



# ТЕХНОСФЕРА

## Лекция 5 Классификация и регрессия

Ксения Стройкова

24 октября 2016 г.

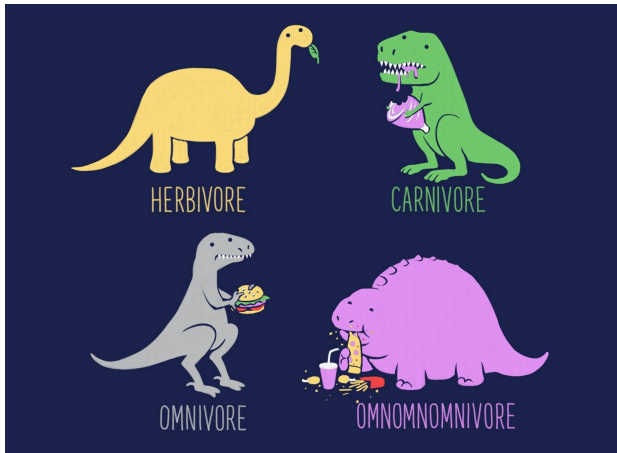
# План занятия

Задачи классификации и регрессии

Некоторые полезные идеи

Оценка качества классификации

# Задачи классификации и регрессии



# Классификация: интуиция

## Задача

Разработать алгоритм, позволяющий определить класс произвольного объекта из некоторого множества

- ▶ Дана *обучающая выборка*, в которой для каждого объекта известен класс

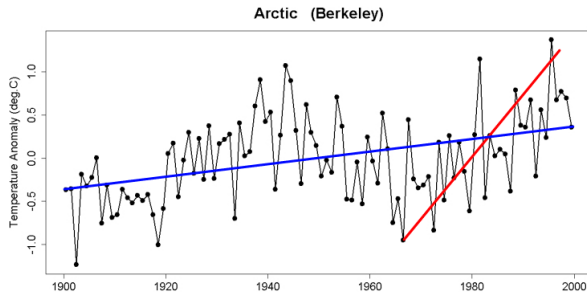


# Регрессия: интуиция

## Задача

Разработать алгоритм, позволяющий предсказать числовую характеристику произвольного объекта из некоторого множества

- ▶ Дана обучающая выборка, в которой для каждого объекта известно значение числовой характеристики



## Обучение с учителем / supervised learning

**Дано.** Признаковые описания  $N$  объектов  $\mathbf{x} = (x_1, \dots, x_m) \in \mathcal{X}$ , образующие тренировочный набор данных  $X$ , и значения целевой переменной  $y = f(\mathbf{x}) \in \mathcal{Y}$  для каждого объекта из  $X$ .

**Найти.** Для семейства параметрических функций

$$H = \{h(\mathbf{x}, \theta) = y : \mathcal{X} \times \Theta \rightarrow \mathcal{Y}\},$$

найти значение вектора параметров  $\theta^*$ , такое что  $h^*(\mathbf{x}) = h(\mathbf{x}, \theta^*)$  наилучшим образом приближает целевую функцию.

$\mathcal{Y} \in \{C_1, C_2, \dots, C_N\}$  – задача классификации

$\mathcal{Y} \in [a, b] \subset \mathcal{R}$  – задача регрессии

$$L = R + E + O$$

- R** Выдвигаем гипотезу насчет **модели** - семейства параметрических функций вида

$$H = \{h(\mathbf{x}, \theta) = y : \mathcal{X} \times \Theta \rightarrow \mathcal{Y}\},$$

которая могла бы решить нашу задачу (representation)

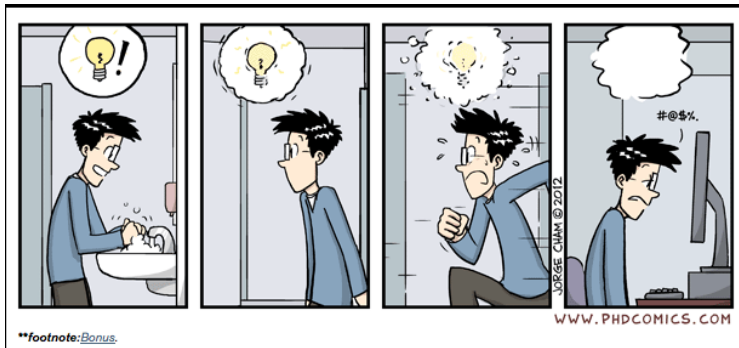
- E** Выбираем критерий, на основании которого будем оценивать качество предсказания (evaluation)
- O** Выбираем наилучшие параметры модели  $\theta^*$ , используя **алгоритм обучения**

$$A(X, Y) : (\mathcal{X}, \mathcal{Y})^N \rightarrow \Theta$$

(optimization)

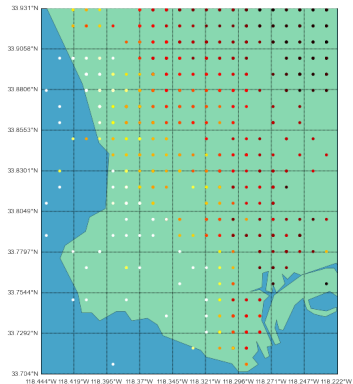
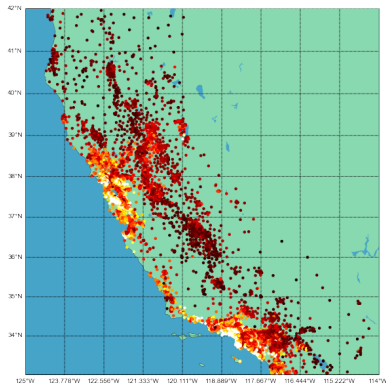
- D** Используя полученную модель  $h^*(\mathbf{x}) = h(\mathbf{x}, \theta^*)$ , решаем, как классифицировать неизвестные объекты (decision making)

## Некоторые полезные идеи





# Цены на недвижимость<sup>1</sup>

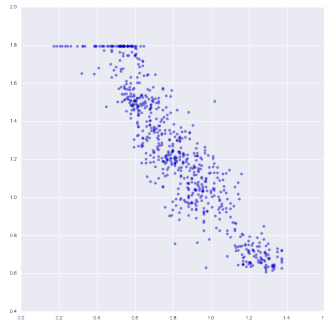
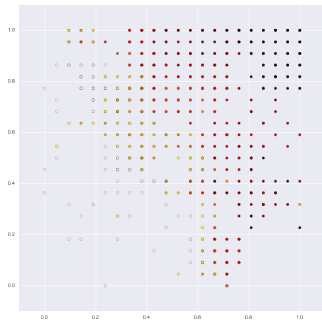


---

<sup>1</sup>California Housing data set

# Преобразование данных

- ▶ Нормализуем широту, долготу
- ▶ Логарифм от целевой переменной



# Метод ближайших соседей

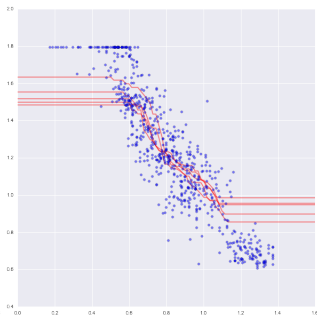
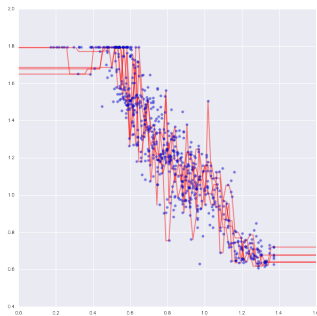
K-Nearest Neighbours

Representation:

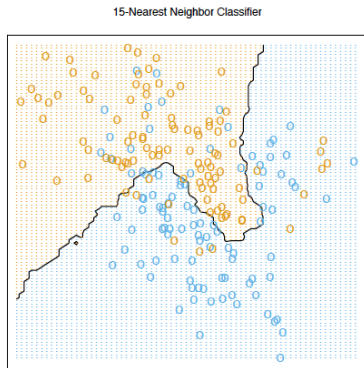
$$h(\mathbf{x}) = \frac{1}{K} \sum_{\mathbf{x}_k \in N_K(\mathbf{x})} f(\mathbf{x}_k)$$

Evaluation: любая

Optimization: не требуется



# Классификация с помощью метода ближайших соседей



## Representation: линейная модель

Идея: предположить, что искомая функция линейно зависит от признаков

$$y = h(\mathbf{x}, \mathbf{w}) = \sum_{d=1}^D x_d w_d + w_0 = \mathbf{x}^T \mathbf{w} + w_0$$

Добавим к  $\mathbf{x}$  фиктивный компонент  $x_0 = 1$

$$y = h(\mathbf{x}, \mathbf{w}) = \sum_{d=0}^D x_d w_d = \mathbf{x}^T \mathbf{w},$$

тогда для всего набора данных

$$\mathbf{Y} = \mathbf{X}\mathbf{w}$$

## Evaluation: метод наименьших квадратов

Идея: выбрать веса так, чтобы сумма квадратов отклонений предсказаний от реальных значений была минимальной

$$RSS(\mathbf{w}) = \sum_{n=1}^N (y_n - h(\mathbf{x}_n, \mathbf{w}))^2 = \sum_{n=1}^N (y_n - \mathbf{x}_n^T \mathbf{w})^2 \rightarrow \min_{\mathbf{w}}$$

## Optimization: аналитически

$$RSS(\mathbf{w}) = (Y - X^T \mathbf{w})^T (Y - X^T \mathbf{w})$$

$\Downarrow$

$$\mathbf{w} = (X^T X)^{-1} X^T Y$$



# Нелинейные зависимости

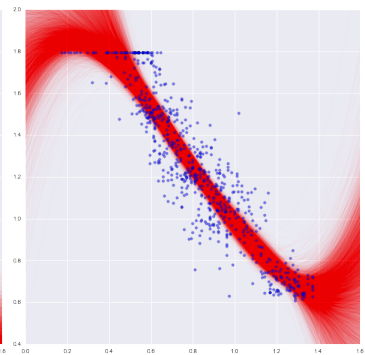
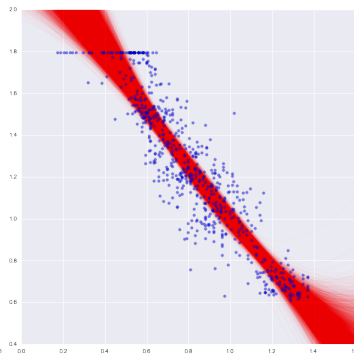
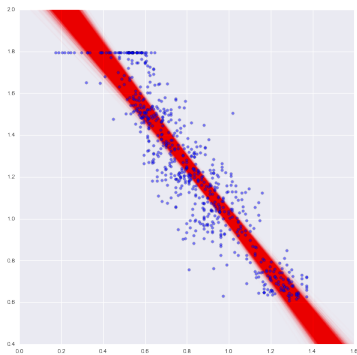
Перейдем в новое пространство признаков

$$\mathbf{x} = (x_1, x_2, \dots, x_D)$$



$$\mathbf{z} = (x_1, x_2, \dots, x_D, x_1^2, x_1x_2, x_1x_3, \dots, x_{D-1}x_D, x_D^2, \dots)$$

и сможем приближать сложные нелинейные функции



## Эмпирический риск

**Функция потерь**  $\mathcal{L}(\mathbf{x}, y, \theta)$  - ошибка, которую для данного  $\mathbf{x}$  дает модель  $h(\mathbf{x}, \theta)$  по сравнению с реальным значением  $y$

**Эмпирический риск** – средняя ошибка на обучающей выборке

$$Q(X, Y, \theta) = \frac{1}{N} \sum_{n=1}^N \mathcal{L}(\mathbf{x}_n, y_n, \theta)$$

**Задача** – найти значение  $\theta^*$ , минимизирующее эмпирический риск

$$\theta^* = \theta^*(X, Y) = \operatorname{argmin}_{\theta} Q(X, Y, \theta)$$

## Некоторые функции потерь

- ▶ Индикатор ошибки

$$\mathcal{L}(\mathbf{x}, y, \theta) = 0 \text{ if } h(\mathbf{x}, \theta) = y \text{ else } 1$$

- ▶ Функция Минковского

$$\mathcal{L}(\mathbf{x}, y, \theta) = |y - h(\mathbf{x}, \theta)|^q$$

Частные случаи: квадратичная  $q = 2$ , абсолютная ошибка  $q = 1$

- ▶ Hinge

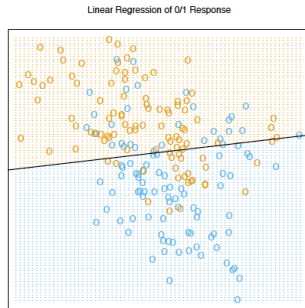
$$\mathcal{L}(\mathbf{x}, y, \theta) = \max(0, 1 - y \times h(\mathbf{x}, \theta))$$

- ▶ Информационная

$$\mathcal{L}(\mathbf{x}, y, \theta) = -\log_2 p(y|\mathbf{x}, \theta)$$

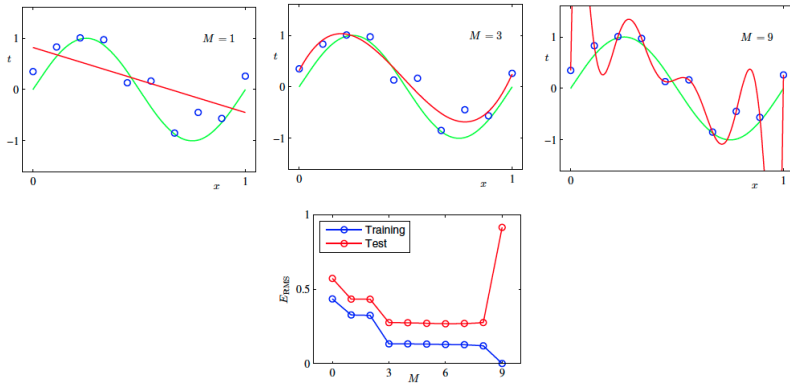
# Классификация с помощью метода наименьших квадратов

Пусть  $\mathcal{Y} = \{0, 1\}$ , тогда 
$$\begin{cases} \text{классифицируем 1, если } h^*(\mathbf{x}) \geq 0.5 \\ \text{классифицируем 0, если } h^*(\mathbf{x}) < 0.5 \end{cases}$$



# Проблема 1. Переобучение

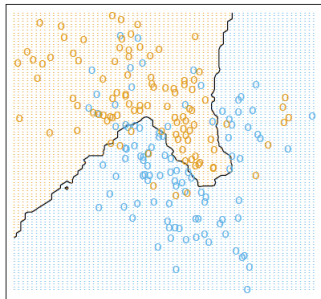
## Метод наименьших квадратов



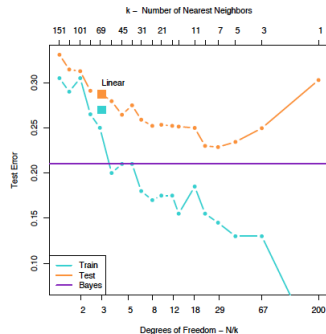
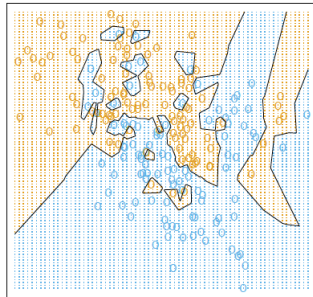
# Проблема 1. Переобучение

## KNN

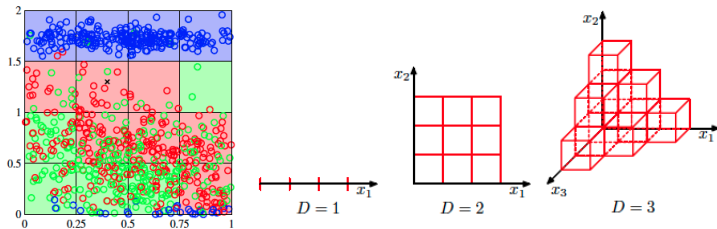
15-Nearest Neighbor Classifier



1-Nearest Neighbor Classifier



## Проблема 2. Проклятие размерности<sup>2</sup>



---

<sup>2</sup>Big Dimensions, and What You Can Do About It



## Оценка качества классификации



# Как оценить различные модели?

## Идея

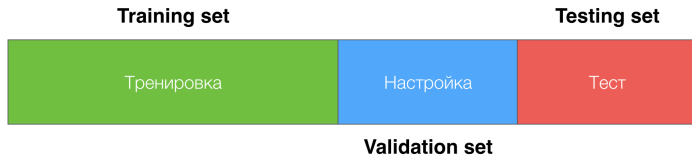
использовать долю неверно классифицированных объектов  
(error rate)

## Важное замечание

error rate на обучающей выборке **НЕ** является хорошим показателем качества модели

## Решение 1: разделение выборки

Делим обучающую выборку на **тренировочную**, **валидационную** и **тестовую**



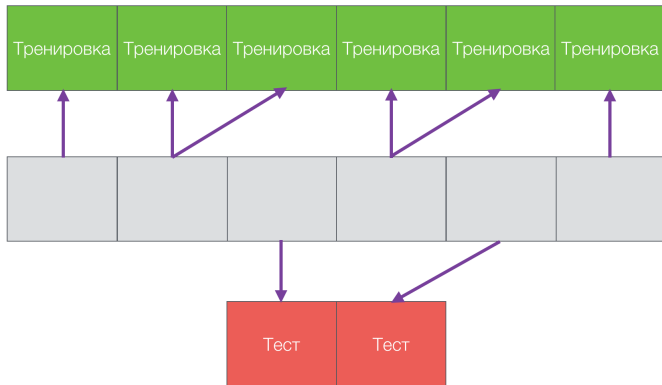
## Решение 2: скользящий контроль (n-times) (stratified) cross-validation



частный случай: leave-one-out

## Решение 3: bootstrap

выбираем в тренировочную выбоку  $n$  объектов с возвращением



## Доверительный интервал для success rate

При тестировании на  $N = 100$  объектах было получено 25 ошибок. Таким образом измеренная вероятность успеха (success rate) составила  $f = 0.75$ . Найти доверительный интервал для действительной вероятности успеха с уровнем доверия  $\alpha = 0.8$ .

### Решение

Пусть  $p$  – действительная вероятность успеха в испытаниях бернулли, тогда

$$f \sim \mathcal{N}(p, p(1-p)/N).$$

Воспользовавшись табличным значением  $P(-z \leq \mathcal{N}(0, 1) \leq z) = \alpha$ , имеем

$$P\left(-z \leq \frac{f - p}{\sqrt{p(1-p)/N}} \leq z\right) = \alpha,$$

откуда

$$p \in \left(f + \frac{z^2}{2N} \pm z \sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right) / \left(1 + \frac{z^2}{N}\right) = [0.69, 0.80]$$

# Метрики качества. Вероятностные модели.

Пусть  $y_i$  - действительный класс для объекта  $\mathbf{x}_i$

- ▶ Information loss

$$-\frac{1}{N} \sum_i \log_2 p(y_i | \mathbf{x}_i)$$

- ▶ Quadratic loss

$$\frac{1}{N} \sum_j (p(y_j | \mathbf{x}_i) - a_j(\mathbf{x}_i))^2,$$

где

$$a_j(\mathbf{x}_i) = \begin{cases} 1, & \text{если } C_j = y_i \\ 0, & \text{иначе} \end{cases}$$

## Метрики качества. Функции решения.

		Предсказанный	
		true	false
Действительный	true	TP	FN
	false	FP	TN

$$success\ rate = accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$recall = TPR = \frac{TP}{TP + FN}; \quad precision = \frac{TP}{TP + FP}$$

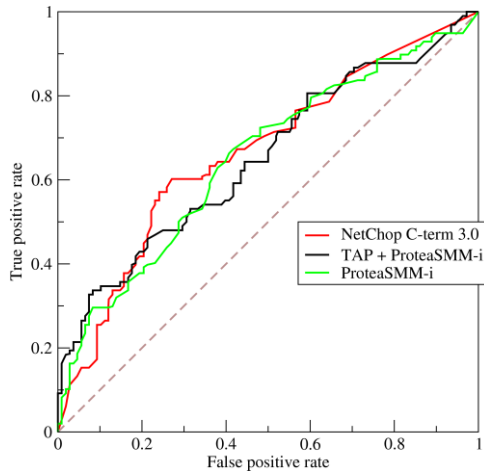
$$FPR = \frac{FP}{FP + TN}$$

$$affinity = lift = \frac{precision}{p}$$



# Receiver Operating Characteristic

$$TPR = \frac{TP}{TP + FN}; \quad FPR = \frac{FP}{FP + TN}$$



## Метрики качества. Регрессия

$$MSE = \frac{1}{N} \sum (h(\mathbf{x}_i) - y_i)^2, \quad RMSE = \sqrt{MSE}$$

$$MAE = \frac{1}{N} \sum |h(\mathbf{x}_i) - y_i|, \quad RMAE = \sqrt{MAE}$$

$$RSE = \frac{\sum (h(\mathbf{x}_i) - y_i)^2}{\sum (y_i - \bar{y})^2}$$

$$correlation = \frac{S_{hy}}{\sqrt{S_h S_y}}; \quad S_{yh} = \frac{\sum (h(i) - \overline{h(i)})(y_i - \bar{y})}{N - 1}$$

$$S_h = \frac{\sum (h(i) - \overline{h(i)})^2}{N - 1}; \quad S_y = \frac{\sum (y_i - \bar{y})^2}{N - 1}$$

NFLT, MDL, AIC и все такое

### No free lunch theorem

Не существует единственной лучшей модели, решающей все задачи

### Minimum description length

Лучшая гипотеза о данных – та, которая ведет к самому краткому их описанию

### Akaike information criterion (AIC)

$$model = \arg \max \ln p(\mathcal{D}|\theta_{ML}) - \|\theta\|$$

Вопросы

