Лекция 3:
метрические методы

Евгений Борисов

метрические методы

общая схема применения методов ML

определяем задачу в общем виде

изучаем предметную область

формализуем задачу

метрические методы

общая схема применения методов ML

определяем задачу в общем виде

изучаем предметную область

формализуем задачу

извлекаем признаки из объекта

собираем и обрабатываем учебный набор

выбираем и обучаем модель

метрические методы

общая схема применения методов ML

определяем задачу в общем виде

изучаем предметную область

формализуем задачу

извлекаем признаки из объекта

собираем и обрабатываем учебный набор

выбираем и обучаем модель

тестируем модель

запускаем модель в работу

метрические методы

Задачи ML

Классификация - разделение на части

Кластеризация - формирование групп

Регрессия - восстановление зависимости

метрические методы

методы ML

Метрические: k-Neighbors

Статистические: Naive Bayes

Логические: Decision Tree

Линейные: SVM, MLP

Композиции: AdaBoost

метрические методы

датасет - размеченная матрица признаков

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} & y_1 \\ x_{21} & x_{22} & \cdots & x_{2n} & y_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} & y_m \end{bmatrix}$$

x - вектор-признак

y - метка класса

n - размер пространства признаков

m - количество примеров

метрические методы

метрика - функция расстояния

$$\rho: X \times X \to [0, \infty)$$

аксиома тождества : $\rho(x, y) = 0 \Leftrightarrow x = y$

симметрия: $\rho(x, y) = \rho(y, x)$

неравенство треугольника: $\rho(x, z) \leqslant \rho(x, y) + \rho(y, z)$

метрические методы

метрика - функция расстояния

Евклидова метрика: $\rho(x,y)=\sqrt{\sum_i (x_i-y_i)^2}$

метрика Минковского: $\rho(x,y)=\sqrt[n]{\sum_i w_i|x_i-y_i|^n}$

метрика Чебышева: $\rho(x,y)=\max_i|x_i-y_i|$

косинусная метрика: $\rho(x,y)=\frac{\sum_i x_i y_i}{\sqrt{\sum_i x_i^2}\cdot\sqrt{\sum_i y_i^2}}$

метрические методы

метрический подход в методах ML

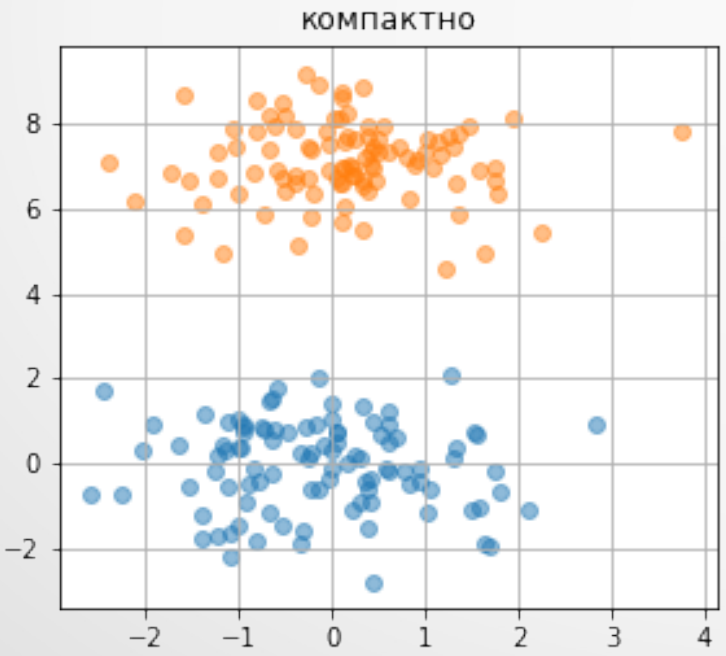
использование расстояний между объектами

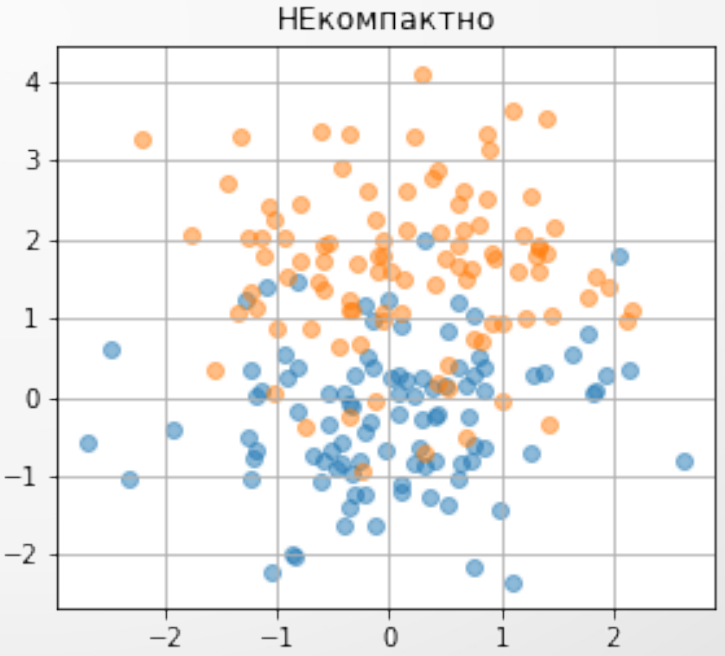
метрические методы

<u>метрический подход в методах ML</u>

использование расстояний между объектами

гипотеза компактности: близкие объекты лежат в одном классе





метрические методы

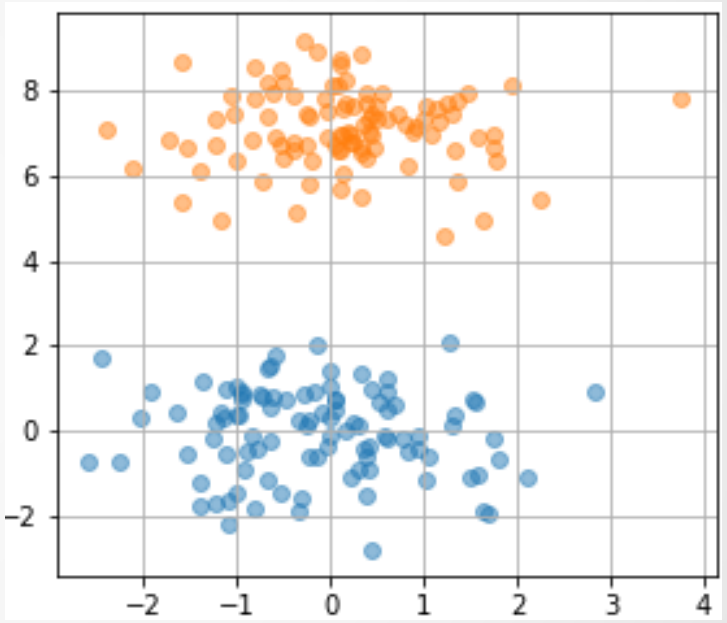
о задаче классификации

разделение данных на части (классы)

Учебный набор: [объект, ответ]

Задача: классификатор

объект → вектор-признак → класс



метрические методы

метрический классификатор

X - пространство признаков размерности m

$X_l \subset X - объекты\ учебной\ выборки$

$y_l - метки\ классов\ учебного\ набора\ X_l$

метрические методы

метрический классификатор

X - пространство признаков размерности m

$X_l \subset X - объекты\ учебной\ выборки$

$y_l - метки\ классов\ учебного\ набора\ X_l$

$u \in X - выберем\ объект$

выстроим соседей из $\mathbf{X_l}$ u объекта **u** по расстоянию (вариационный ряд)

$$\rho(u, x_u^1) \leqslant \rho(u, x_u^2) \leqslant \cdots \leqslant \rho(u, x_u^n)$$

метрические методы

метрический классификатор

X - пространство признаков размерности m

$X_l \subset X - объекты\ учебной\ выборки$

$y_l - метки\ классов\ учебного\ набора\ X_l$

$u \in X - выберем\ объект$

выстроим соседей из $\mathbf{X_l}$ u объекта **u** по расстоянию (вариационный ряд)

$$\rho(u, x_u^1) \leqslant \rho(u, x_u^2) \leqslant \cdots \leqslant \rho(u, x_u^n)$$

$\nu(i, u)$ - ф-ция оценки важности i-того соседа объекта u, убывает по мере удаления от u

метрические методы

метрический классификатор

X - пространство признаков размерности m

$X_l \subset X - объекты\ учебной\ выборки$

$y_l - метки\ классов\ учебного\ набора\ X_l$

$u \in X - выберем\ объект$

выстроим соседей из $\mathbf{X_l}$ u объекта **u** по расстоянию (вариационный ряд)

$$\rho(u, x_u^1) \leqslant \rho(u, x_u^2) \leqslant \cdots \leqslant \rho(u, x_u^n)$$

$v(i,u)$ - ф-ция оценки важности i-того соседа объекта u, убывает по мере удаления от u

$\Gamma_y(u) = \sum_i [y = y_i] v(i,u)$ - оценка близости **u** к классу **y**

метрические методы

метрический классификатор

X - пространство признаков размерности m

$X_l \subset X - объекты\ учебной\ выборки$

$y_l - метки\ классов\ учебного\ набора\ X_l$

$u \in X - выберем\ объект$

выстроим соседей из $\mathbf{X_l}$ u объекта **u** по расстоянию (вариационный ряд)

$$\rho(u, x_u^1) \leqslant \rho(u, x_u^2) \leqslant \cdots \leqslant \rho(u, x_u^n)$$

$v(i,u)$ - ф-ция оценки важности i-того соседа объекта u, убывает по мере удаления от u

$\Gamma_y(u) = \sum_i [y = y_i] v(i,u)$ - оценка близости **u** к классу **y**

$$a(u, X_l) = \underset{y \in y_l}{argmax}\, \Gamma_y(u)$$

метрические методы

метод ближайшего соседа (1NN)

$v(i,u) = [i=1]$

метрические методы

метод ближайшего соседа (1NN)

$v(i,u) = [i=1]$

достоинства:

- простота
- интерпретируемость

метрические методы

метод ближайшего соседа (1NN)

$v(i,u) = [i=1]$

достоинства:

- простота
- интерпретируемость

недостатки:

- неустойчив к шуму
- нет параметров
- недостаточная точность
- выборка хранится целиком

метрические методы

метод ближайшего соседа (1NN)

$v(i,u) = [i=1]$

<u>достоинства</u>:

- простота
- интерпретируемость

<u>недостатки</u>:

- неустойчив к шуму
- нет параметров
- недостаточная точность
- выборка хранится целиком

метод k-соседей (kNN)

$v(i,u) = [i<k]$

метрические методы

метод ближайшего соседа (1NN)

$v(i,u) = [i=1]$

достоинства:

- простота
- интерпретируемость

недостатки:

- неустойчив к шуму
- нет параметров
- недостаточная точность
- выборка хранится целиком

метод k-соседей (kNN)

$v(i,u) = [i<k]$

достоинства:

- более устойчив к шуму чем 1NN
- есть параметр - количество соседей k

недостатки:

- возможны неоднозначности

метрические методы

метод ближайшего соседа (1NN)

$v(i,u) = [i=1]$

<u>достоинства</u>:

- простота
- интерпретируемость

<u>недостатки</u>:

- неустойчив к шуму
- нет параметров
- недостаточная точность
- выборка хранится целиком

метод k-соседей (kNN)

$v(i,u) = [i<k]$

<u>достоинства</u>:

- более устойчив к шуму чем 1NN
- есть параметр - количество соседей k

<u>недостатки</u>:

- возможны неоднозначности

метод взвешеных k-соседей

$v(i,u) = [i<k]w_i$

w_i - вес соседа

метрические методы

метод взвешеных k-соседей

$v(i,u) = [i<k]w_i$ w_i - вес соседа

как выбирать вес w_i?

метрические методы

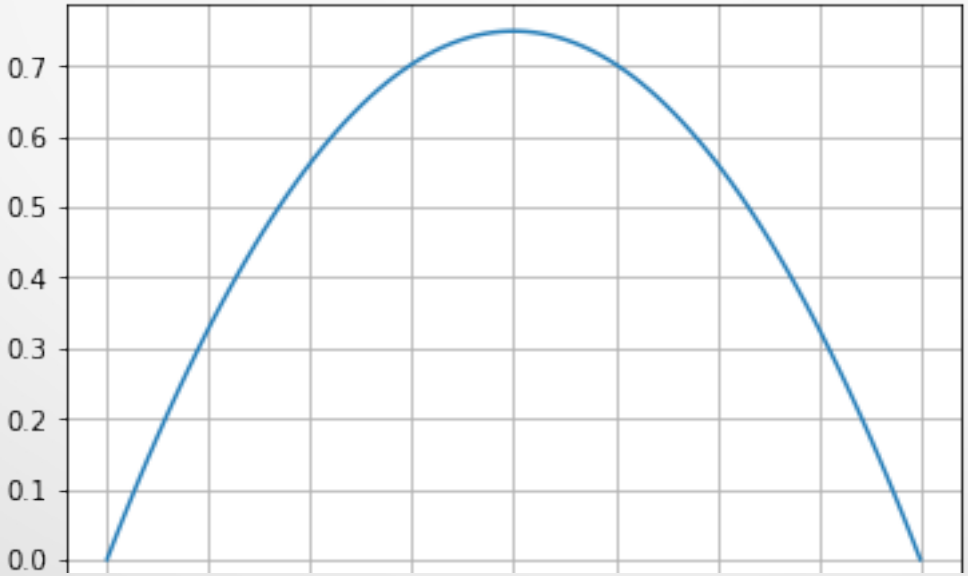
метод взвешеных k-соседей

$v(i,u) = [i<k]w_i$ w_i - вес соседа

как выбирать вес w_i?

$$v(i,u)=K\left(\frac{\rho(u,x_u^i)}{h}\right)$$

выбираем степень важности i-того соседа на основании расстояния до него

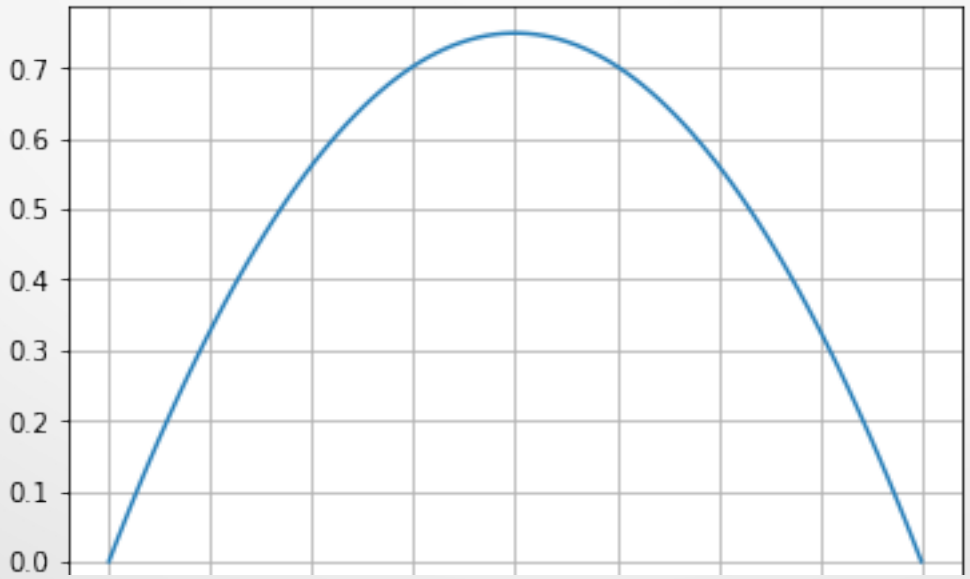


метрические методы

метод взвешеных k-соседей - парзеновское окно

выбираем степень важности i-того соседа на основании расстояния

$$a(u, X_l) = \underset{y \in y_l}{argmax} \sum_i [y(i) = y] K\left(\frac{\rho(u, x_u^i)}{h}\right)$$



метрические методы

профиль компактности - метод оценки данных и метрик на них

доля объектов, у которых m-тый сосед из другого класса

метрические методы

профиль компактности - метод оценки данных и метрик на них

доля объектов, у которых m-тый сосед из другого класса

$$K(m,X)=\frac{1}{L}\sum_{i=1}^{L}[y_i \neq y_i^m]$$

x_i^m - m-тый сосед x_i

y_i^m - ответ на m-том соседе x_i

метрические методы

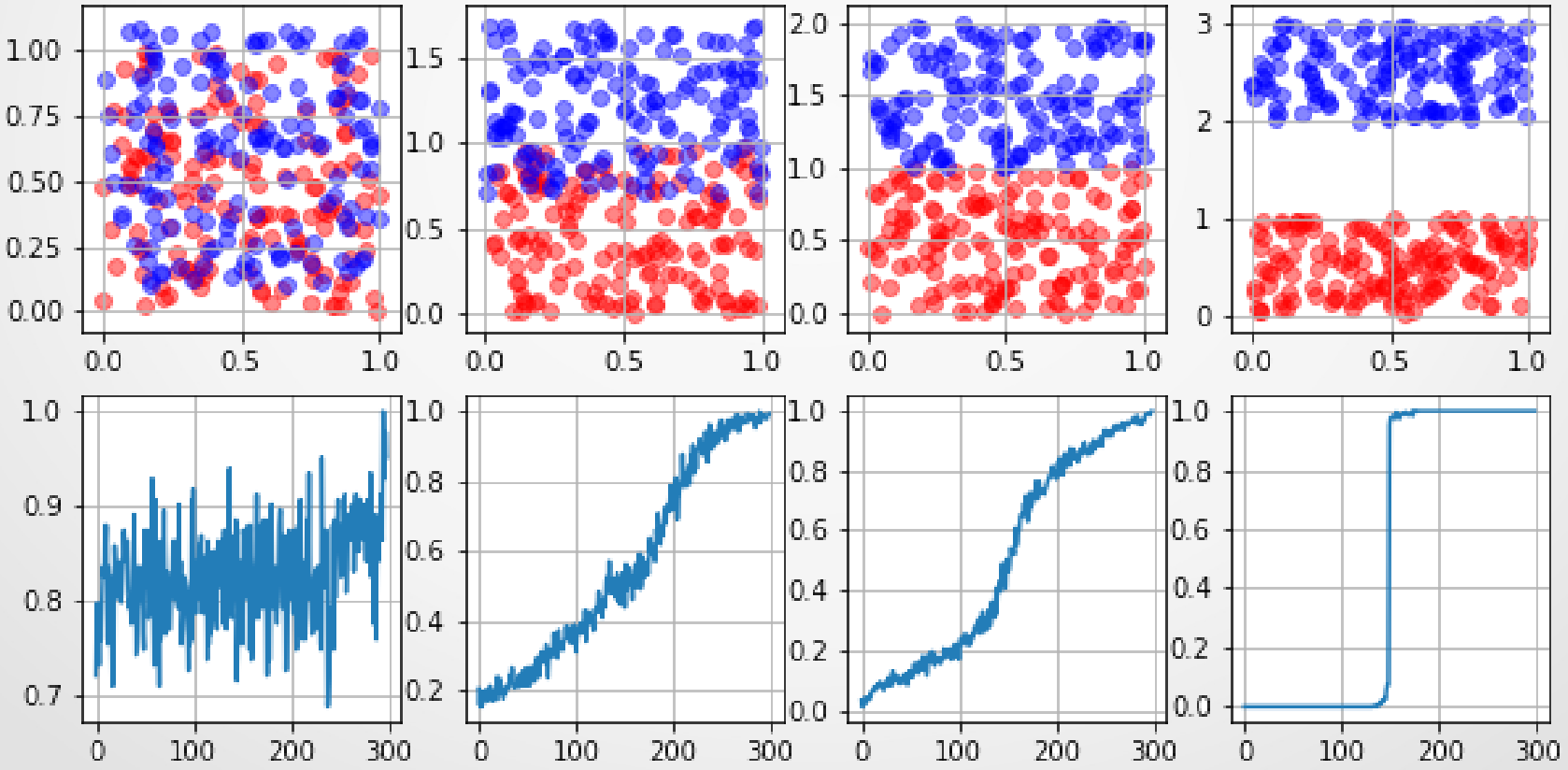
профиль компактности - метод оценки данных и метрик на них

доля объектов, у которых m-тый сосед из другого класса

$$K(m,X)=\frac{1}{L}\sum_{i=1}^{L}[y_i \neq y_i^m]$$

x_i^m - m-тый сосед x_i

y_i^m - ответ на m-том соседе x_i



метрические методы: литература

git clone *https://github.com/mechanoid5/ml_lectorium.git*

К.В. Воронцов Метрические методы классификации. - курс "Машинное обучение" ШАД Яндекс 2014

метрические методы

Вопросы ?

метрические методы: практика

источники данных для экспериментов

sklearn.datasets
UCI Repository
kaggle

реализовать
1NN
kNN