Введение в многомерный статистический анализ
2021-2022 учебный год

Автор лекции: Тамбовцева А.А.

Лекция 2. Кластерный анализ методом k-средних

1.1 Алгоритм кластерного анализа

Общий алгоритм кластерного анализа можно сформулировать в виде следующих шагов:

1. Реализовать иерархический кластерный анализ.
2. По итогам иерархического кластерного анализа выбрать число кластеров k, исходя из содержательных соображений и более формальных методов. Примеры методов:
   - метод согнутого локтя или согнутого колена (*Elbow method*);
   - силуэтный метод (*Silhouette method*).
3. Проверить с помощью формальных методов, насколько различия между группами существенны.
4. Реализовать более точный кластерный анализ с помощью метода k-средних (*k-means*) с выбранным числом групп k.
5. Сравнить результаты разных реализаций кластерного анализа в качестве проверки устойчивости, стабильности результатов.

1.2 Кластеризация методом k-means

Общая идея метода k-means (k-средних):

- На входе p-мерный массив, число кластеров k.
- На выходе хотим получить такое деление на кластеры, при котором внутригрупповой разброс минимален.

Остановимся на идее минимизации поподробнее. Для начала зафиксируем обозначения для кластеров и наблюдений.

Пусть $C_1, C_2, \dots, C_k$ – наборы (множества) наблюдений в каждом кластере, обладающие следующими свойствами:

1. $C_1 \cup C_2 \cup C_k = \{1, 2, \dots, n\}$ (все наблюдения распределены по кластерам);
2. $C_i \cap C_j \neq \emptyset$ для всех $i \neq j$ (никакие два кластера не пересекаются).

Пользуясь введёнными обозначениями, определим функцию внутригруппового разброса для кластера k:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2,$$

где $|C_k|$ – мощность множества C_k, то есть количество наблюдений в кластере k.

Содержательно, эта функция – просто мера среднего расстояния между точками кластера k: мы вычисляем квадраты евклидовых расстояний между всеми парами точек i и i' с учётом всех p измерений, суммируем результаты и делим на количество элементов.

Как будет выглядеть мера общего внутригруппового разброса? Очень просто – это сумма внутригрупповых разбросов всех кластеров от 1 до K:

$$W(C_1, \ldots, C_k) = \sum_{k=1}^{K} W(C_k).$$

Теперь задача метода k-means сводится к тому, чтобы подобрать такое разбиение на кластеры $C_1, C_2, \ldots, C_k$, чтобы этот суммарный внутригрупповой разброс минимизировать:

$$\sum_{k=1}^{k} W(C_k) \xrightarrow[C_1 \ldots C_K]{} \min .$$

Решить эту оптимизационную задачу точно довольно сложно, высока вычислительная сложность – всего существует K^n способов разбить n наблюдений на K групп. Поэтому на практике используется приближённый алгоритм нахождения разбиения на кластеры, который включает элемент случайности.

Приближённый алгоритм нахождения разбиения на кластеры методом k-means:

1. Случайным образом закрепляем за наблюдениями метки кластеров – числа от 1 до K.
2. Повторяем до тех пор, пока не получим наилучшее возможное качество – пока метки кластеров не перестанут меняться – следующие шаги:
   - определить центроид кластера;
   - приписать точку к кластеру, расстояние до центроида которого минимальное из всех возможных.

Приведённый выше алгоритм является эффективным и даёт хорошие результаты, однако стоит помнить, что, во-первых, он несёт в себе элемент случайности, а во-вторых, находит не глобальный, а локальный минимум функции $\sum_{k=1}^{k} W(C_k)$. Поэтому разумным решением будет запускать кластеризацию методом k-means несколько раз, сравнивать результаты и выбирать тот вариант, который лучшим образом подходит по содержательным соображениям и итогам предварительного иерархического кластерного анализа.