Методы машинного обучения

Лекция 3

Операции в векторных пространствах. Оценки обобщающей способности. Метод k ближайших соседей.

Эльвира Зиннурова

elvirazinnurova@gmail.com

НИУ ВШЭ, 2019

Напоминание

- $\mathbb{X} = \mathbb{R}^d$ — пространство объектов, $\mathbb{Y}$ — пространство ответов
- $X = (x_i, y_i)_{i=1}^{\ell}$ — обучающая выборка
- $Q(a, X)$ — функционал ошибки алгоритма a на выборке X
- Обучение — поиск $a(x) = \arg\min_{a \in \mathcal{A}} Q(a, X)$
- kNN: обучение как таковое отсутствует, предсказываем самый популярный среди соседей объекта класс
- «Близость» объектов — евклидова метрика

Операции в векторных пространствах

Евклидово пространство

- Векторное пространство — множество элементов, для которых определены операции:
  - сложения друг с другом
  - умножения на число

1. $\mathbf{x}+\mathbf{y}=\mathbf{y}+\mathbf{x}$, для любых $\mathbf{x},\mathbf{y}\in V$ (*коммутативность сложения*);
2. $\mathbf{x}+(\mathbf{y}+\mathbf{z})=(\mathbf{x}+\mathbf{y})+\mathbf{z}$, для любых $\mathbf{x},\mathbf{y},\mathbf{z}\in V$ (*ассоциативность сложения*);
3. существует такой элемент $\mathbf{0}\in V$, что $\mathbf{x}+\mathbf{0}=\mathbf{0}+\mathbf{x}=\mathbf{x}$ для любого $\mathbf{x}\in V$ (*существование нейтрального элемента относительно сложения*), называемый **нулевым вектором** или просто **нулём** пространства V;
4. для любого $\mathbf{x}\in V$ существует такой элемент $-\mathbf{x}\in V$, что $\mathbf{x}+(-\mathbf{x})=\mathbf{0}$, называемый вектором, **противоположным** вектору $\mathbf{x}$;
5. $\alpha(\beta\mathbf{x})=(\alpha\beta)\mathbf{x}$ (*ассоциативность умножения на скаляр*);
6. $1\cdot\mathbf{x}=\mathbf{x}$ (*унитарность: умножение на нейтральный (по умножению) элемент поля F сохраняет вектор*).
7. $(\alpha+\beta)\mathbf{x}=\alpha\mathbf{x}+\beta\mathbf{x}$ (*дистрибутивность умножения вектора на скаляр относительно сложения скаляров*);
8. $\alpha(\mathbf{x}+\mathbf{y})=\alpha\mathbf{x}+\alpha\mathbf{y}$(*дистрибутивность умножения вектора на скаляр относительно сложения векторов*).

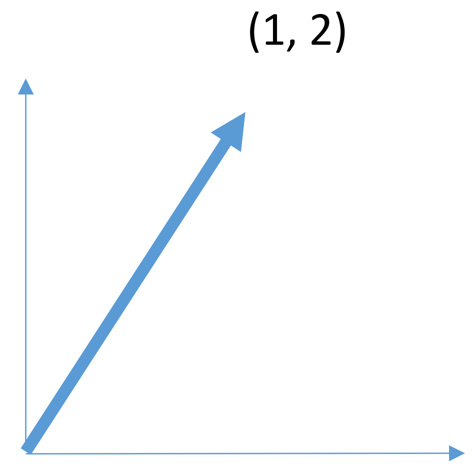
Евклидово пространство

- Векторное пространство — множество элементов, для которых определены операции:
  - сложения друг с другом
  - умножения на число
- Пример: пространство наборов из d вещественных чисел — евклидово пространство $\mathbb{R}^d$
- Бывают пространства с более сложными элементами: многочленами, уравнениями, функциями

Евклидово пространство

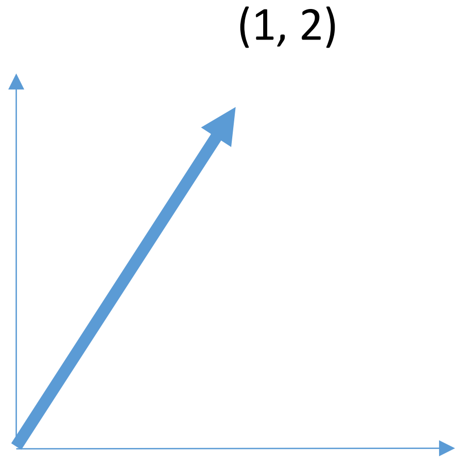
- Сложение и умножение на число — покоординатно
- Сложение
  - $a = (a_1, \dots, a_d)$
  - $b = (b_1, \dots, b_d)$
  - $a + b = (a_1 + b_1, \dots, a_d + b_d)$
- Умножение на число:
  - $a = (a_1, \dots, a_d)$
  - $\beta \in \mathbb{R}$
  - $\beta a = (\beta a_1, \dots, \beta a_d)$

- Вектор — точка и стрелка, идущая к ней из нуля

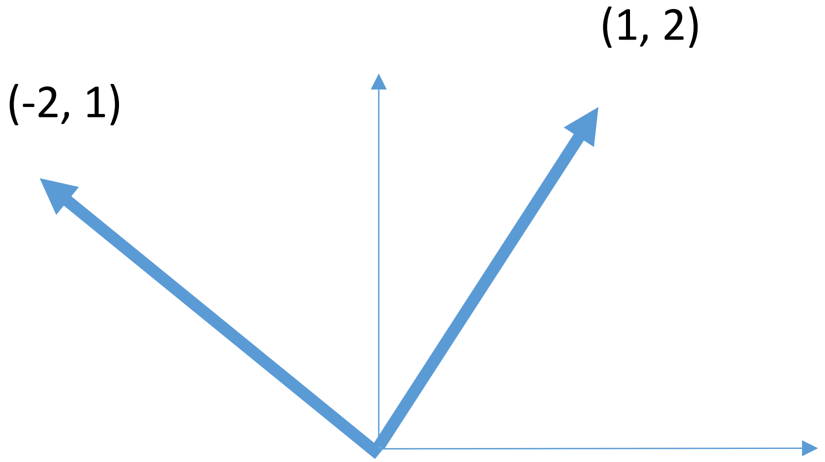


- Длина вектора:

$$\sqrt{1^2 + 2^2} = \sqrt{5}$$

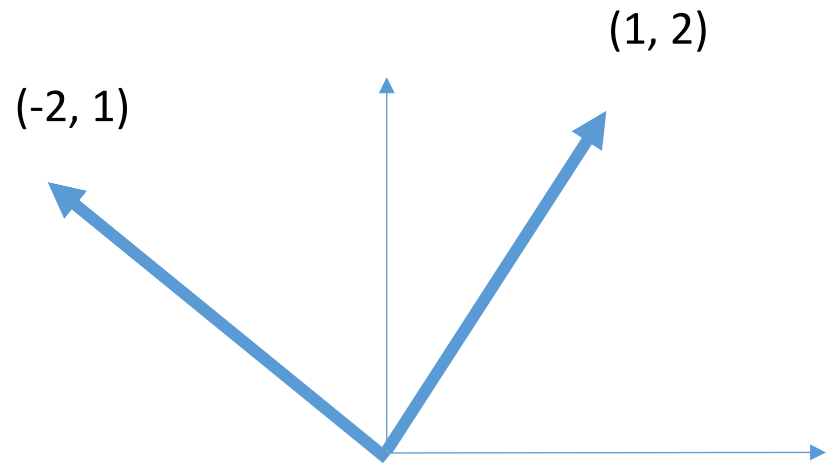


- Можем измерить угол с помощью транспортира: 90 градусов



- Можем измерить расстояние между точками:

$$\sqrt{\left(1-(-2)\right)^2+(2-1)^2}=\sqrt{10}$$



Норма

- Обобщение понятия длины вектора
- Функция $\|x\|$ от вектора
- Если $\|x\| = 0$, то $x = 0$
- $\|x + y\| \leq \|x\| + \|y\|$
- $\|\alpha x\| = |\alpha| \, \|x\|$

- Векторное пространство с нормой — нормированное

Примеры норм

- Евклидова норма:

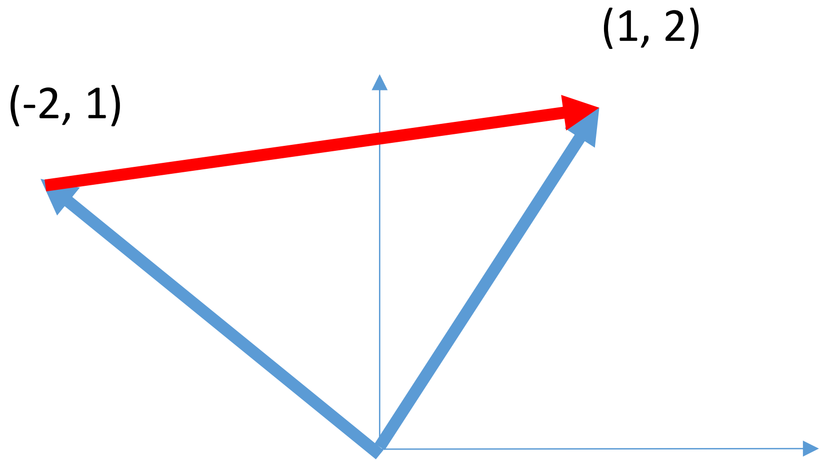
$$\|x\|_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$$

- Манхэттенская норма:

$$\|x\|_1 = \sum_{i=1}^{d} |x_i|$$

Метрика

- Обобщение понятия расстояния
- $\rho(x, y) = \|x - y\|$
- Соответствует геометрическим представлениям
- Векторное пространство с метрикой — метрическое



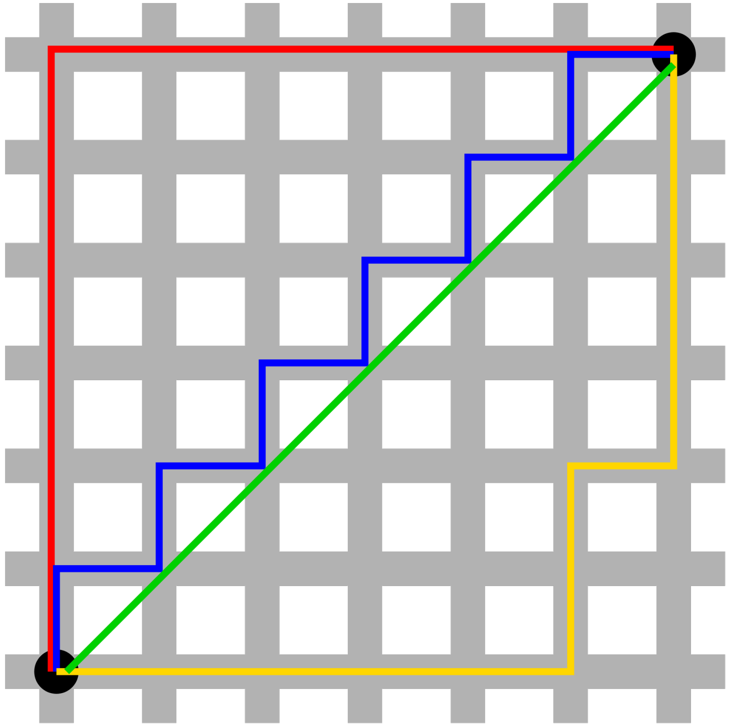
Примеры метрик

- Евклидова метрика:

$$\rho_2(x, z) = \sqrt{\sum_{i=1}^{d} (x_i - z_i)^2}$$

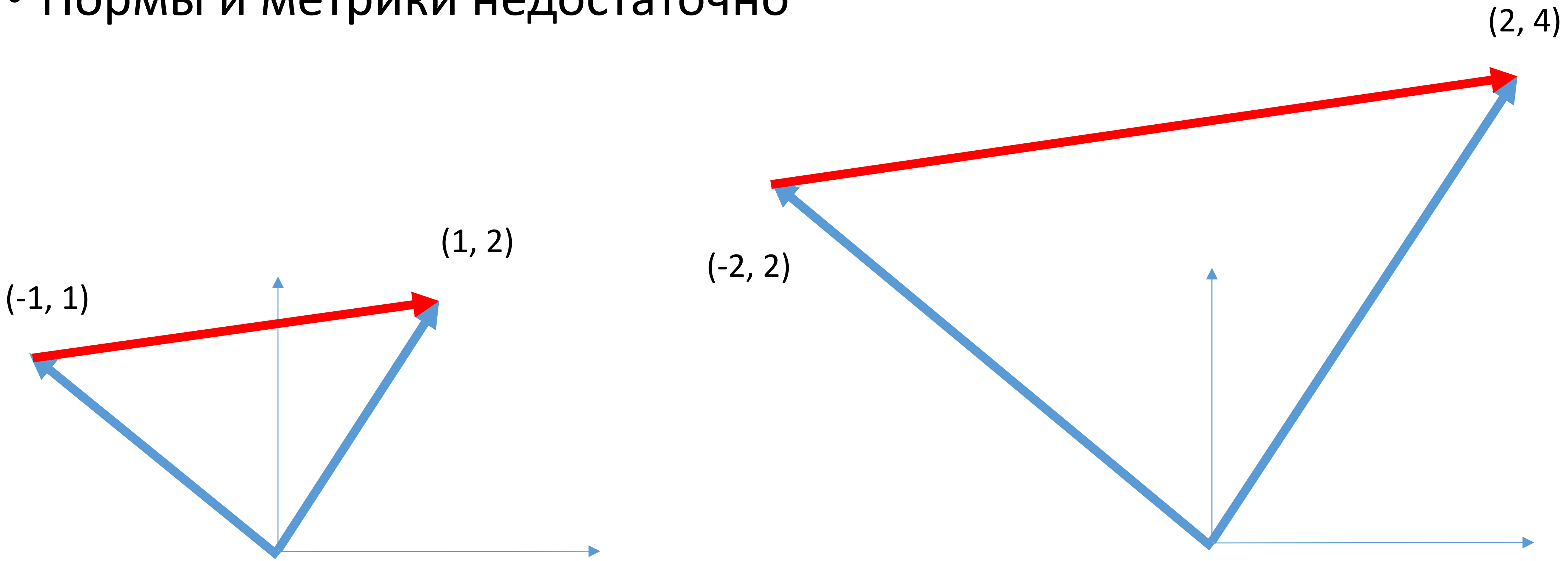
- Манхэттенская метрика:

$$\rho_1(x, z) = \sum_{i=1}^{d} |x_i - z_i|$$



Как искать углы?

- Нормы и метрики недостаточно



Скалярное произведение

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

Скалярное произведение

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

- Норма: $\|x\|_2 = \sqrt{\langle x, x \rangle}$

Скалярное произведение

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$$

- Норма: $\|x\|_2 = \sqrt{\langle x, x\rangle}$
- Расстояние: $\rho_2(x, z) = \|x - z\| = \sqrt{\langle x - z, x - z\rangle}$

Скалярное произведение

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

- Норма: $\|x\|_2 = \sqrt{\langle x, x \rangle}$
- Расстояние: $\rho_2(x, z) = \|x - z\| = \sqrt{\langle x - z, x - z \rangle}$
- Угол?

Скалярное произведение

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

- Важное соотношение: $\langle x, y \rangle = \|x\| \|y\| \cos \angle(x, y)$

Скалярное произведение

$$\langle x, y\rangle = \sum_{i=1}^{n} x_i y_i$$

- Важное соотношение: $\langle x, y\rangle = \|x\|\|y\| \boxed{\cos\angle(x, y)}$

Косинус угла

Скалярное произведение

- Косинус угла: $\cos \angle(x, y) = \frac{\langle x, y \rangle}{\|x\| \|y\|}$

Скалярное произведение

- Косинус угла: $\cos\angle(x,y) = \frac{\langle x,y\rangle}{\|x\|\|y\|}$
- Мера сонаправленности векторов
- Для параллельных векторов $\cos\angle(x,y) = 1$
- Для перпендикулярных векторов $\cos\angle(x,y) = 0$

Функционал ошибки для классификации

Ошибка классификации

- Доля неправильных ответов:

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

- Нотация Айверсона:
  - $[\text{истина}] = 1$
  - $[\text{ложь}] = 0$

Ошибка классификации

$a(x)$	y
-1	-1
$+1$	$+1$
-1	-1
$\mathbf{+1}$	$\mathbf{-1}$
$+1$	$+1$

- Доля неправильных ответов:

$$?$$

Ошибка классификации

$\boldsymbol{a(x)}$	$\boldsymbol{y}$
-1	-1
$+1$	$+1$
-1	-1
$+\mathbf{1}$	$-\mathbf{1}$
$+1$	$+1$

- Доля неправильных ответов:

$$\frac{1}{5} = 0.2$$

Accuracy

- Доля правильных ответов:

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

- На английском: accuracy

Accuracy

- Доля правильных ответов:

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

- На английском: accuracy
- ВАЖНО: не переводите это как «точность»!

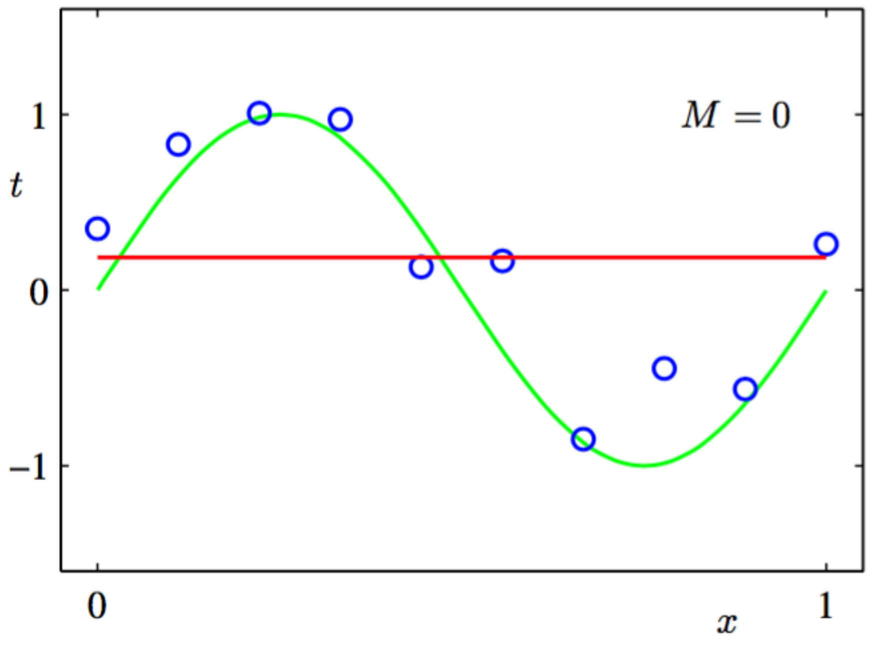
Обобщающая способность

Обобщающая способность

- Выбираем алгоритм с лучшим качеством на обучающей выборке
- Как он будет вести себя на новых данных?
- Смог ли он выразить y через x?

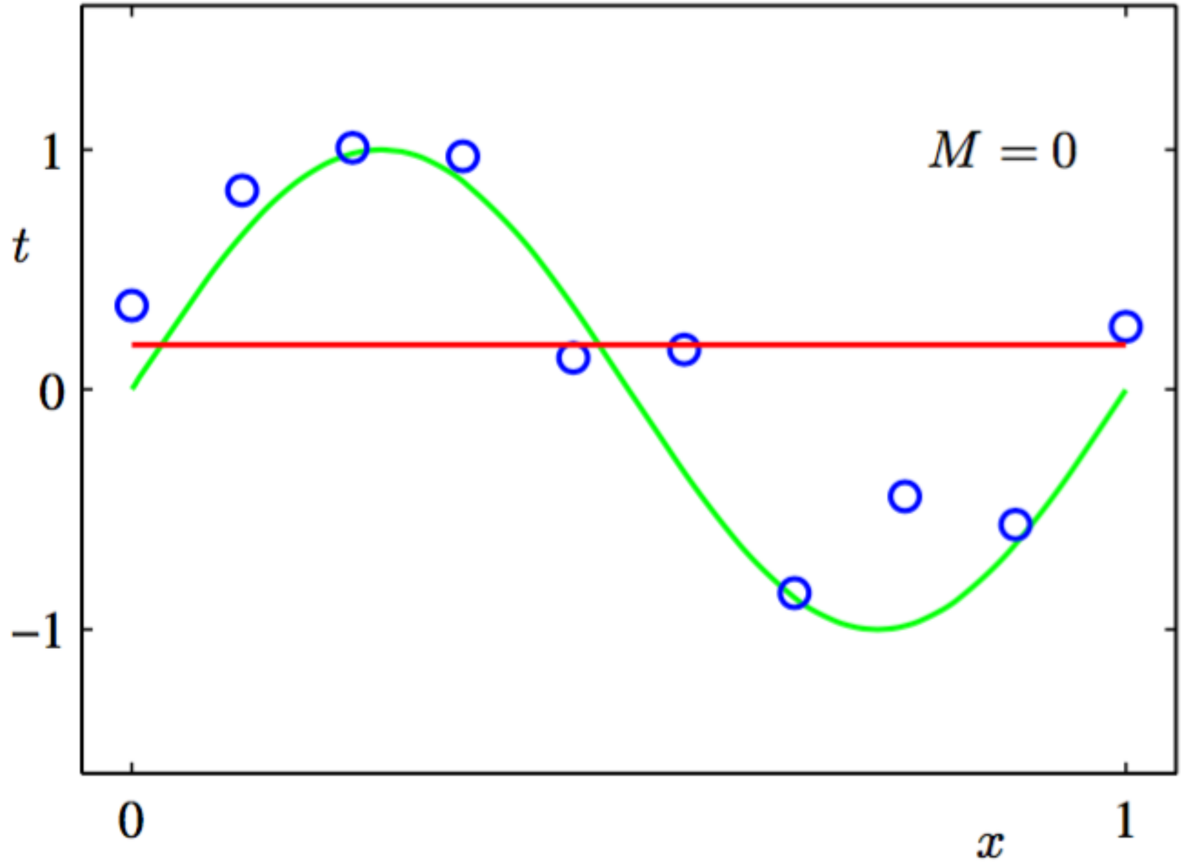
Обобщающая способность

- Зеленый — истинная зависимость
- Красный — прогноз алгоритма
- Синий — выборка
- Линейный алгоритм



Обобщающая способность

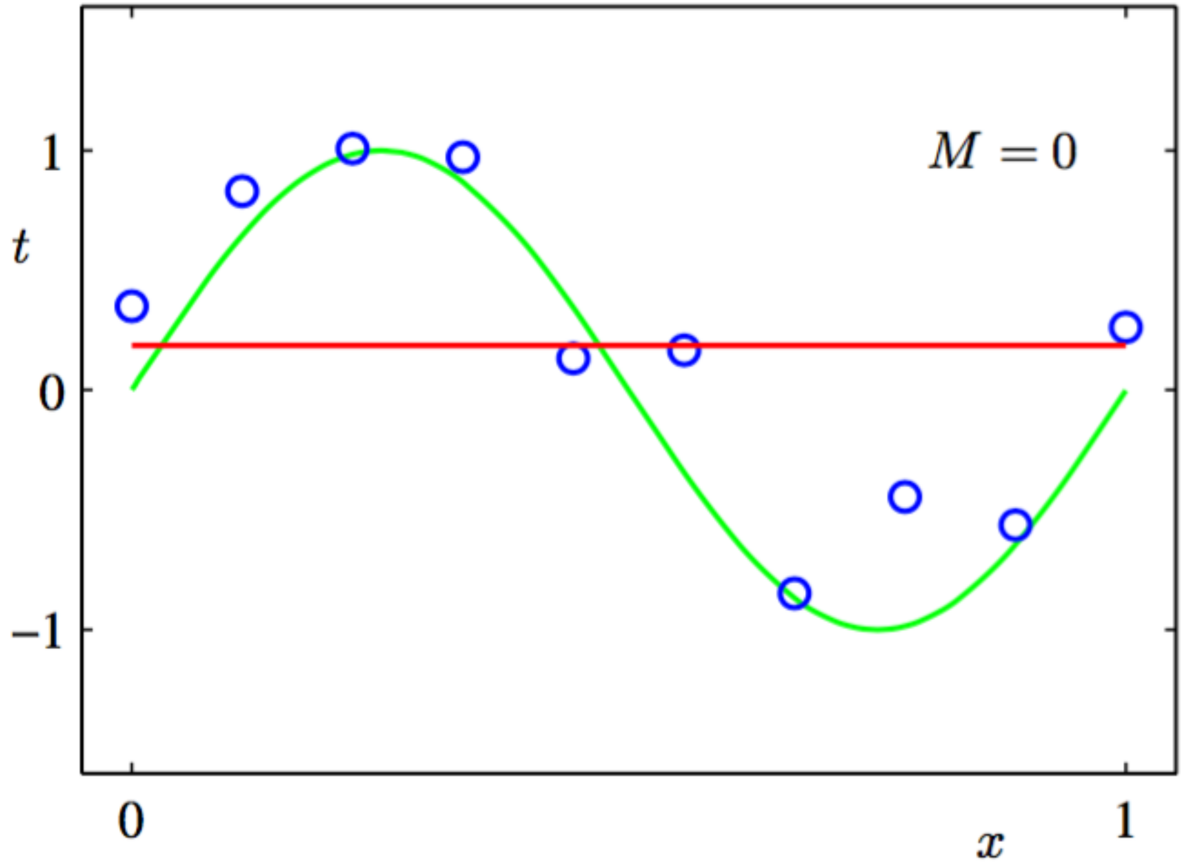
- Без признаков
- Константный алгоритм
- $a(x) = w_0$



Обобщающая способность

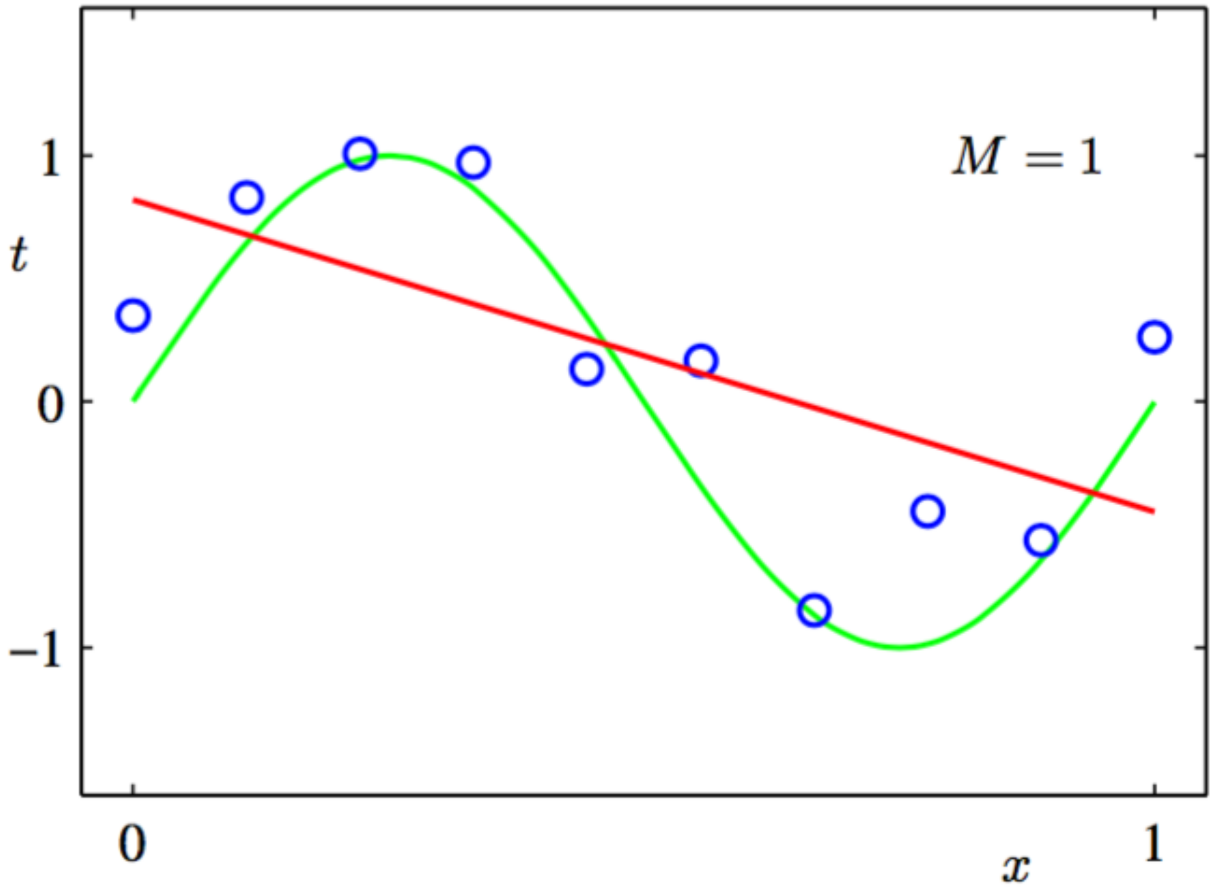
- Без признаков
- Константный алгоритм
- $a(x) = w_0$

Недообучение



Обобщающая способность

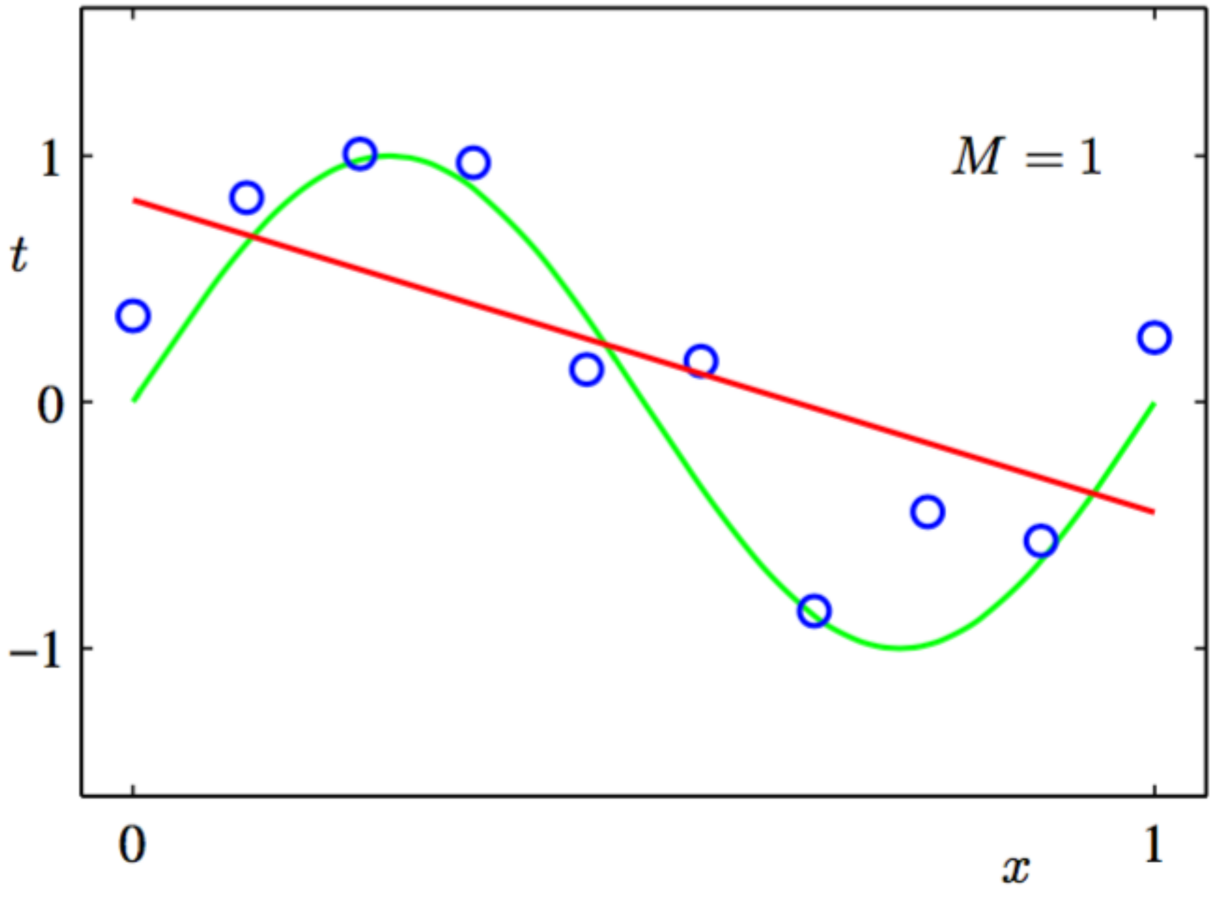
- 1 признак
- x
- $\mathrm{a}(x) = w_0 + w_1 x$



Обобщающая способность

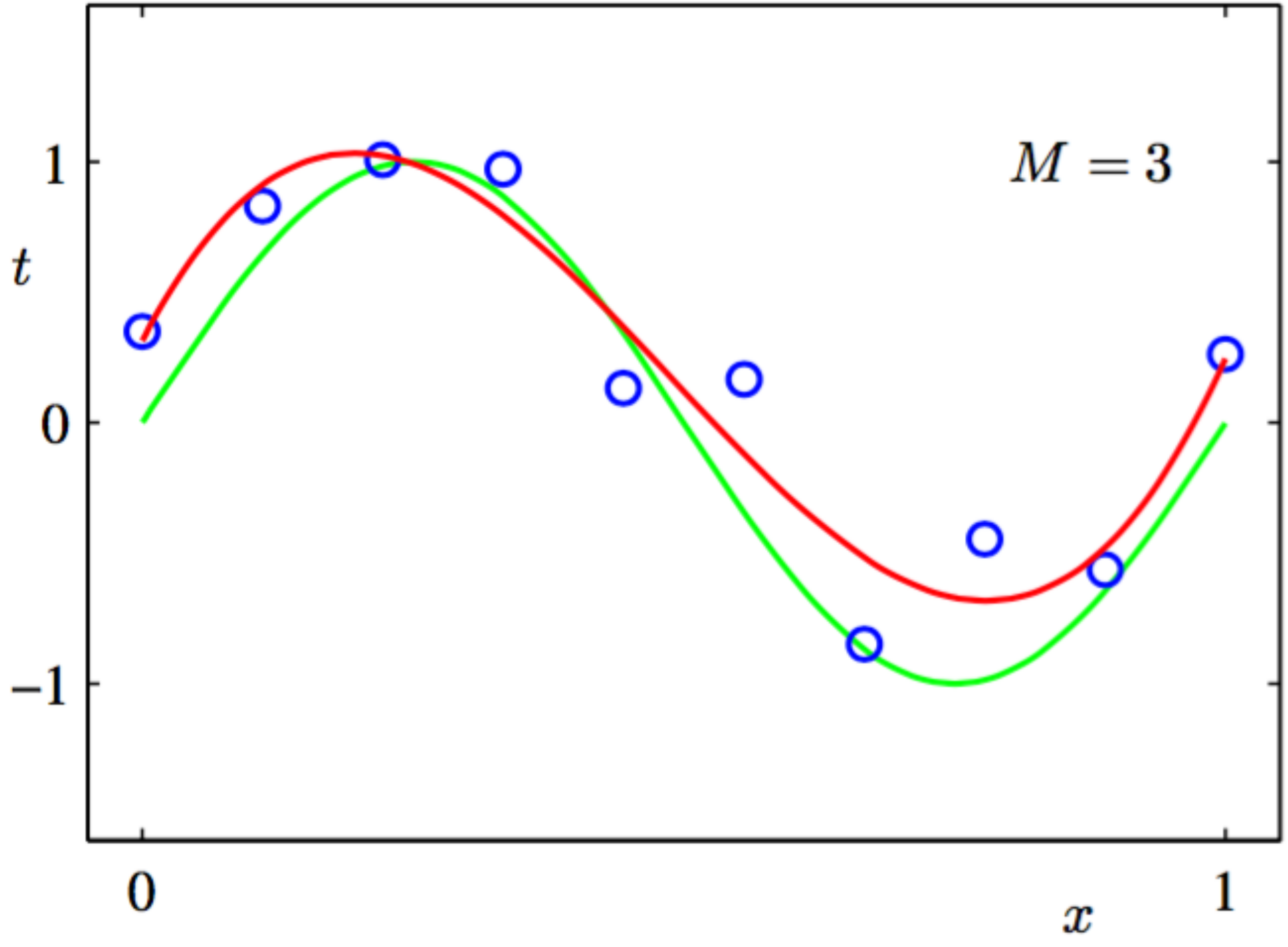
- 1 признак
- x
- $a(x) = w_0 + w_1 x$

Недообучение



Обобщающая способность

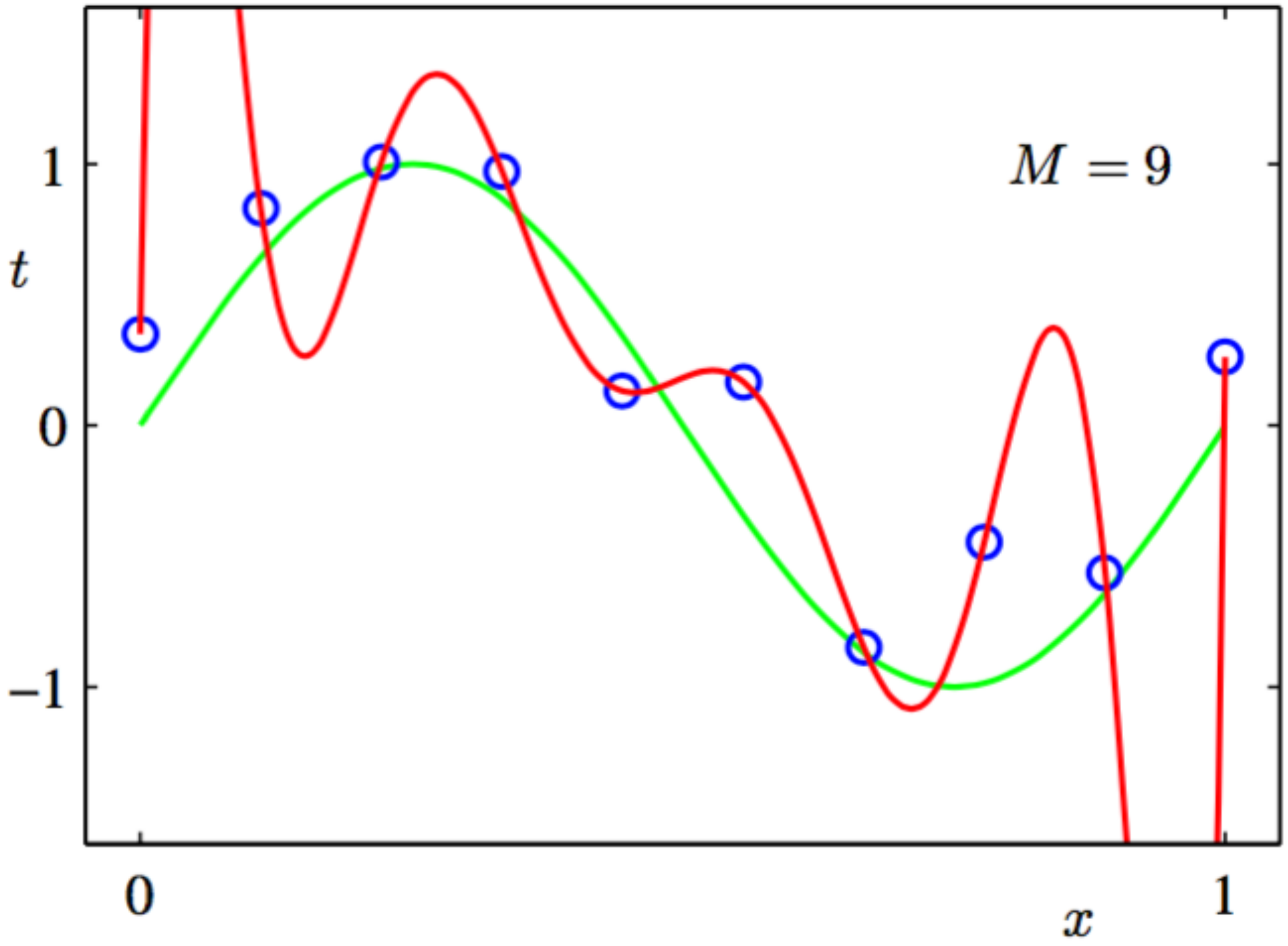
- 3 признака
- x, x^2, x^3
- $a(x) = w_0 + w_1 x + w_2 x^2 + w_3 x^3$



Обобщающая способность

- 9 признаков
- $x, x^2, x^3, x^4, \dots, x^9$
- $\mathrm{a}(x) = w_0 + w_1 x + \cdots + w_9 x^9$

Переобучение (overfitting)



Обобщающая способность

- Недообучение — плохое качество на обучении и на новых данных
- Переобучение — хорошее качество на обучении, плохое на новых данных

- Переобучение — алгоритм запоминает ответы, а не находит закономерности

Как выявить переобучение?

- Хороший алгоритм — хорошее качество на обучении
- Переобученный алгоритм — хорошее качество на обучении
- По обучающей выборке очень сложно выявить переобучение

Как выявить переобучение?

- Нужен способ оценки качества на новых данных, а не только на обучающей выборке

Оценивание обобщающей способности

Как оценить качество?

- Как алгоритм будет вести себя на новых данных?
- Какая у него будет доля ошибок?
- …или другая метрика качества
- По обучающей выборке нельзя это оценить

Отложенная выборка

- Разбиваем выборку на две части
  - Обучающая выборка
  - Отложенная выборка
- На первой обучаем алгоритм
- На второй измеряем качество

Пропорции разбиения

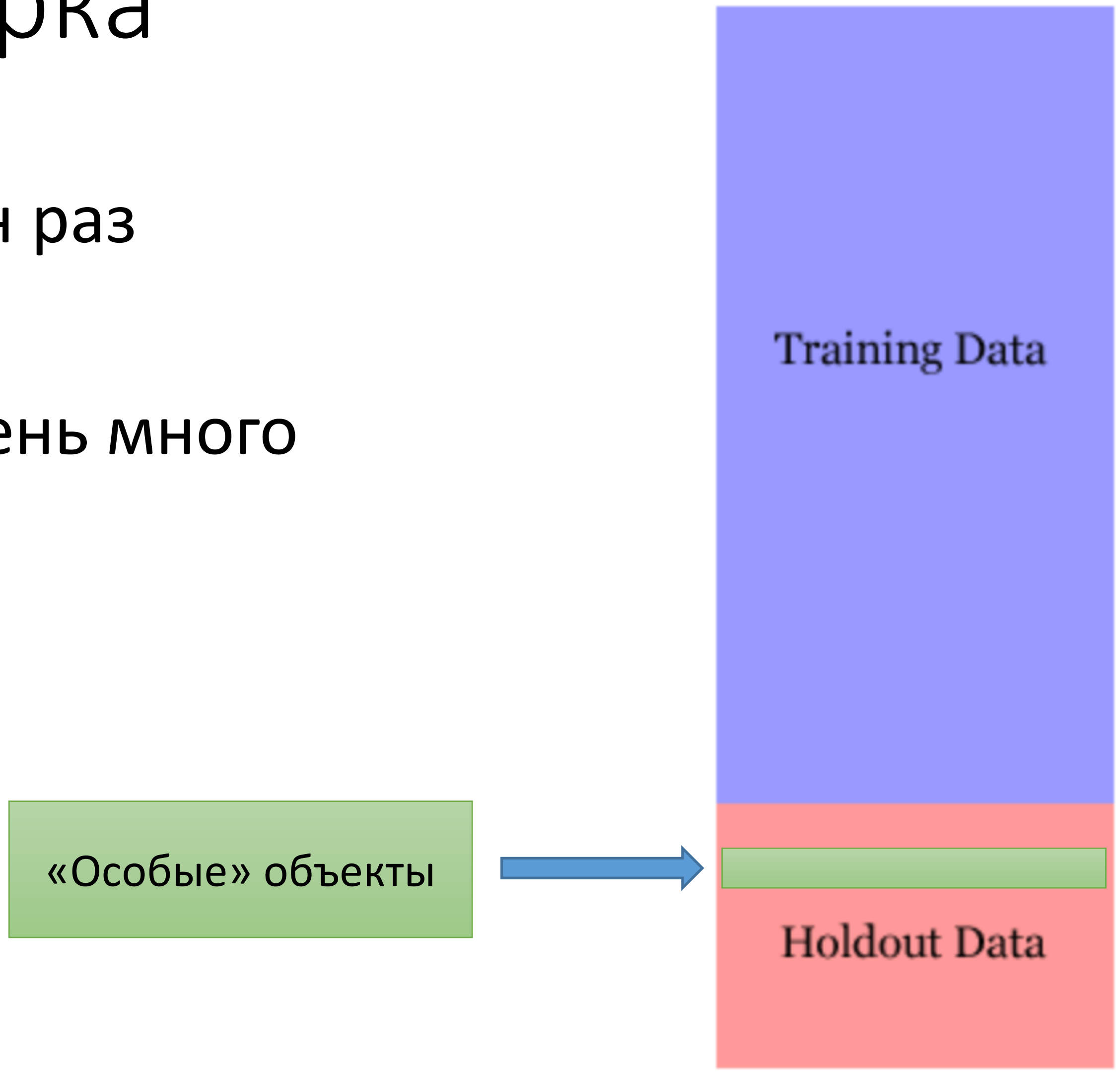
- Маленькая отложенная часть
  - (+) Обучающая выборка репрезентативная
  - (-) Оценка качества ненадежная
- Большая отложенная часть
  - (+) Оценка качества надежная
  - (-) Оценка качества смещенная
- Обычно: 70/30, 80/20, 0.632/0.368

Отложенная выборка

- (+) Обучаем алгоритм один раз
- (-) Зависит от разбиения
- Подходит, если данных очень много

Отложенная выборка

- (+) Обучаем алгоритм один раз
- (-) Зависит от разбиения
- Подходит, если данных очень много



Много отложенных выборок

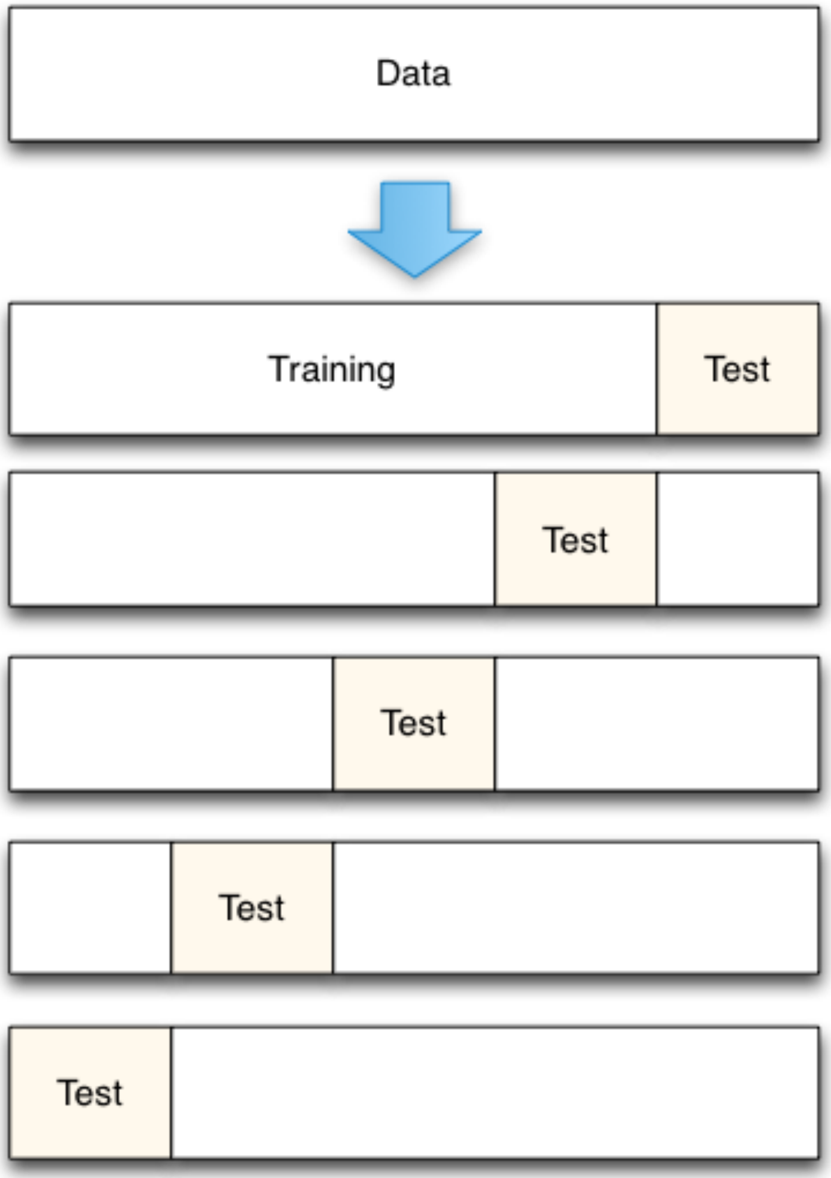
- Улучшение: разбиваем выборку на две части n раз
- Усредняем оценку качества

Много отложенных выборок

- Нет гарантий, что каждый объект побывает в обучении

Кросс-валидация

- Разбиваем выборку на k блоков
- Каждая по очереди выступает как тестовая



Число блоков

- Мало блоков
  - Тестовая выборка всегда большая — (+) надежные оценки
  - Обучение маленькое — (-) смещенные оценки
- Много блоков
  - (-) Ненадежные оценки
  - (+) Несмещенные оценки

Число блоков

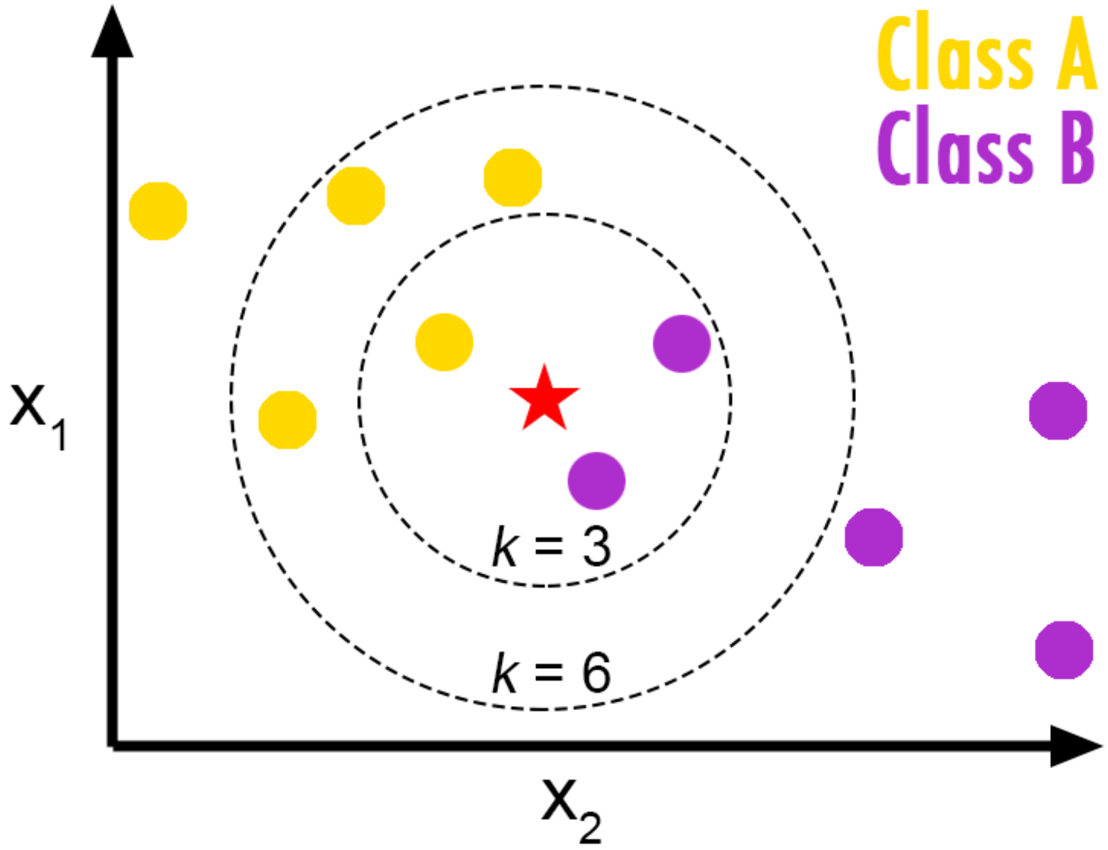
- Обычно: $k = 3, 5, 10$
- Чем больше выборка, тем меньше нужно k
- Чем больше k, тем больше раз надо обучать алгоритм
- Крайний случай $k = l$ называется leave-one-out оценкой

Совет

- Перемешивайте выборку!
- Объекты могут быть отсортированы
- При разбиении в обучении могут оказаться только мальчики, в контроле — только девочки

Метрические методы классификации

Метод k ближайших соседей



kNN с весами

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

Варианты:

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

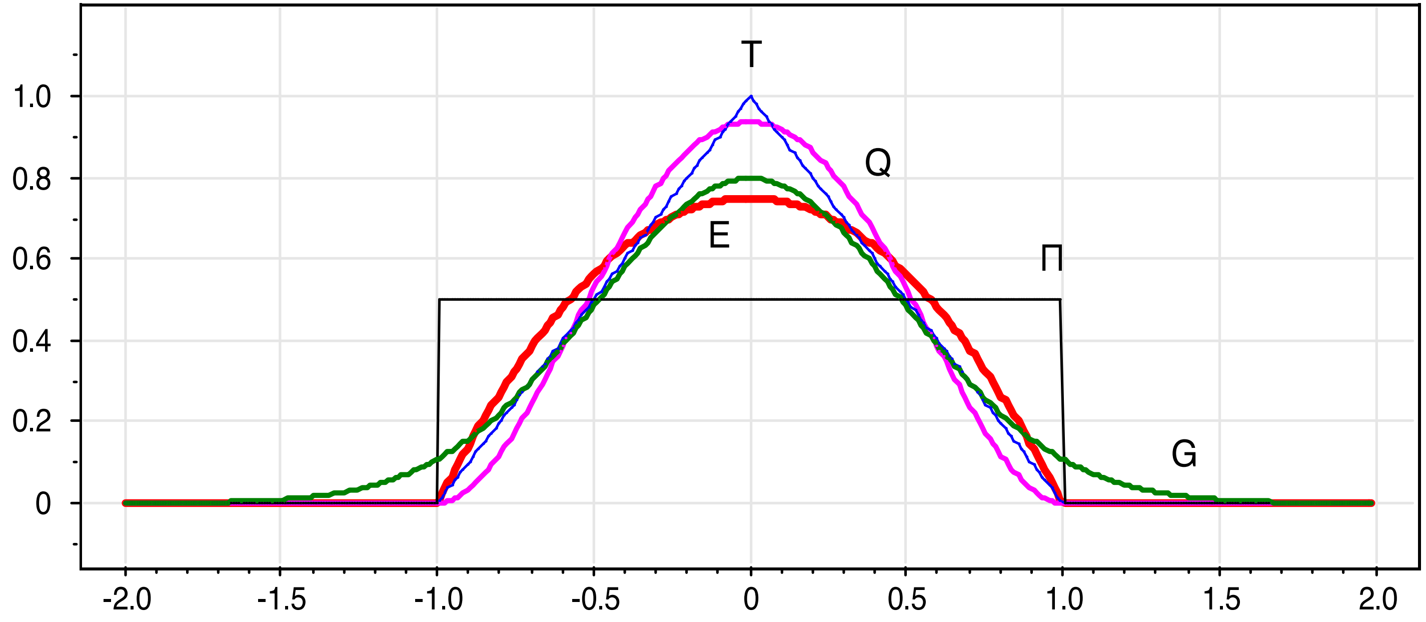
Все еще учитывается **номер** соседа, а не само расстояние до него

kNN с весами

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

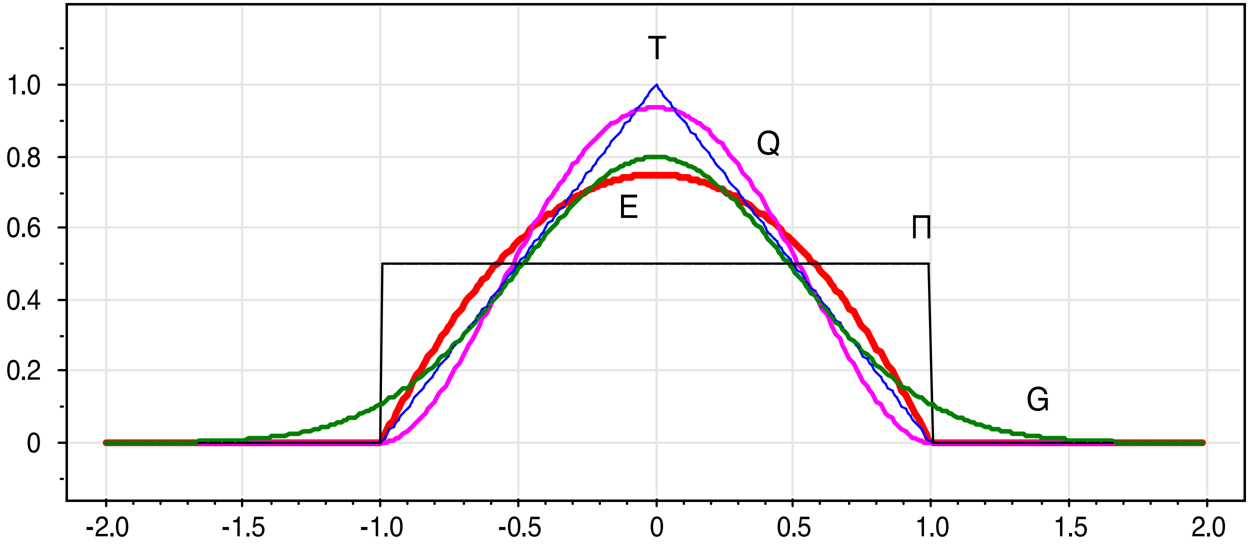
Парзеновское окно:

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

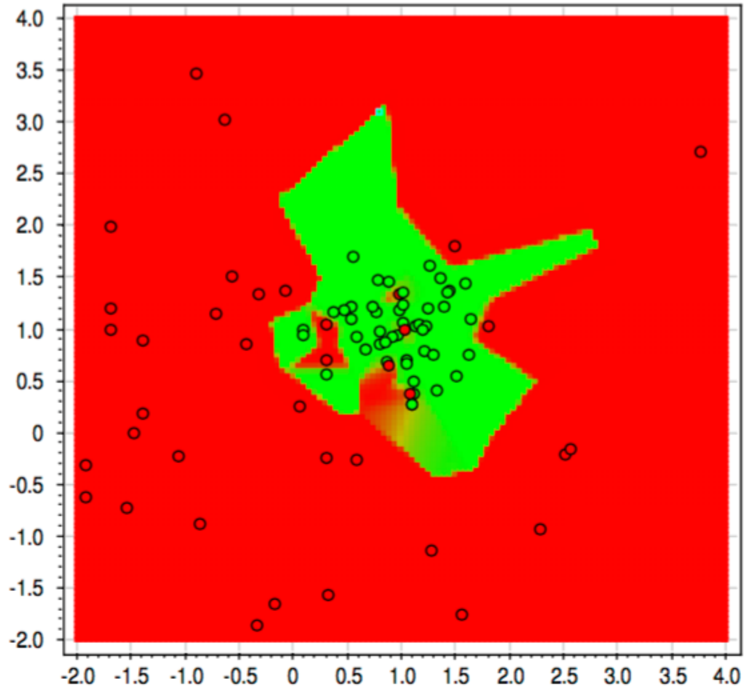


Ядра

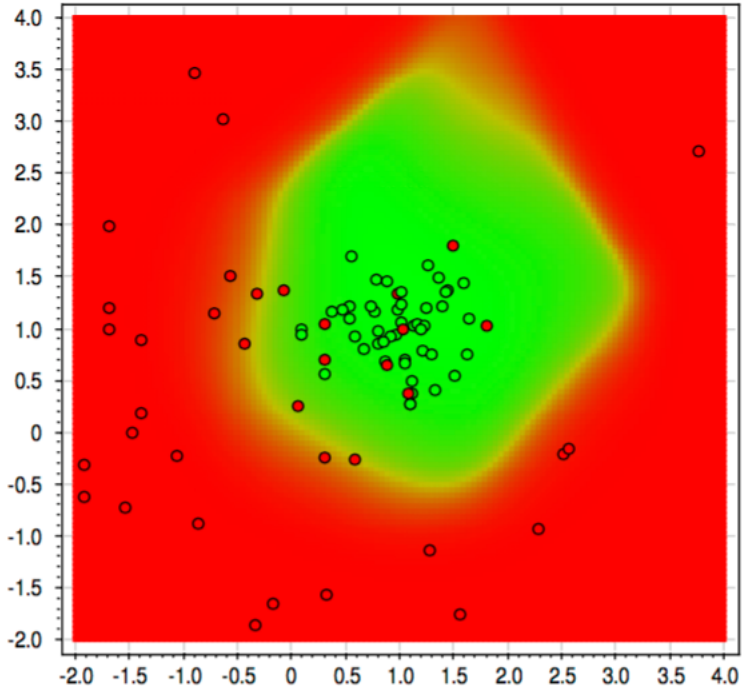
- Гауссовское ядро: $K(z) = (2\pi)^{-0.5} \exp\left(-\frac{1}{2} z^2\right)$
- И много других



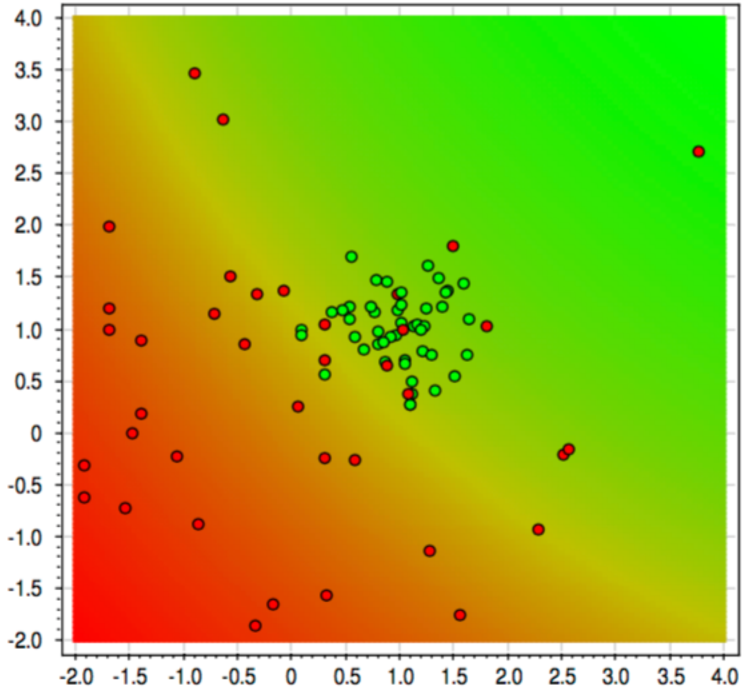
Ядра



$h = 0.05$



$h = 0.5$



$h = 5$

kNN для регрессии

- Классификация:

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

- Регрессия:

kNN для регрессии

- Классификация:

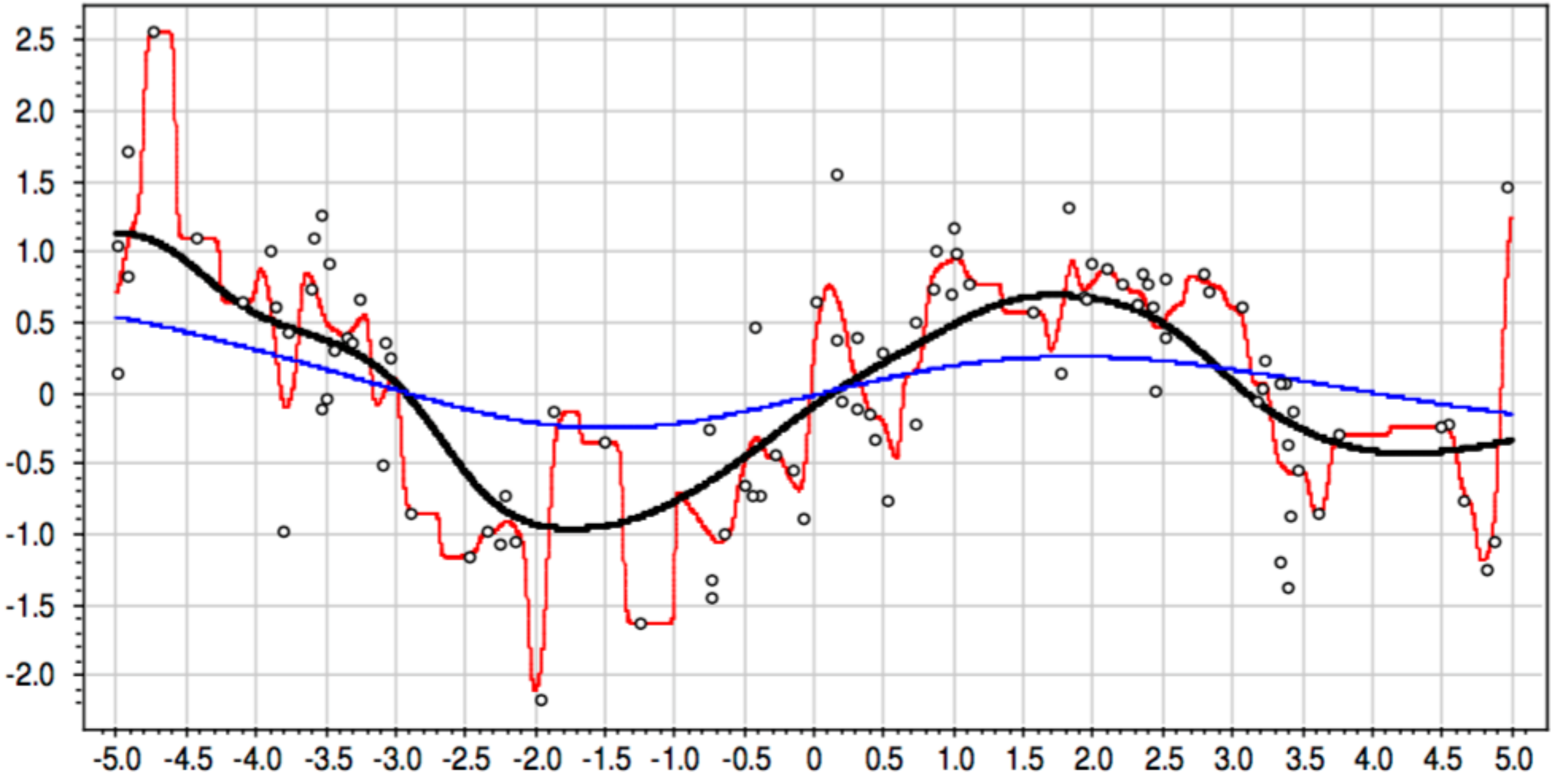
$$a(x) = \arg\max_{y \in \mathbb{Y}} \sum_{i=1}^{k} w_i [y_{(i)} = y]$$

- Регрессия:

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

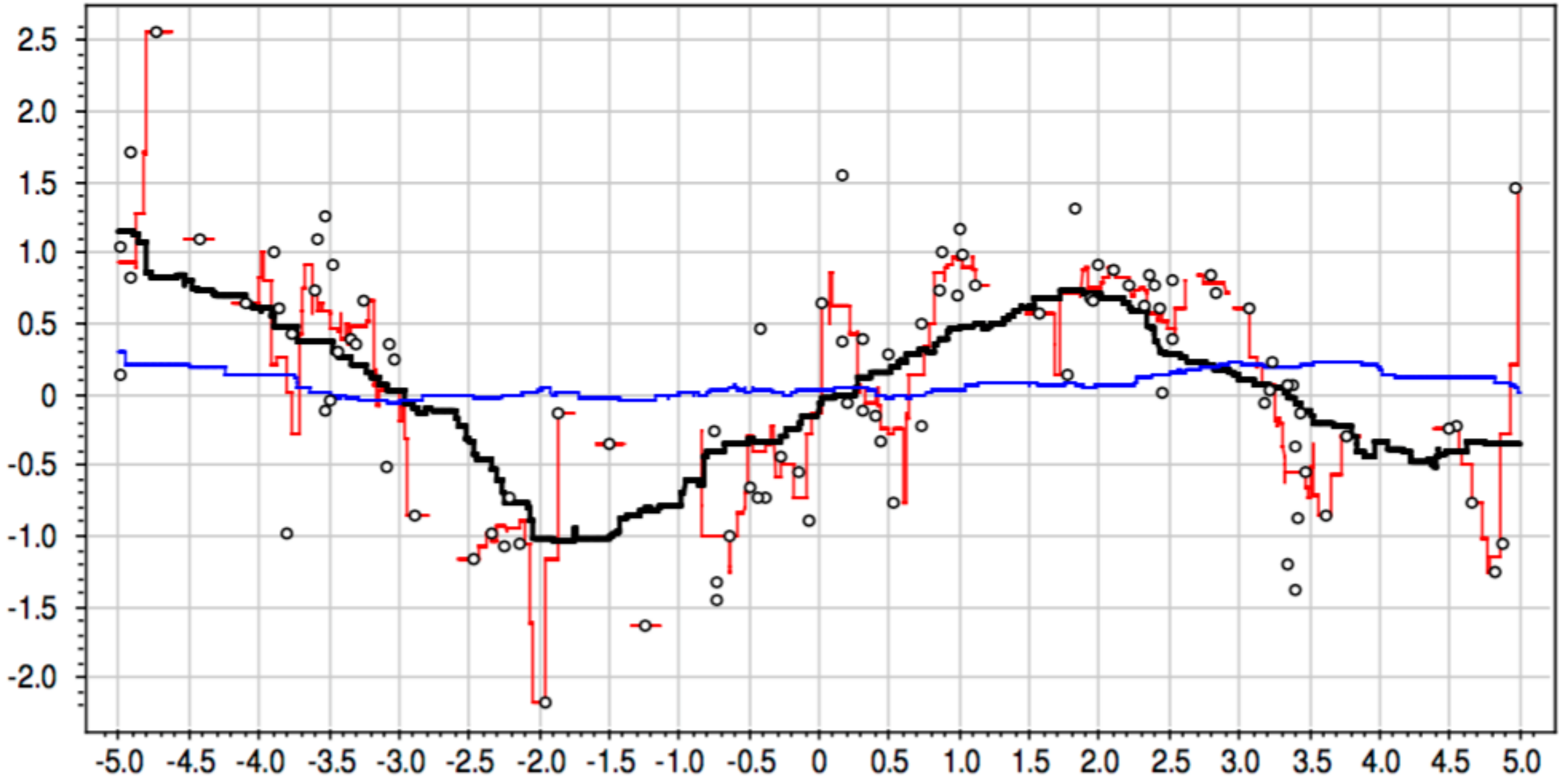
kNN для регрессии

- Гауссовское ядро
- $h \in \{0.1, 1.0, 3.0\}$



kNN для регрессии

- Прямоугольное ядро $K(z) = [|z| \leq 1]$
- $h \in \{0.1, 1.0, 3.0\}$



Особенности kNN

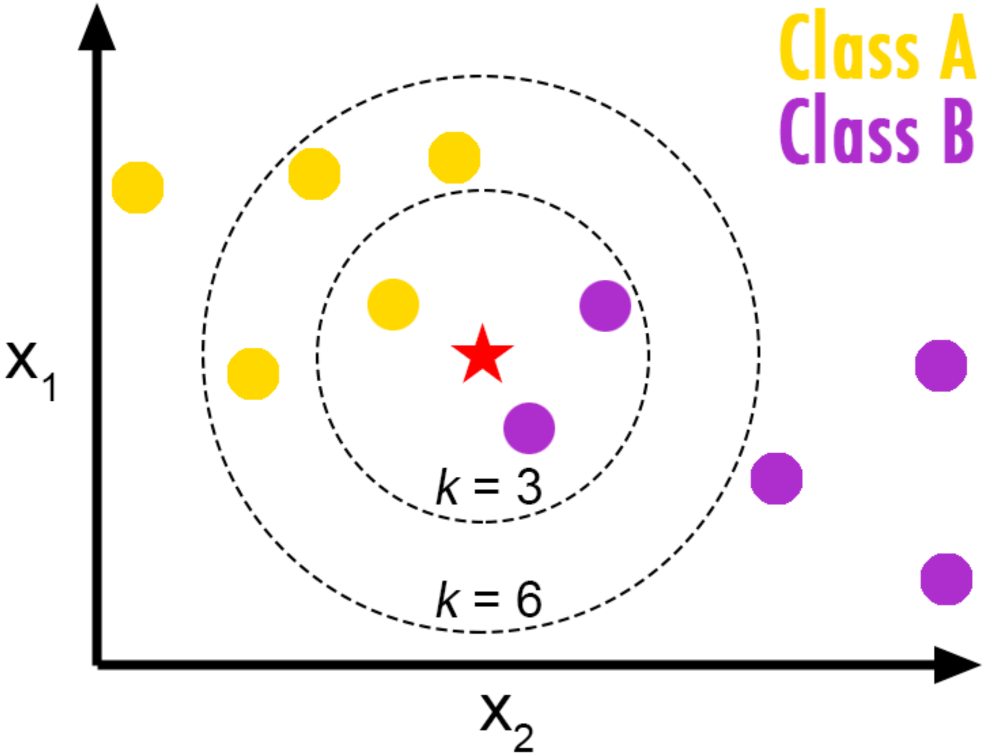
- Обучение как таковое отсутствует — нужно лишь запомнить обучающую выборку
- Для применения модели необходимо вычислить расстояния от нового объекта до всех обучающих объектов
- Применение требует ℓd операций
- Существуют специальные методы для поиска ближайших соседей

Как выбрать k?

- Как все-таки выбрать k?
- Хотим выбрать так, чтобы качество было хорошим
- …на какой-то выборке

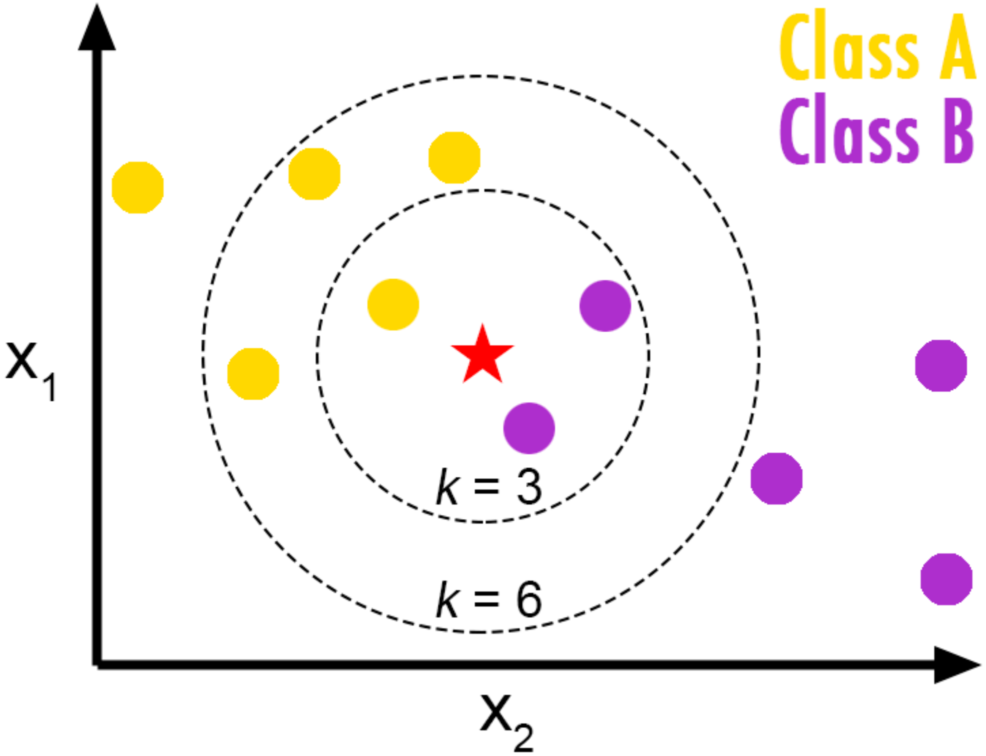
Как выбрать k?

- Выберем так, чтобы минимизировать ошибку на обучающей выборке — какое k будет оптимальным?



Как выбрать k?

- Выберем так, чтобы минимизировать ошибку на обучающей выборке — какое k будет оптимальным?



k=1

- Ближайший сосед совпадает с самим объектом!
- Поэтому при k=1 мы гарантированно будем относить каждый объект к правильному классу
- **Вывод**: подбирать k по обучающей выборке нельзя

Как выбрать k?

- Идея: использовать отложенную выборку
- Тогда качество будет оцениваться по объектам, не входящим в обучение
- Для каждого потенциального значения k обучаем модель на обучающей выборке, считаем качество на тестовой
- k = 1, 3, 5, 7, 10
- Выбираем то значение, для которого качество на тестовой выборке лучше всего

Как выбрать k?

k	$accuracy(holdout)$
1	0.80
3	0.87
5	0.93
7	0.92
10	0.89

Как выбрать k?

- Аналогично можно использовать другие способы оценки обобщающей способности (например, кросс-валидацию)
- Для каждого потенциального значения k считаем качество на кросс-валидации
- Берем лучшее значение k

Как выбрать k?

- Количество соседей k — гиперпараметр (или структурный параметр) модели
- Оптимальные значения гиперпараметров подбираются по валидационной выборке (не по обучающей!)
- Подробнее — в следующих лекциях

Как выбрать k?

- Количество соседей k — гиперпараметр (или структурный параметр) модели
- Оптимальные значения гиперпараметров подбираются по валидационной выборке (не по обучающей!)
- Подробнее — в следующих лекциях
- Ширина окна h, вид ядра $K(z)$ — тоже гиперпараметры

Резюме

- Операции в векторных пространствах
    - Норма
    - Метрика
    - Скалярное произведение
- Переобучение не отличить от хорошей модели на обучающей выборке — нужны оценки обобщающей способности
    - Их же можно использовать для выбора k в kNN
    - …и любых других гиперпараметров
- kNN
    - Для регрессии берем среднее
    - Ядра — чтобы веса соседей зависели от расстояния до них