Задание 14. Кластеризация

Санкт-Петербургский государственный университет

11 апреля 2025

Хочется: эффективно разбить n многомерных точек в p-мерном пространстве обоснованным, надежныи и экономным способом.

Что требуется в задаче:

- реализовать методы руками;
- сравнить сравниваемое; понимать, почему одно сравнимо, а другое нет;
- уметь говорить об интерпретации полученных результатов; т. е. отображать результаты и их интерпретировать;
- обязательно использование не только своих, но и прилагаемых тестов.

k-means

- Количество k кластеров задано.
- Начальные центры кластеров выбирают из каких-либо соображений.
- Перераспределяют данные по кластерам, используя какое-либо расстояние.
- Пересчитывают центры кластеров (центры масс) по отнесенным к ним точкам.
- Повторяют перераспределение данных и пересчет центров до тех пор, пока происходят изменения.

Можно выбирать оптимальное k.

Среднесвязывающий метод

- Это иерархический метод.
- О расстоянии между группами судят по расстоянию между центрами масс.
- Объединяют кластеры с ближайшими центрами.
- Т. е. каждое объединение уменьшает количество групп на единицу.

Вычисления проводить до получения какого-то заданного k количества кластеров.

Замечания: односвязывающий метод (где расстояние между группами определяется как расстояние между ближайшими членами групп) и полносвязывающий метод (расстояние между группами определяется как расстояние между самыми удаленными членами групп) считаются менее разумными, чем среднесвязывающий метод. Можно попробовать перепроверить это предположение.

Задание 14

- Реализовать два метода кластеризации.
- Каждый из прилагаемых тестов содержит матрицу целочисленных значений признаков (строка — информация по одному элементу; столбец — значения признаков для всех элементов).
  Первый столбец — количественный признак, остальные — номинальные. Необходимо провести кластеризацию:
  1. на основании только первого признака (первый столбец данных)
  2. на основании первого и одного или нескольких следующих (в идеале рассмотреть все возможные комбинации)
  3. помимо разноцветных картинок хочется знать центры и границы получившихся кластеров, а также количество элементов в каждом кластере.

ЗЫ. Тестов может стать больше...