

# Метрические методы

Виктор Владимирович Китов

МГУ им. Ломоносова, ф-т ВМиК, кафедра ММП.

I семестр 2015 г.

# Содержание

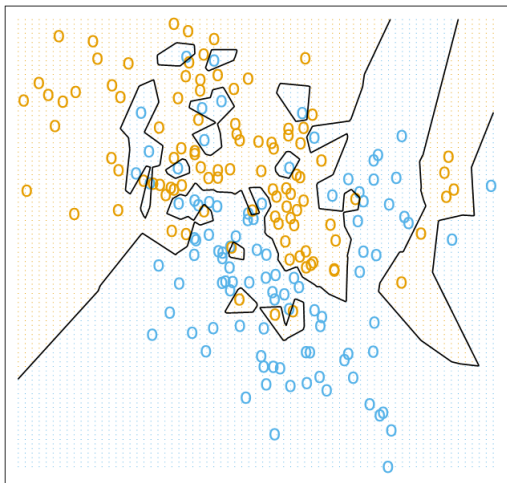
- 1 Простейший вариант
- 2 Выбор метрики
- 3 Взвешенный учет ближайших соседей

## Метод K-ближайших соседей (K-nearest neighbours)

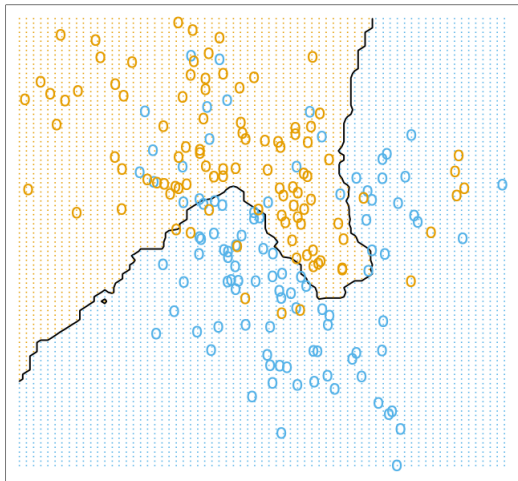
### Классификация методом k ближайших соседей

- 1 Найти  $k$  ближайших объектов к интересующему объекту  $x$  в обучающей выборке.
  - 2 Сопоставить  $x$  самый часто встречающийся класс среди  $k$  соседей.
- Случай  $k = 1$ : алгоритм ближайшего соседа (nearest neighbour)
  - Случай  $k = N$ : константный прогноз наиболее частым классом в выборке.
  - В случае регрессии нужно усреднить характеристики  $k$  ближайших соседей.
  - Основное предположение метода:
    - близкие объекты выдают похожие ответы

## Пример: классификация одним ближайшим соседом



## Пример: классификация 15 ближайшими соседями



## Параметры метода

- Параметры:
  - число соседей  $k$
  - функция близости  $\rho(x, y)$
- Вариант метода с адаптивным  $k$  по точности ближайших соседей.

## Свойства

- Преимущества:
  - нужно знать только ф-цию близости между объектами, сами признаки не нужны.
    - может быть применен к объектам любой сложности, если задана ф-ция близости
  - простая логика работы, легко объяснить и реализовать
  - интерпретируемость (case based reasoning)
  - не требует обучения
    - может применяться в online случаях.
    - K-CV можно заменить на LOO оценивание.
- Недостатки:
  - медленная классификация со сложностью  $O(N)$
  - требования по памяти тоже  $O(N)$ , т.к. нужно хранить всю выборку
  - точность ухудшается с ростом размерности пространства

## Проклятие размерности (curse of dimensionality)

- Проклятие размерности - явление, когда с ростом размерности признакового пространства  $D$  методу для поддержания точности требуется обучающая выборка, которая растет экспоненциально от  $D$ .
- Пример: оценка гистограмм



## Проклятие размерности (curse of dimensionality)

- Случай K-ближайших соседей:
  - допущение: объекты распределены равномерно в признаковом пространстве
  - шар радиуса  $R$  имеет объем  $V(R) = CR^D$ ,  $C = \frac{\pi^{D/2}}{\Gamma(D/2+1)}$ .
  - отношение объемов шаров радиуса  $R$  и  $R - \varepsilon$ :

$$\frac{V(R - \varepsilon)}{V(R)} = \left( \frac{R - \varepsilon}{R} \right)^D \xrightarrow{D \rightarrow \infty} 0$$

- основной объем концентрируется на поверхности сферы, значит там лежат ближайшие соседи.
  - ближайшие соседи перестают быть близкими по расстоянию.
- В реальных задачах признаки могут быть распределены не равномерно по всему объему, а быть зависимыми и лежать на многообразии меньшей размерности.

## Классификация - пограничный случай

Когда несколько классов имеют одинаковый ранг, можно сопоставить класс:

- случайным образом
- имеющий большую априорную вероятность
- чей ближайший представитель ближе к  $x$
- для которого среднее значение представителей этого класса среди ближайших соседей ближе
- чей самый дальний представитель среди ближайших соседей ближе (т.е. класс представлен более компактно)

# Содержание

- 1 Простейший вариант
- 2 Выбор метрики**
- 3 Взвешенный учет ближайших соседей

# Нормализация признаков

- Чаще всего используется Евклидова метрика близости:

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

- Необходимо нормализовывать признаки.
  - Определим  $\mu_j$ ,  $\sigma_j$ ,  $L_j$ ,  $U_j$  как среднее значение, стандартное отклонение, минимальное и максимальное значение  $j$ -го признака.

Преобразование	Свойства результата
$x'_j = \frac{x_j - \mu_j}{\sigma_j}$	среднее=0, дисперсия=1.
$x'_j = \frac{x_j - L_j}{U_j - L_j}$	принадлежит интервалу [0, 1].

## Нормализация признаков

- Нелинейные трансформации для признаков, допускающих редкие большие значения:
  - $\tilde{x}^i = \log(x^i)$
  - $\tilde{x}^i = (x^i)^p, 0 \leq p < 1$
- Для  $F_i(\alpha) = P(x^i \leq \alpha)$  преобразование  $\tilde{x}^i \rightarrow F_i(x^i)$  даст признак, равномерно распределенный на  $[0, 1]$ .

# Выбор функции расстояния

Metric	$d(x, z)$
Евклидова	$\sqrt{\sum_{i=1}^D (x^i - z^i)^2}$
$L_p$	$\sqrt[p]{\sum_{i=1}^D (x^i - z^i)^p}$
$L_\infty$	$\max_{i=1,2,\dots,D}  x^i - z^i $
$L_1$	$\sum_{i=1}^D  x^i - z^i $
Canberra	$\frac{1}{D} \sum_{i=1}^D \frac{ x^i - z^i }{x^i + z^i}$
Ланса-Уильямса	$\frac{\sum_{i=1}^D  x^i - z^i }{\sum_{i=1}^D x^i + z^i}$
косинусная мера	$1 - \frac{\sum_{i=1}^D x^i z^i}{\sqrt{\sum_{i=1}^D (x^i)^2} \sqrt{\sum_{i=1}^D (z^i)^2}}$

## Нормализующее преобразование

- $x$  распределен согласно некоторому распределению со средним  $\mu$  и невырожденной матрицей ковариации  $\Sigma$  ( $\mu \in \mathbb{R}^D$ ,  $\Sigma \in \mathbb{R}^{D \times D}$ )
- Тогда нормализующее преобразование (whitening)  $z = \Sigma^{-1/2}(x - \mu)$  даст новый вектор признаков со средним  $0$  и ковариацией  $I \in \mathbb{R}^{D \times D}$ , где  $I$  - это единичная матрица.
- Доказательство:

$$\mathbb{E}z = \mathbb{E} \left\{ \Sigma^{-1/2}(x - \mu) \right\} = \Sigma^{-1/2} \mathbb{E} \{x - \mu\} = \mathbf{0} \in \mathbb{R}^D$$

$$\begin{aligned} \text{cov}[z] &= \mathbb{E}(z - \mathbb{E}z)(z - \mathbb{E}z)^T = \mathbb{E}zz^T \\ &= \mathbb{E} \left\{ \Sigma^{-1/2}(x - \mu)(x - \mu)^T (\Sigma^{-1/2})^T \right\} = \\ &= \Sigma^{-1/2} \mathbb{E} \left\{ (x - \mu)(x - \mu)^T \right\} (\Sigma^{-1/2})^T = \\ &= \Sigma^{-1/2} \Sigma \Sigma^{-1/2} = I \end{aligned}$$

## Расстояние между нормализациями

- Расстояние между  $x$  и  $x'$  можно считать в нормализованном пространстве как Евклидово расстояние между  $z = \Sigma^{-1/2}(x - \mu)$  и  $z' = \Sigma^{-1/2}(x' - \mu)$ :

$$\begin{aligned} \rho_M(x, x') &= \rho_E(z, z') = \sqrt{(z - z')^T (z - z')} = \\ &= \sqrt{(x - x')^T \Sigma^{-1/2} \Sigma^{-1/2} (x - x')} \\ &= \sqrt{(x - x')^T \Sigma^{-1} (x - x')} \end{aligned}$$

- Это называется *расстоянием Махаланобиса*.
- Частный случай: если признаки нескоррелированы и

$$\text{Var } x^i = \sigma_i^2, \text{ то } \rho(x, \tilde{x}) = \sqrt{\sum_i \frac{(x^i - \tilde{x}^i)^2}{\sigma_i^2}}$$



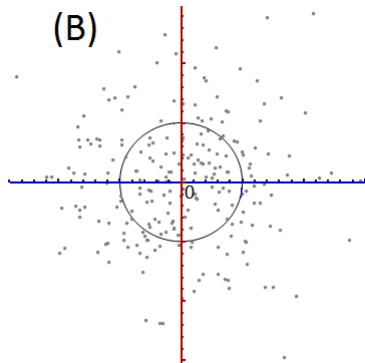
## Расстояние Махаланобиса - иллюстрация

(A): объекты в исходном признаковом пространстве и множество точек  $G_\alpha = \{x : \rho_M(x, \mu) = \alpha\}$ . (B): объекты в нормализованном признаковом пространстве и образ  $G_\alpha$  - множество точек  $\{z : \rho_E(z, 0) = \alpha\}$ .

(A)



(B)



# Содержание

- 1 Простейший вариант
- 2 Выбор метрики
- 3 Взвешенный учет ближайших соседей**

## Взвешенный учет

Определим  $x_{i_1}, x_{i_2}, \dots, x_{i_N}$  как переупорядочивание  $x_1, x_2, \dots, x_N$  по расстоянию до  $x$ :  $\rho(x, x_{i_1}) \leq \rho(x, x_{i_2}) \leq \dots \leq \rho(x, x_{i_N})$ .

Определим  $z_1 = x_{i_1}, z_2 = x_{i_2}, \dots, z_K = x_{i_K}$ .

Метод ближайших соседей можно определить через  $C$  дискриминантных функций:

$$g_c(x) = \sum_{k=1}^K \mathbb{I}[z_k \in \omega_c], \quad c = 1, 2, \dots, C.$$

Взвешенный учет ближайших соседей:

$$g_c(x) = \sum_{k=1}^K w(k, \rho(x, z_k)) \mathbb{I}[z_k \in \omega_c], \quad c = 1, 2, \dots, C.$$

## Часто выбираемые веса

Независимые от  $x$ :

$$w_k = \alpha^k, \quad \alpha \in (0, 1)$$

$$w_k = \frac{K + 1 - k}{K}$$

Зависимые от  $x$ :

$$w_k = \begin{cases} \frac{\rho(z_K, x) - \rho(z_k, x)}{\rho(z_K, x) - \rho(z_1, x)}, & \rho(z_K, x) \neq \rho(z_1, x) \\ 1 & \rho(z_K, x) = \rho(z_1, x) \end{cases}$$

$$w_k = \frac{1}{\rho(z_k, x)}$$

## Понятие отступа

- Рассмотрим обучающую выборку:  
 $(x_1, c_1), (x_2, c_2), \dots, (x_N, c_N)$ , где  $c_i$  - правильный класс для объекта  $x_i$ , а  $\mathbf{C} = \{1, 2, \dots, C\}$  - множество допустимых классов.
- Определим понятие отступа:

$$M(x_i, c_i) = g_{c_i}(x_i) - \max_{c \in \mathbf{C} \setminus \{c_i\}} g_c(x_i)$$

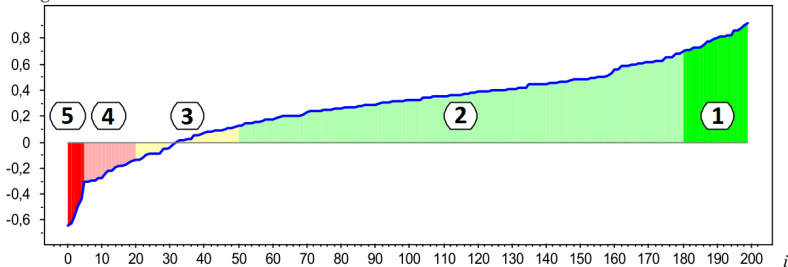
- отступ отрицательный  $\Leftrightarrow$  объект  $x_i$  был неправильно классифицирован
- величина отступа показывает, насколько классификатор уверен, что объект  $x_i$  может быть отнесен к истинному классу  $c_i$

# Классификация объектов по отступам

По величине отступа объекты делятся на следующие категории:

- 1 Объекты, классифицированные верно с высокой степенью уверенности
- 2 Правильно классифицированные объекты
- 3 Пограничные объекты (их классификация неустойчива)
- 4 Неправильно классифицированные объекты
- 5 Объекты, сильно выбивающиеся из закономерности, определенной классификатором

*Margin*



## Критерии качества по отступу

Хороший классификатор:

- минимизирует область отрицательных отступов
- классифицирует правильно с высокой степенью уверенности (высоким  $M(x_i, c_i)$ )