Меры сходства. Меры расстояния

Введение в кластерный анализ

Novosibirsk State University
*THE REAL SCIENCE

2GIS

Кластерный анализ

Кластерный анализ — это методы разбиения заданной задачи выборки объектов (ситуаций) на непересекающиеся подмножества (кластеры), так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Какие бывают меры сходства?

- Меры расстояния.
- Коэффициенты корреляции.
- Коэффициенты ассоциативности.

Меры расстояния

Меры расстояния (метрики) представляют объекты как точки в k-мерном простанстве, где размерность пространства определяется количеством переменных, которые используются для описания объектов.

Симметрия

Расстояние от объекта **x** до **y** должно быть таким же, как от **y** до **x**.

$$d(x,y) = d(y,x) \geq 0$$

Неравенство треугольника

$$d(x,y) \leq d(x,z) + d(y,z)$$

Различимость нетождественных объектов

$$x \neq y \Rightarrow d(x, y) \neq 0$$

Неразличимость идентичных объектов

$$x \equiv y \Rightarrow d(x,y) = 0$$

Евклидово расстояние

$$d_{ij} = \sqrt{\sum_{k=1}^{m} \left(x_{ik} - x_{jk} \right)^2}$$

Квадрат Евклидова расстояния

$$d_{ij} = \sum_{k=1}^{m} \left(x_{ik} - x_{jk} \right)^2$$

Манхеттенское расстояние

$$d_{ij} = \sum_{k=1}^{m} \left| x_{ik} - x_{jk} \right|$$

Расстояние Чебышёва

$$d_{ij} = \max_{k} \left| x_{ik} - x_{jk} \right|$$