Статистический анализ данных. Спецкурс.

Лекция 5. Классификация

Ботанический сад-институт ДВО РАН

Кислов Д.Е.

27 ноября 2016 г.

Задачи классификации

- Классификация в отсутствии обучающей выборки (кластеризация);

Задачи классификации

- Классификация в отсутствии обучающей выборки (кластеризация);
- Классификация при наличии обучающей выборки (по прецедентам);

Кластеризация

- Иерархический (агломеративная, дивизивные);

Кластеризация

- Иерархический (агломеративная, дивизивные);
- Логическая кластеризация

Кластеризация

- Иерархический (агломеративная, дивизивные);
- Логическая кластеризация
- Вероятностные

Кластеризация

- Иерархический (агломеративная, дивизивные);
- Логическая кластеризация
- Вероятностные
- Нейросетевые

Общая структура алгоритмов

- задание расстояние между кластеризуемыми объектами;
- задание расстояние между группами объектов;

Расстояние между объектами

Полагается, что объекты x, y имеют координаты $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$.

- Евклидово расстояние: $\rho(x, y) = \sum\limits_j (x_j - y_j)^2$;

Расстояние между объектами

Полагается, что объекты x, y имеют координаты $x_1, \dots, x_n$ и $y_1, \dots, y_n$.

- Евклидово расстояние: $\rho(x, y) = \sum_j (x_j - y_j)^2$;
- Расстояние Чебышева: $\rho(x, y) = \max_j |x_j - y_j|$;

Расстояние между объектами

Полагается, что объекты x, y имеют координаты $x_1, \ldots, x_n$ и $y_1, \ldots, y_n$.

- Евклидово расстояние: $\rho(x,y) = \sum\limits_j (x_j - y_j)^2$;
- Расстояние Чебышева: $\rho(x,y) = \max_j |x_j - y_j|$;
- Расстояние city-block: $\rho(x,y) = \sum\limits_i |x_i - y_i|$;

Расстояние между объектами

Полагается, что объекты x, y имеют координаты $x_1, \dots, x_n$ и $y_1, \dots, y_n$.

- Евклидово расстояние: $\rho(x, y) = \sum\limits_j (x_j - y_j)^2$;
- Расстояние Чебышева: $\rho(x, y) = \max_j |x_j - y_j|$;
- Расстояние city-block: $\rho(x, y) = \sum\limits_i |x_i - y_i|$;
- Расстояние Минковского: $\rho(x, y)^p = \sum\limits_i (x_i - y_i)^p$;

Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);

Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);
- метод максимального расстояния (complete method);

Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);
- метод максимального расстояния (complete method);
- попарное среднее;

Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);
- метод максимального расстояния (complete method);
- попарное среднее;
- центроидный метод;

Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);

Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);
- метод максимального расстояния (complete method);

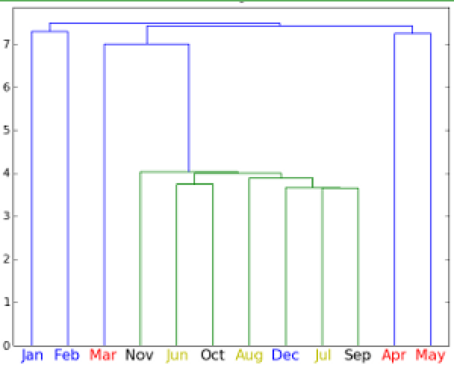
Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);
- метод максимального расстояния (complete method);
- попарное среднее;

Подходы для вычисления расстояний между группами объектов

- метод минимального расстояния (single method);
- метод максимального расстояния (complete method);
- попарное среднее;
- центроидный метод;

Представление иерархической кластеризации в виде дендрограммы



Метод k-средних

- задается число кластеров k;
- в факторном пространстве случайным образом выбираются начальные приближения центров кластеров;
- ближайшие к j-му центру точки помещаются в j-й кластер;
- пересчитываются центроиды кластеров;

Последние 2 шага повторяются пока алгоритм не сойдется.

Сравнение кластерных структур

Задача

Можно ли построить какую-либо меру, чтобы сравнить, например, кластерные структуры?:

$$a, a, a, b, b, c, c, c, c$$
$$1, 1, 1, 3, 3, 2, 2, 2, 2$$

или

$$a, a, a, b, b, c, c, c, c$$
$$2, 2, 2, 1, 1, 3, 3, 3, 1$$

Индекс Рэнда

Пусть $X = (x_1, \dots, x_n)$ – n-элементное множество;
$P = (A_1, \dots, A_r)$ и $Q = (B_1, \dots, B_s)$ – два его разбиения.
Определим

- a – число пар элементов попавших в один кластер в разбиениях P и Q одновременно;
- b – число пар элементов, находящихся в разных кластерах в разбиения P и Q;
- c – число пар элементов, находящихся в одном кластере в P разбиении, но в разных в Q;
- d – число пар элементов, находящихся в разных кластерах в P разбиении, но в одном в Q;

В этом случае $a + b$ – характеризует степень совпадения кластеров, если $c = d = 0$, то кластера совпадают. Индекс Рэнда:

$$I_R = \frac{a+b}{a+b+c+d} = \frac{a+b}{C_n^2}$$