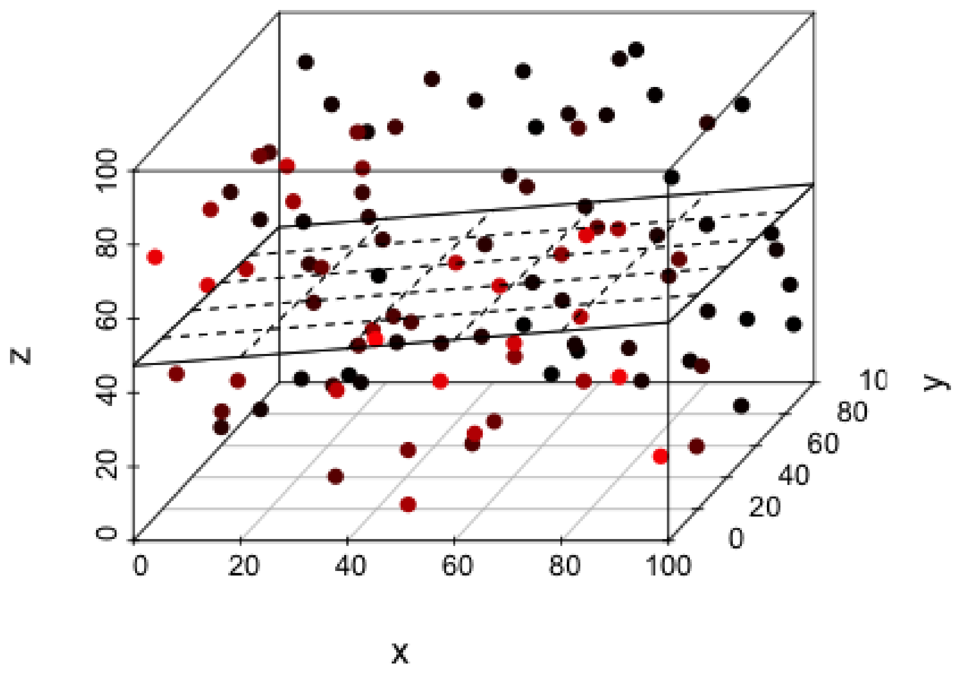
- $x = 1, a(x) = 3$
- $x = 2, a(x) = 5$
- $x = 10, a(x) = 21$
- $x = 20, a(x) = 41$



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

$$a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d$$

- Количество параметров: $d + 1$

Много признаков

- Общий случай: d признаков
- Модель

$$a(x) = w_0 + w_1 x_1 + \cdots + w_d x_d$$

Веса/коэффициенты

Свободный коэффициент/сдвиг/bias

- Количество параметров: $d + 1$

Много признаков

Запишем через скалярное произведение:

$$a(x) = w_0 + w_1 x_1 + \dots + w_d x_d =$$

$$= w_0 + \langle w, x \rangle$$

Будем считать, что есть признак, всегда равный единице:

$$a(x) = w_1 x_1 + \dots + w_d x_d =$$

$$= w_1 * 1 + w_2 x_2 + \dots + w_d x_d =$$

$$= \langle w, x \rangle$$

Применимость линейной регрессии

Модель линейной регрессии

$$a(x) = w_1 x_1 + \dots + w_d x_d = \langle w, x \rangle$$

- Нет гарантий, что целевая переменная именно так зависит от признаков
- Надо формировать признаки так, чтобы модель подходила

Предсказание стоимости квартиры

- Признаки: площадь, район, расстояние до метро
- Целевая переменная: рыночная стоимость квартиры
- Линейная модель:

$$a(x) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{район}) + w_3 * (\text{расстояние до метро})$$

Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$

$$+ w_2 * (\text{район})$$

$$+ w_3 * (\text{расстояние до метро})$$

Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$
$$+ w_2 * (\text{район})$$
$$+ w_3 * (\text{расстояние до метро})$$

- За каждый квадратный метр добавляем w_1 к прогнозу

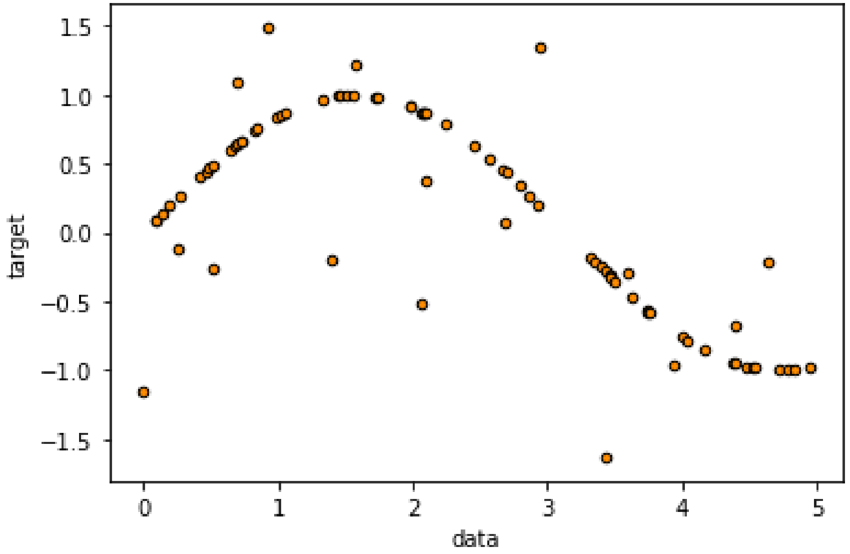
Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$
$$+ w_2 * (\text{район})$$
$$+ w_3 * (\text{расстояние до метро})$$

- Что-то странное

Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$
$$+ w_2 * (\text{район})$$
$$+ w_3 * (\text{расстояние до метро})$$



Кодирование категориальных признаков

- Значения признака «район»: $U = \{u_1, \ldots, u_m\}$
- Новые признаки вместо x_j: $[x_j = u_1], \ldots, [x_j = u_m]$
- One-hot кодирование

Кодирование категориальных признаков

Район
ЦАО
ЮАО
ЦАО
САО
ЮАО

ЦАО	ЮАО	САО
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0

Кодирование категориальных признаков

Район
ЦАО
ЮАО
ЦАО
САО
ЮАО

ЦАО	ЮАО	САО
1	0	0
0	1	0
1	0	0
0	0	1
0	1	0

$$a(x) = w_0 + w_1 * (\text{площадь})$$

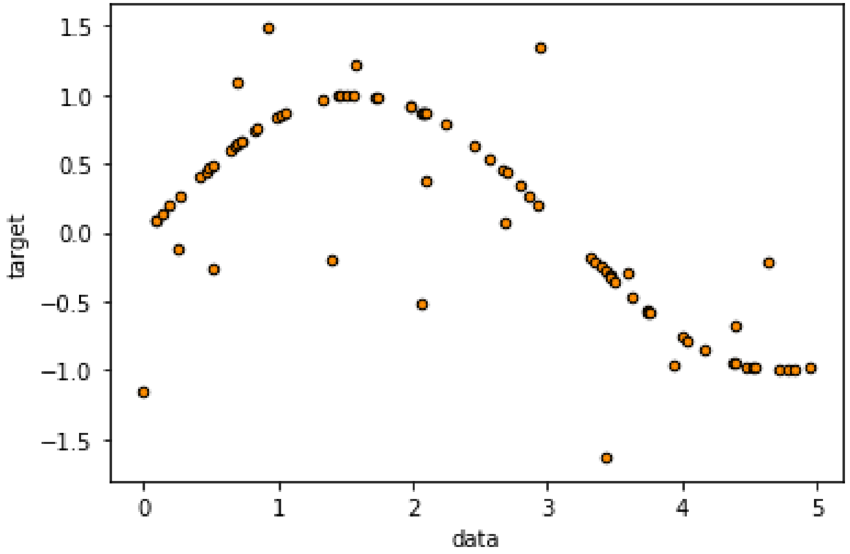
$$+ w_2 * (\text{квартира в ЦАО?})$$

$$+ w_3 * (\text{квартира в ЮАО?})$$

$$+ w_4 * (\text{квартира в САО?})$$

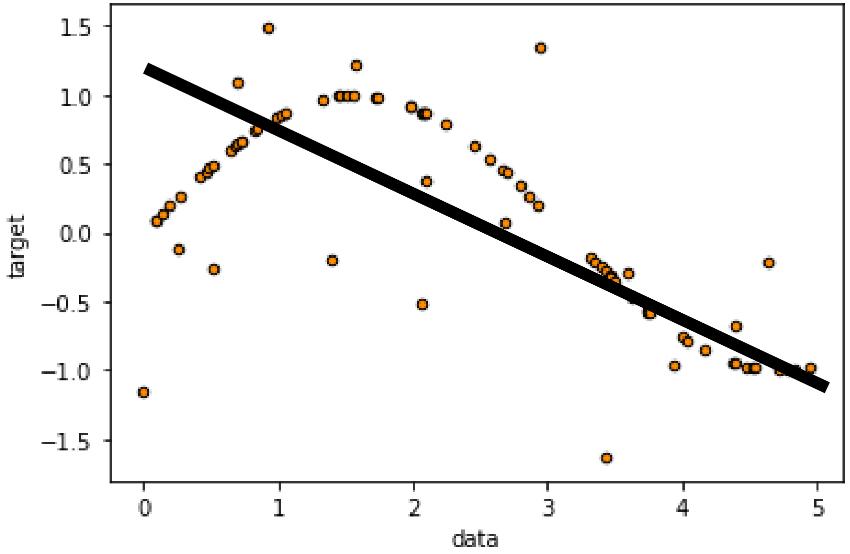
Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$
$$+ w_2 * (\text{район})$$
$$+ w_3 * (\text{расстояние до метро})$$



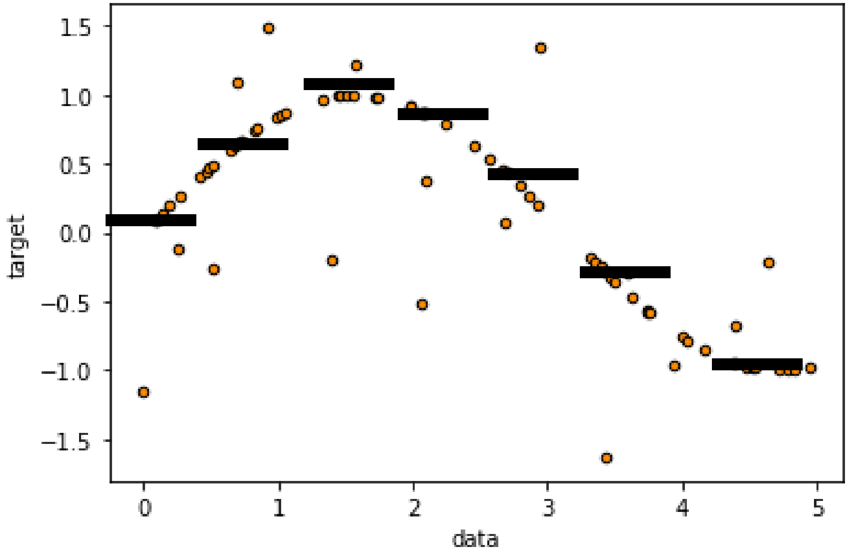
Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$
$$+ w_2 * (\text{район})$$
$$+ \color{red}{w_3 * (\text{расстояние до метро})}$$



Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$
$$+ w_2 * (\text{район})$$
$$+ w_3 * (\text{расстояние до метро})$$

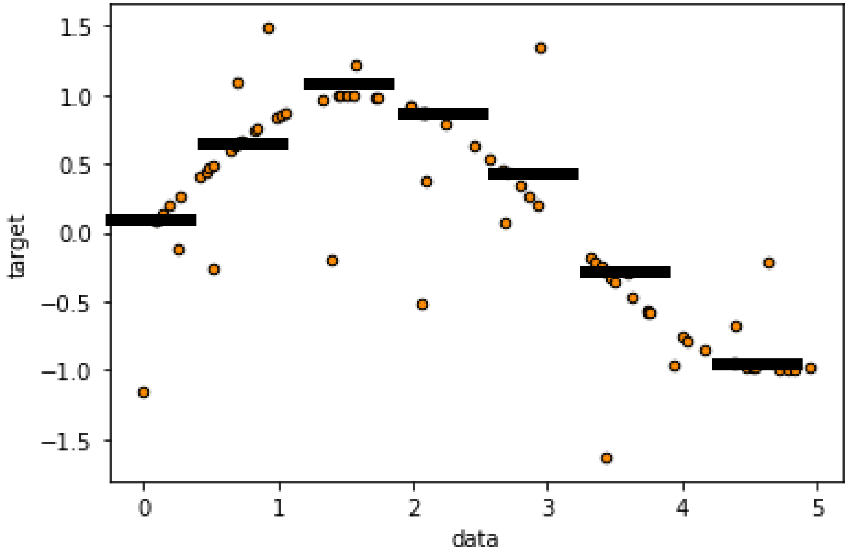


Предсказание стоимости квартиры

$$a(x) = w_0 + w_1 * (\text{площадь})$$

$$+ w_2 * (\text{район})$$

$$+ w_3 * [t_0 \leq x_3 < t_1] + \cdots + w_{3+n}[t_{n-1} \leq x_3 < t_n]$$



Нелинейные признаки

- Линейная модель с полиномиальными признаками:

$$a(x) = w_0 + w_1 * (\text{площадь}) + w_2 * (\text{этаж})$$
$$+ w_3 * (\text{расстояние до метро}) + w_4 * (\text{площадь})^2$$
$$+ w_5 * (\text{этаж})^2 + w_6 * (\text{расстояние до метро})^2$$
$$+ w_7 * (\text{площадь}) * (\text{этаж}) + \cdots$$

Линейные модели

- Модель линейной регрессии хороша, если признаки сделаны специально под неё
- Пример: one-hot кодирование категориальных признаков или бинаризация числовых признаков

Линейная регрессия в векторном виде

Модель линейной регрессии

$$a(x) = \langle w, x \rangle$$

- Среднеквадратичная ошибка и задача обучения:

$$\frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2 \to \min_w$$

Матрицы

- Матрица — таблица с числами (для простоты)
- Матрица «объекты-признаки»:

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix} \in \mathbb{R}^{\ell \times d}$$

Матрицы

- Матрица — таблица с числами (для простоты)
- Матрица «объекты-признаки»:

$$\begin{array}{r} \\ \text{объект и его признаки} \\ \\ \\ \end{array} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

Матрицы

- Матрица — таблица с числами (для простоты)
- Матрица «объекты-признаки»:

$$\begin{pmatrix} x_{11} & \boxed{x_{12}} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

значения признака на всех объектах

Векторы

- Вектор размера d — тоже матрица

- Вектор-строка: $w = (w_1, \dots, w_d) \in \mathbb{R}^{1 \times d}$

- Вектор-столбец: $w = \begin{pmatrix} w_1 \\ \cdots \\ w_d \end{pmatrix} \in \mathbb{R}^{d \times 1}$

Матричное умножение

- Только для матриц $A \in \mathbb{R}^{m \times k}$ и $B \in \mathbb{R}^{k \times n}$
- Результат: $AB = C \in \mathbb{R}^{m \times n}$
- Правило:

$$c_{ij} = \sum_{p=1}^{k} a_{ip} b_{pj}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} \boxed{1} & & \\ & & \\ & & \end{pmatrix}$$

Пример

$$\begin{pmatrix} \boxed{1 \quad 2} \\ 0 \quad 1 \\ 10 \quad 0 \end{pmatrix} * \begin{pmatrix} 1 & \boxed{\begin{matrix} 0 \\ 0 \end{matrix}} & 0 \\ 0 & & 2 \end{pmatrix} = \begin{pmatrix} 1 & \boxed{0} & \\ & & \\ & & \end{pmatrix}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ 0 & 1 \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 2 \end{pmatrix} = \begin{pmatrix} 1 & 0 & \boxed{4} \\ & & \\ & & \end{pmatrix}$$

Пример

$$\begin{pmatrix} 1 & 2 \\ \boxed{0 \quad 1} \\ 10 & 0 \end{pmatrix} * \begin{pmatrix} \boxed{\begin{matrix} 1 \\ 0 \end{matrix}} & \begin{matrix} 0 \\ 0 \end{matrix} & \begin{matrix} 0 \\ 2 \end{matrix} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 4 \\ \boxed{0} & & \\ & & \end{pmatrix}$$

Применение линейной модели

- $a(x) = \langle w, x \rangle = w_1 x_1 + \dots + w_d x_d$
- Как применить модель к обучающей выборке?

$$\begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{\ell 1} & x_{\ell 2} & \cdots & x_{\ell d} \end{pmatrix}$$

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_d \end{pmatrix}$$

$$\begin{pmatrix} \sum_{i=1}^{d} w_i x_{1i} \\ \sum_{i=1}^{d} w_i x_{2i} \\ \vdots \\ \sum_{i=1}^{d} w_i x_{\ell i} \end{pmatrix}$$

Модель линейной регрессии

- Среднеквадратичная ошибка и задача обучения:

$$\frac{1}{\ell}\sum_{i=1}^{\ell}(\langle w, x_i \rangle - y_i)^2 \to \min_w$$

Вычисление ошибки

- Отклонения прогнозов от ответов:

$$Xw - y = \begin{pmatrix} \langle w, x_1 \rangle - y_1 \\ \vdots \\ \langle w, x_\ell \rangle - y_\ell \end{pmatrix}$$

Вычисление ошибки

- Евклидова норма:

$$\|z\| = \sqrt{\sum_{j=1}^{n} z_j^2}$$

$$\|z\|^2 = \sum_{j=1}^{n} z_j^2$$

Вычисление ошибки

- Отклонения прогнозов от ответов:

$$Xw - y = \begin{pmatrix} \langle w, x_1 \rangle - y_1 \\ \vdots \\ \langle w, x_\ell \rangle - y_\ell \end{pmatrix}$$

- Среднеквадратичная ошибка:

$$\frac{1}{\ell} \|Xw - y\|^2 = \frac{1}{\ell} \sum_{i=1}^{\ell} (\langle w, x_i \rangle - y_i)^2$$

Обучение линейной регрессии

$$\frac{1}{\ell}\|Xw - y\|^2 \to \min_w$$

- Вычисление MSE в NumPy:

```
np.square(X.dot(w) - y).mean()
```