

Лекция 5. Задача кластеризации

2024/2025 учебный год

Доцент кафедры ИВЭ, Махно В.В.

©Создано при помощи <https://sberuniversity.ru/>





Задача кластеризации

Задача заключается в том, что машина делит данные на кластеры по степени сходства между собой. Обратите внимание: если в задаче классификации нужно было заранее определить будущие классы, то при кластеризации количество будущих кластеров не известно. Сколько кластеров получится и какими они будут алгоритм определит самостоятельно.

Перед началом работы нужно задать машине признаки, по которым она будет кластеризовать данные. Машина видит данные как точки в пространстве. Расстояние между точками часто берется за метрику сходства в данных внутри кластера.

Проверить качество работы алгоритма кластеризации достаточно сложно, поскольку нам не с чем сравнить результат. Но некоторые параметры измерить все-таки можно. Например, компактность показывает, насколько близко друг к другу находятся объекты внутри одного кластера, а отделимость — насколько далеко находятся объекты из разных кластеров. Существуют также гибриды этих метрик и другие, более сложные индексы.

Кластеризация — это деление объектов на группы похожих объектов, эти группы называют кластерами.



Пример кластеризации экспертным методом

Рассмотрим клиентов кофейни (объект — клиент), признаки: сколько раз посещал(-а) кофейню утром, днем и вечером и сколько раз покупал позиции из меню (столько же признаков, сколько позиций: эспрессо, капучино, латте, маффины, сэндвичи и т. д.). В этом примере кластеризация задает деление клиентов на группы, соответственно, в каждой группе должны оказаться клиенты с похожими значениями признаков. Например, у нас могла бы получиться такая кластеризация клиентов кофейни на три группы:

- первая группа — работники близлежащего офисного центра, приходят в обеденное время, вместе с кофе обычно покупают салат или сэндвич;
- вторая группа — пассажиры автобусов, останавливающихся рядом с кофейней, приходят в разное время и покупают только кофе;
- третья группа — клиенты, приехавшие целенаправленно в данную кофейню после просмотра рекламы о ней, приходят в основном вечером и покупают фирменный напиток кофейни.

Приложения кластеризации

При анализе изображений кластеризацию можно использовать для сегментации изображений (выделения различных объектов на изображении и фона) — для этого в качестве объектов используют пиксели, в качестве признаков — цвет и расположение пикселей, и ищут группы похожих пикселей.

Кластеризация текстов делит тексты по темам, что может упростить навигацию по большому набору текстов (пользователь читает только интересные ему/ей темы).

В биоинформатике кластеризацию используют для группировки схожих геномных последовательностей в семейство генов, в медицине — для автоматического выделения различных типов тканей при диагностике.

Визуализация кластеризации

Геокластеризация, то есть кластеризация точек на карте, например точек продаж. Такая кластеризация позволяет, к примеру, найти удобные расположения для складов (склады размещают в центрах кластеров), чтобы было удобнее развозить товары со складов в точки продаж.



Алгоритмы кластеризации

k-Means

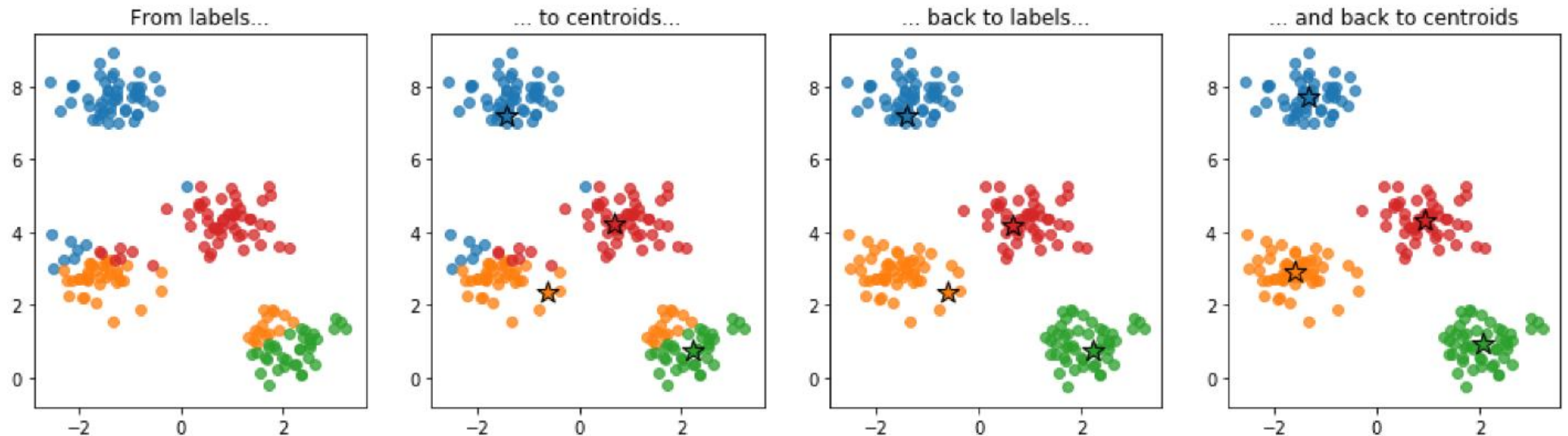
Самый известный и часто используемый метод кластеризации называется k-Means (метод k средних). Число k в названии говорит о том, что алгоритм находит k кластеров по данным. Схема работы алгоритма довольно проста. Каждый кластер (группа клиентов, которую мы хотим выделить) задается центром — неким типичным клиентом этой группы.

Алгоритм k-Means очень хорошо работает на несложных данных, например на данных, изображенных на следующем слайде. Если же данные сложнее, например требуется выделить кластеры с несколькими «центрами», то этот метод может не очень хорошо справиться с задачей. Кроме того, метод не очень хорошо работает, если число кластеров задано неправильно: тогда он будет разбивать кластеры на более мелкие или объединять несколько кластеров в один.

Алгоритмы кластеризации

k-Means

- В начале работы алгоритма мы назначаем эти центры случайно, например, выбираем произвольных клиентов в качестве центров.
- Затем мы выполняем кластеризацию: каждого клиента из данных записываем в тот кластер, на центр которого он больше всего похож. Получается k групп (кластеров) клиентов.
- Теперь, когда клиенты распределены по кластерам, мы находим новые центры кластеров: в каждой группе находим самого репрезентативного клиента.
- И далее все повторяется снова: заново распределяем клиентов по кластерам, используя новые центры, заново ищем центры и т.д.

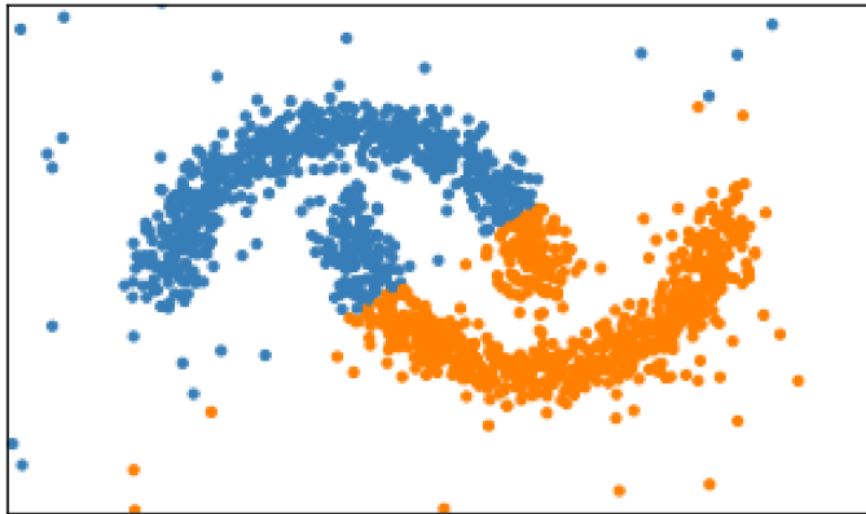


Алгоритмы кластеризации

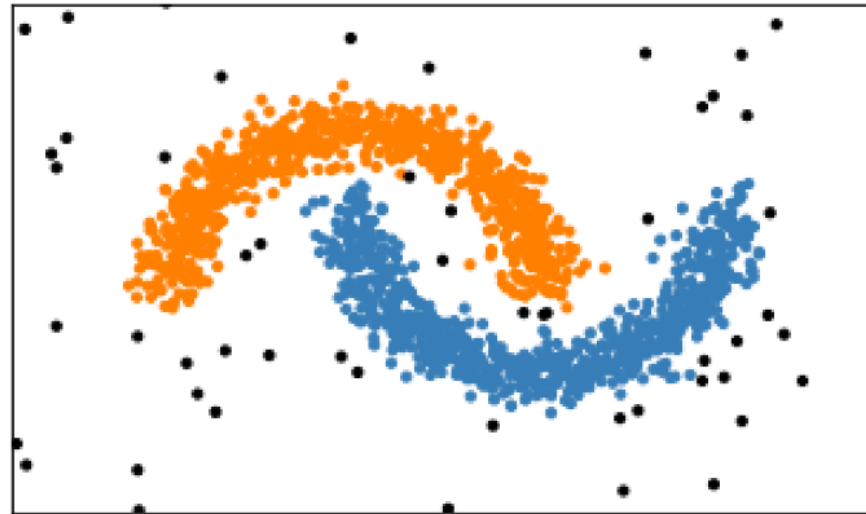
DBSCAN

Другой известный метод — DBSCAN (Density-Based Spatial Clustering of Applications with Noise) — требует задания минимального числа объектов в кластере и минимальной схожести объектов в кластере и затем сам определяет число кластеров. Интересной особенностью DBSCAN является то, что некоторые объекты он называет шумовыми и не относит их ни к одному кластеру. Пример данных, для которых хорошо работает DBSCAN (шумовые точки отмечены черным) и плохо работает k-Means:

KMeans



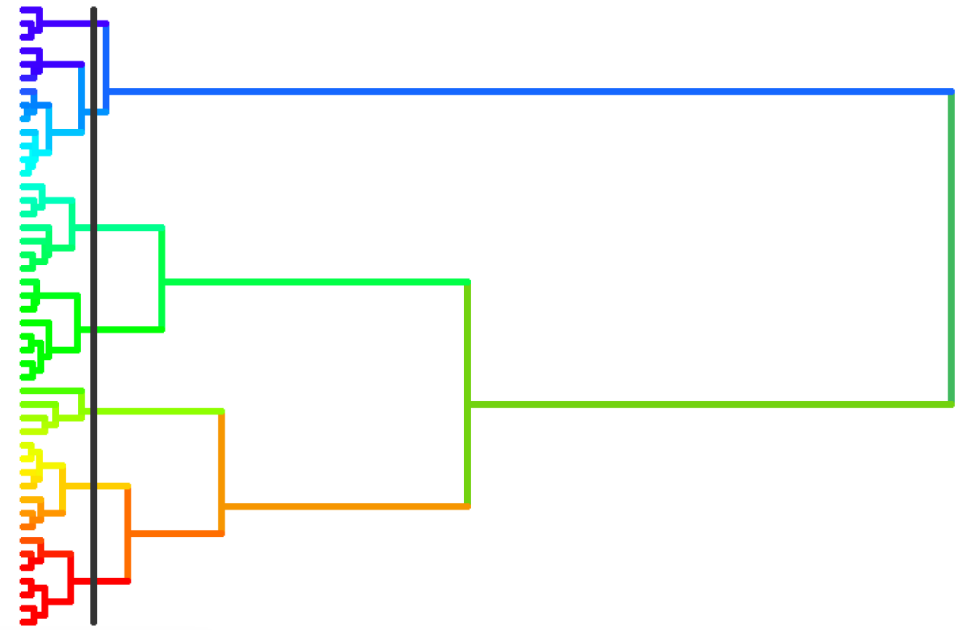
DBSCAN



Алгоритмы кластеризации

Иерархическая кластеризация

Иерархическая кластеризация (Agglomerative clustering), находит вложенные кластеры: например, в кластере «клиенты-студенты» могут быть выделены подкластеры «работающие студенты», «студенты с большими тратами в индустрии развлечений» и «иностранцы-студенты». Иерархическая кластеризация строит диаграммы следующего вида:



Инструменты для выполнения кластеризации

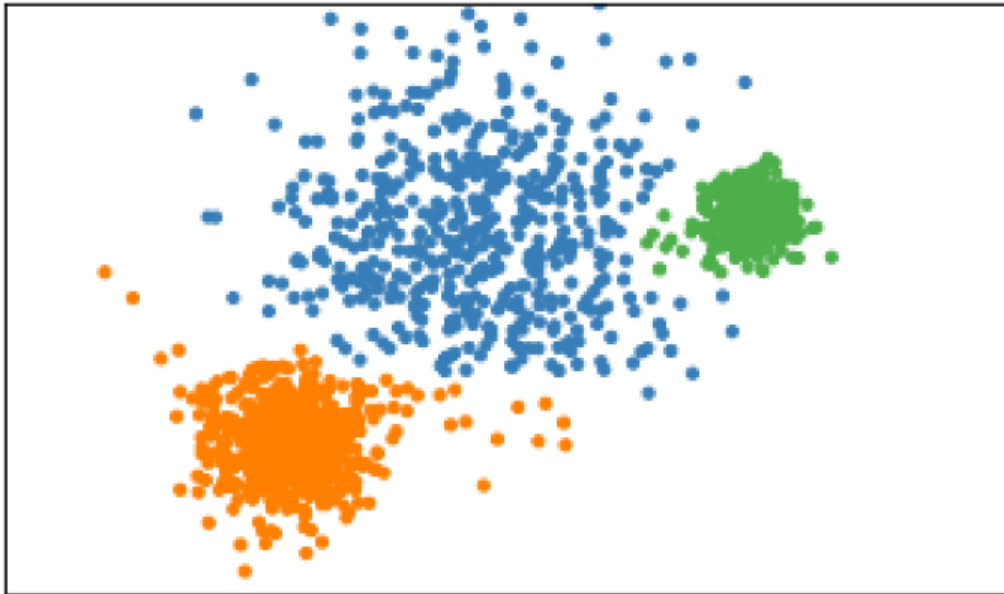
Для выполнения кластеризации по конкретным данным чаще всего используется язык программирования **Python** и библиотека **Scikit-learn** — в ней реализовано множество различных алгоритмов кластеризаций, в частности все, описанные выше.



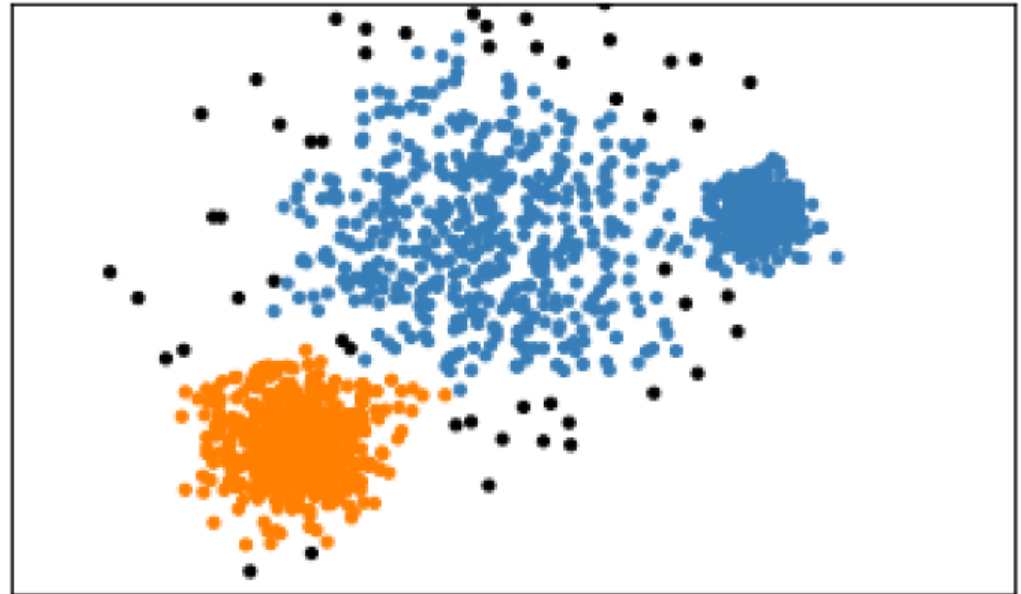
Измерение качества кластеризации

Задача кластеризации не является четко поставленной: как правило, в одних и тех же данных можно выделить разные кластеры, и не ясно, какая кластеризация лучше. Например, на изображении ниже можно выделить три кластера, а можно два (маленький кластер присоединится к большому).

Три кластера



Два кластера



Онлайн-курс СберУниверситета

Генеративное искусство

Подробнее о курсе



Бесплатный курс от Сбера
по генеративному
искусству

https://courses.sberuniversity.ru/generative-art?utm_source=tg&utm_medium=organic&utm_campaign=courses&utm_content=gen_i&utm_term=01_09_2023