1 Методы кластерного анализа

2 Иерархические алгоритмы

3 Расстояния между кластерами

Возможные расстояния:

1. **Расстояние "ближайшего соседа"** (одиночная связь):

$$\rho_{\min}(K_i, K_j) = \min_{x_i \in K_i, x_j \in K_j} \rho(x_i, x_j).$$

   Равно расстоянию между самыми ближними объектами кластеров.

2. **Расстояние "дальнего соседа"** (полная связь):

$$\rho_{\max}(K_i, K_j) = \max_{x_i \in K_i, x_j \in K_j} \rho(x_i, x_j).$$

   Равно расстоянию между самыми дальними объектами кластеров.

3. **Невзвешенное попарное среднее**:

$$\rho_{\text{mean}}(K_i, K_j) = \text{mean}_{x_i \in K_i, x_j \in K_j} \rho(x_i, x_j).$$

   Равно среднему между всеми парами расстояний.

4. **Взвешенное попарное среднее**:

$$\rho_{\text{mean2}}(K_i, K_j) = \text{mean}_{x_i \in K_i, x_j \in K_j} \rho(k_1 \cdot x_i, k_2 \cdot x_j) = \text{mean}_{x_i \in K_i, x_j \in K_j} \frac{k_1 x_i + k_2 x_j}{k_1 + k_2}.$$

   Равно среднему между всеми парами расстояний с учетом весов k_1, k_2, равных ёмкости кластеров.

5. **Незвешенный центроидный метод**: расстояние между кластерами равно расстоянию между их центрами тяжести, где центр тяжести есть среднее арифметическое всех объектов в кластере:

$$x_c = \frac{\sum_{i=1}^{k} 1 \cdot x_i}{k} = \text{mean}_{x_i \in K_i}(x_i)$$

6. **Взвешенный центроидный метод**. Как я понял, это расстояние между центрами, но с учетом весов, как в пункте 4. Но вообще начиная с пункта 4 нигде нет формул, так что не факт, что они у меня правильные.

7. **Метод Варда**. Целевой функцией является внутригрупповая сумма квадратов:

$$\text{SS} = \sum_i \sum_{x_j \in K_i} \rho(x_j, x_i),$$

   где сумма берётся по всем кластерам, x_i – среднее по кластеру K_i. Объединение в кластеры на каждой итерации происходит так, чтобы увеличение этой функции было минимальным.

4 Процедуры эталонного типа

Наряду с иерархическими методами кластеризации существуют итеративные (k-средних), суть которых заключается в следующем (возможны модификации):

1. Среди объектов x_i некоторым образом выбираются k штук – центры будущих кластеров.
2. Объекты, не отнесённые к какому-либо кластеру, приписываются тому кластеру, до которого будет наименьшее расстрояние.
3. Центры кластеров пересчитываются.
4. Для всех точек пересматривается их принадлежность к кластеру.
5. Пункты 3-4 повторяются, пока точки могут менять принадлежность.

На этом сайте хорошо показано, как работает метод k-средних: `https://www.naftaliharris.com/blog/visualizing-k-means-clustering/`

5 Методика дискриминантного анализа

6 Что характеризует Лямбда Уилкса?

7 Что показывают квадраты расстояний Махаланобиса?

8 Какое максимальное число канонических дискриминантных функций допустимо в дискриминантном анализе?

9 Какую информацию дают стандартизованные и структурные коэффициенты дискриминантной функции?

10 Опишите процедуру отбора переменных с помощью стандартизованных и структурных коэффициентов

11 Какова интерпретация канонического коэффициента корреляции?

12 В каком случае учет априорных вероятностей может сильно изменить результаты классификации?

13 Методика факторного анализа

14 Суть задачи вращения общих факторов

15 Критерий Кайзера

16 Критерий Каменистой осыпи

17 Метод главных компонент