СТАТИСТИКА БОЛЬШИХ ДАННЫХ

Лекция 26. Задача классификации. Часть1

Постановка задачи

Классификация — раздел машинного обучения.
Имеется множество объектов (ситуаций), разделённых на классы. Задано конечное множество объектов **(обучающая выборка)**, для которых известно, к каким классам они относятся.
Требуется построить алгоритм, классифицирующий произвольный объект из исходного множества.

Классифицировать объект — указать номер (или наименование класса), к которому относится данный объект.

Постановка задачи

Задача обучения с учителем

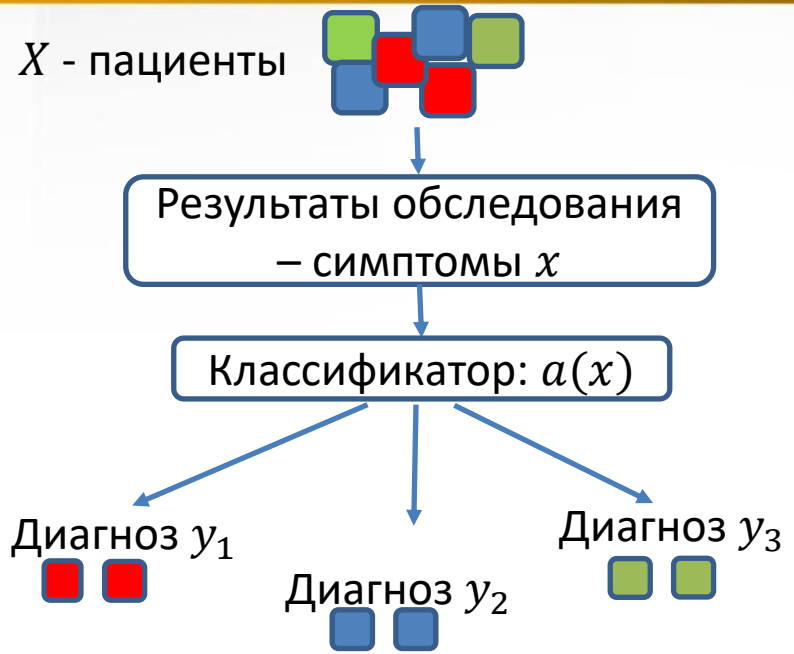
Рассматриваются: вектор $x \in X$ — набор признаков, X — **множество объектов,** Y — **множество ответов (откликов, меток).** Задана выборка $X^l = \{x_1, \dots, x_l\} \subset X$ и **множество известных ответов** $y_i = a^*(x_i)$. Совокупность упорядоченных пар "объект-ответ" (x_i, y_i) называется **обучающей выборкой.**

Задача. Построить **решающее правило (решающую функцию, алгоритм)** (decision function) $a: X \to Y$, которая приближала бы функцию $a^*(x)$ на всём множестве X.

Задача обучения с учителем, в которой множество Y — это конечное множество номеров (имён, меток) классов, называется **задачей классификации**. Требуется построить алгоритм, сопоставляющий объекту $x \in X$ метку $y \in Y$.

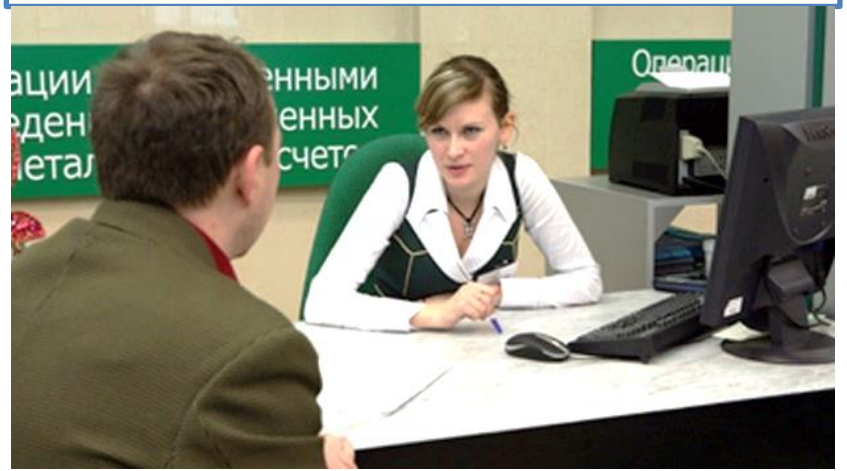
Пример задачи классификации

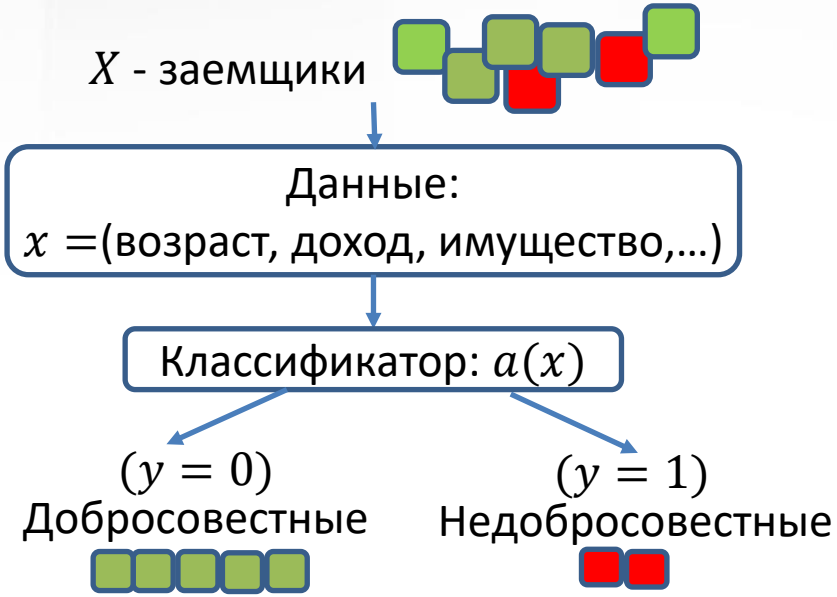
Медицинская диагностика



Пример задачи классификации

Распознавание недобросовестного заемщика



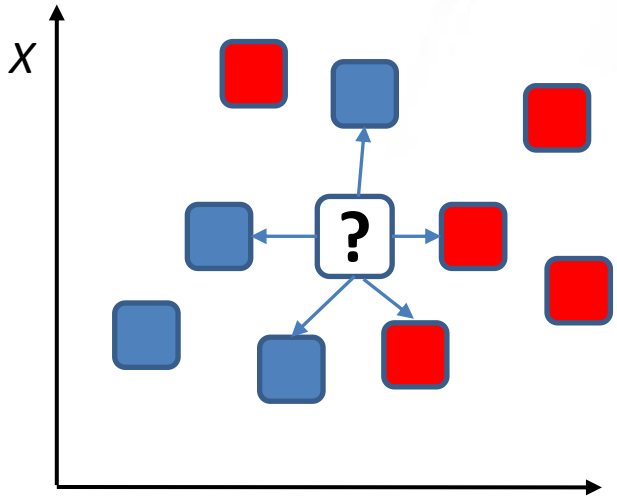


Методы решения

- Метод k ближайших соседей
- Байесовский классификатор
- Метод опорных векторов (SVM)
- Дерево решений

Метод k ближайших соседей

Идея метода: классифицируемый объект относится к тому же классу, к которому принадлежат большинство из ближайших к нему объектов обучающей выборки.



Размерность вектора x не имеет значения, что выгодно отличает этот метод от других.

Для применения метода k ближайших соседей нужно определить меру близости двух векторов из множества X.

Метод k ближайших соседей. Меры близости

1.Манхэттенское расстояние

$$\rho_1(x, x') = \sum_i^n |x_i - x_i'|$$

2.Евклидово расстояние

$$\rho_2(x, x') = \sqrt{\sum_i^n (x_i - x_i')^2}$$

3. Расстояние Чебышева

$$\rho_\infty(x, x') = max_i(|x_i - x_i'|)$$

4. Косинусная мера сходства

$$\rho_{\cos}(x, x') = \frac{\sum_i^n x_i x_i'}{\sqrt{\sum_i^n x_i^2}\sqrt{\sum_i^n x_i'^2}}$$

5. Частота несовпадений

$$\rho_I(x, x') = \frac{1}{n}\sum_i^n I(x_i \neq x_i')$$

Алгоритм k ближайших соседей

0. Задана обучающая выборка из l пар (x_i, y_i). На вход подается вектор x.
1. Задать число k соседей, по которым производится сравнение. Число k должно быть много меньше l - объема обучающей выборки.
2. Выбрать меру близости ρ и вычислить расстояния $\rho(x, x_i)$ от вектора x до каждого из объектов обучающей выборки.
3. Отобрать k ближайших соседей – k объектов обучающей выборки с минимальным расстоянием до вектора x.
4. Определить класс объекта x — это класс y, наиболее часто встречающийся среди k ближайших соседей.

Алгоритм k ближайших соседей

Достоинства и недостатки метода.

Достоинства:

- Простота реализации.
- Не чувствителен к увеличению размерности пространства признаков.
- Классификацию, проведенную алгоритмом, легко интерпретировать путем предъявления пользователю нескольких ближайших объектов.

Недостатки:

- Необходимость хранения обучающей выборки целиком.
- Поиск ближайшего соседа предполагает сравнение классифицируемого объекта со всеми объектами выборки.

Алгоритм k ближайших соседей

Выбор числа k ближайших соседей

1. Малые значения k приведут к тому, что “шум” (выбросы) будет существенно влиять на результаты.
2. Большие значения усложняют вычисления и искажают логику ближайших соседей, в соответствии с которой ближайшие точки могут принадлежать одному классу (гипотеза компактности).
3. Эвристика: $k = \left[\sqrt{l}\right]$

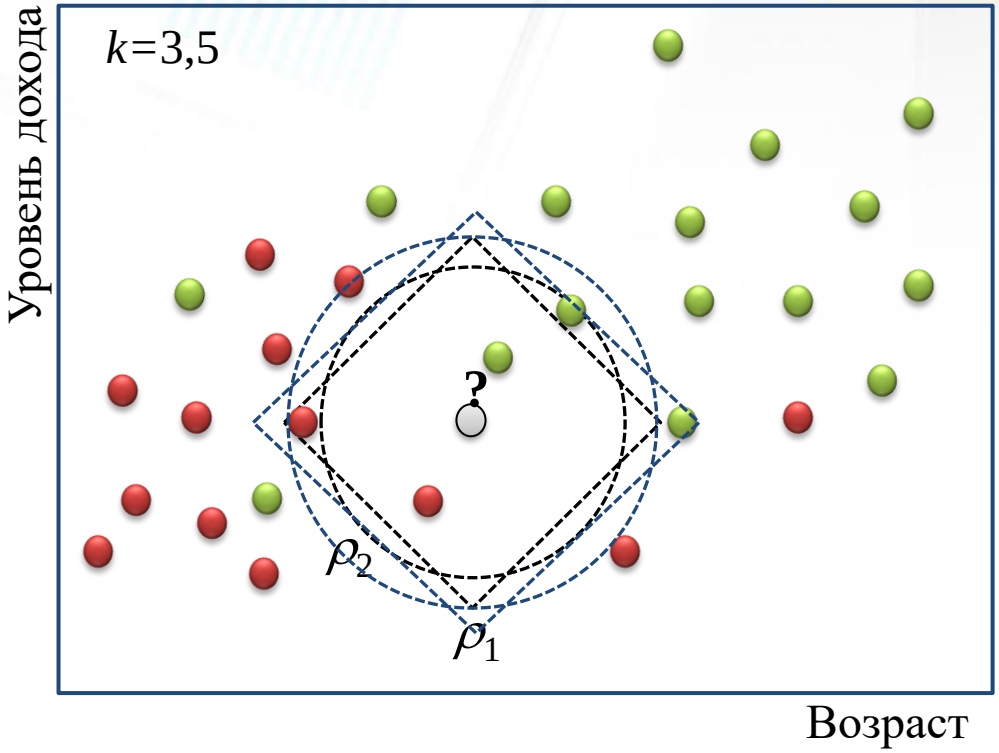
Предварительная подготовка данных.

1. Нормализация

$$x'_{ij} = \frac{x_{ij} - \min\limits_{i=1..l} x_{ij}}{\max\limits_{i=1..l} x_{ij} - \min\limits_{i=1..l} x_{ij}}$$

2. Отбраковка выбросов и шума

Метод k ближайших соседей. Пример.

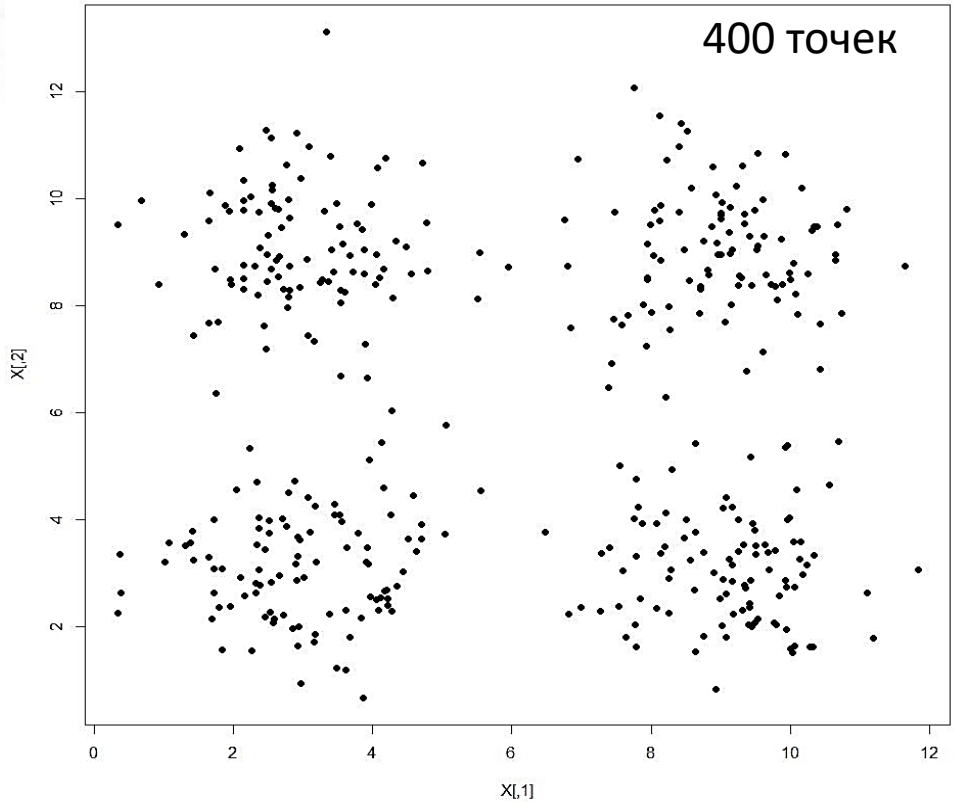


Задача распознавания недобросовестного заемщика

$$\rho_1(x, x') = \sum_i^n |x_i - x_i'|$$

$$\rho_2(x, x') = \sqrt{\sum_i^n (x_i - x_i')^2}$$

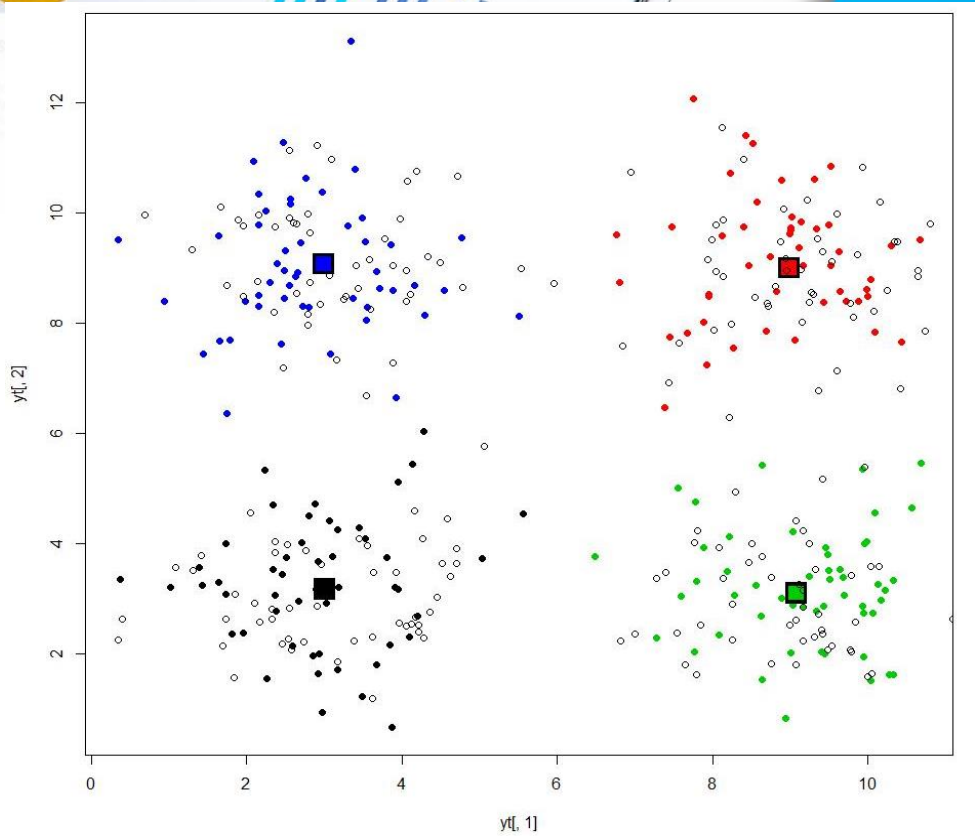
Метод k ближайших соседей. Пример.



Смесь из четырех нормальных двумерных распределений с центрами в точках (3,3) (3,9) (9,3) (9,9)

Решена задача кластеризации - выборка разбита на 4 класса.

Метод k ближайших соседей. Пример.




```
library(class)
....

knn_class<-knn(train=xt,test=xte,cl=ytl,k=5)

#xt - обучающая выборка
# xte- тестовые данные,
# ytl - метки для обучающей выборки
```

Предсказанное

Фактическое	1	2	3	4
1	51	0	0	0
2	0	50	0	0
3	0	0	50	0
4	0	0	0	49