Кластеризация

RS School Machine Learning course

План

1. Векторные и метрические пространства
2. Методы кластеризации
   - Иерархическая кластеризация
   - Алгоритм k средних
   - DBSCAN
3. Оценивание качества кластеризации
   - Силуэт
   - Adjusted Rand index

Напоминание

- Когда целевой переменной нет или её значения неизвестны, говорят, что перед нами задача машинного обучения «без учителя» (unsupervised learning).
- Типичный пример – кластеризация: разбить данные на несколько групп (классов, кластеров) наблюдений, похожих между собой.
- Количество кластеров иногда известно, иногда нет.
- Приложения кластеризации:
  - Сегментация аудитории
  - Опознавание аномалий в поведении пользователей
  - Моделирование тематики текстов
  - ...

Векторные пространства

- Данные для кластеризации (наблюдения, точки) обычно заданы в $\mathbb{R}^d$, вещественном d-мерном пространстве.
- Это одно из векторных пространств – объектов с богатой алгебраической и геометрической структурой.
- Элементы векторного пространства обычно называют векторами или точками.
- Обычная евклидова плоскость $\mathbb{R}^2$ – модельный пример, которым можно пользоваться, обсуждая свойства векторных пространств.

Векторные пространства

Алгебраические свойства:

- Векторы можно складывать.
- Можно умножать на число (будем считать, что из $\mathbb{R}$).
- Поэтому линейная комбинация векторов из какого-нибудь векторного пространства V принадлежит этому пространству:

$$\forall a,\ b \in \mathbb{R},\ u,\ v \in V : au + bv \in V$$

Векторные пространства

Геометрические свойства:

- В некоторых векторных пространствах, в частности, в $\mathbb{R}^d$, определено скалярное произведение:

$$\langle u,\ v\rangle = \sum_{i=1}^{d} u_i v_i \qquad (u,\ v \in \mathbb{R}^d)$$

- Скалярное произведение порождает норму:

$$\|u\| = \sqrt{\langle u,\ u\rangle}$$

- Норма порождает метрику (берётся линейная комбинация векторов):

$$\rho(u,\ v) = \|u - v\|$$

Метрические пространства

- Метрическим пространством называется пара (X, ρ), где
  X – произвольное множество;
  $\rho : X \times X \to \mathbb{R}$ – функция с набором хороших свойств (далее о них).
- Элементы множества X называют точками.
- Функцию ρ называют метрикой.
- О величине $\rho(x, y)$ можно думать как о расстоянии или длине кратчайшего пути между точками x и y.

Метрические пространства

Свойства метрики:

1. Неотрицательность:

$$\forall x,\, y \in X \quad \rho(x,\, y) \geq 0$$

(У любого пути неотрицательная длина.)

2. Условие обращения в ноль:

$$\forall x,\, y \in X \quad \rho(x,\, y) = 0 \Leftrightarrow x = y$$

(У пути нулевой длины совпадают начало и конец, и наоборот.)

3. Симметричность:

$$\forall x,\, y \in X \quad \rho(x,\, y) = \rho(y,\, x)$$

(Путь между точками имеет одинаковую длину в обе стороны.)

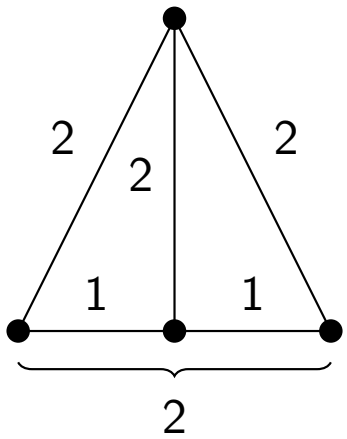
4. Неравенство треугольника:

$$\forall x,\, y,\, z \in X \quad \rho(x,\, y) + \rho(y,\, z) \geq \rho(x,\, z)$$

(Путь через промежуточный пункт не бывает короче, чем напрямик.)

Метрические пространства

- Таким образом, векторное пространство со скалярным произведением (такое, как $\mathbb{R}^d$) сразу является метрическим пространством.
- Но не на любом метрическом пространстве можно задать структуру векторного пространства.
- Есть метрические пространства, которые невозможно вложить в $\mathbb{R}^d$ при любом d:



Опять к кластеризации

- Метрическая структура «беднее», чем векторная
  $\Rightarrow$ больше разных объектов обладают ею.
- Алгоритмы, о которых мы будем говорить:
  - работают в произвольных метрических пространствах
  - либо нуждаются для этого в незначительных изменениях
- Поэтому мы можем считать, что наблюдения живут в метрическом пространстве (X, ρ).
- Совокупность данных обозначим $U = \{x_1, \ldots, x_n\}$.
- У многих алгоритмов кластеризации один настраиваемый параметр k – количество кластеров, которое нужно получить.

Иерархическая кластеризация

- ▷ Инициализируем n кластеров, в каждом по одному наблюдению.
- ▷ На каждом шаге будем брать два кластера, самых близких друг к другу, и объединять их.
- ▷ Таким образом, на каждом шаге число кластеров уменьшается на единицу.
- ▷ Когда останется ровно k кластеров, остановимся.
- Это иерархическая (или агломеративная) кластеризация.
- **Проблема:** Что такое «кластеры, самые близкие друг к другу»?

Критерии близости кластеров

- Как задать «расстояние» D между двумя кластерами M и N?

▷ Single-linkage clustering:

$$D(M,\, N) = \min\{\, \rho(x,\, y) \,|\, x \in M,\, y \in N \,\}$$

▷ Complete-linkage clustering:

$$D(M,\, N) = \max\{\, \rho(x,\, y) \,|\, x \in M,\, y \in N \,\}$$

▷ Average-linkage clustering:

$$D(M,\, N) = \frac{1}{|M||N|} \sum_{x \in M} \sum_{y \in N} \rho(x,\, y)$$

- Эти и другие критерии близости кластеров можно представить как частные случаи формулы Ланса–Уильямса (мы не будем её обсуждать подробнее).

Алгоритм k средних

- Будем считать, что наблюдения даны в $\mathbb{R}^d$.
- ▷ Случайным образом выберем k наблюдений x_{i_1}, ..., x_{i_k} в качестве центров.
- ▷ Сформируем k кластеров: в кластер с центром x_i входят наблюдения, которые ближе к x_i, чем к любому другому из центров.
- ▷ На каждом шаге:
  - Пересчитаем центры, усреднив каждый кластер.
  - Переформируем кластеры относительно новых центров.
- ▷ Когда кластеры перестают меняться, остановимся.

Алгоритм k средних

- Это алгоритм k средних (k-means).
- Модификация для произвольного метрического пространства: центрами могут быть только точки, представленные в данных (алгоритм k-medoids).
- **Проблема:** При разном выборе начальных точек получаются разные результаты.
- Улучшенный способ первоначального выбора центров:

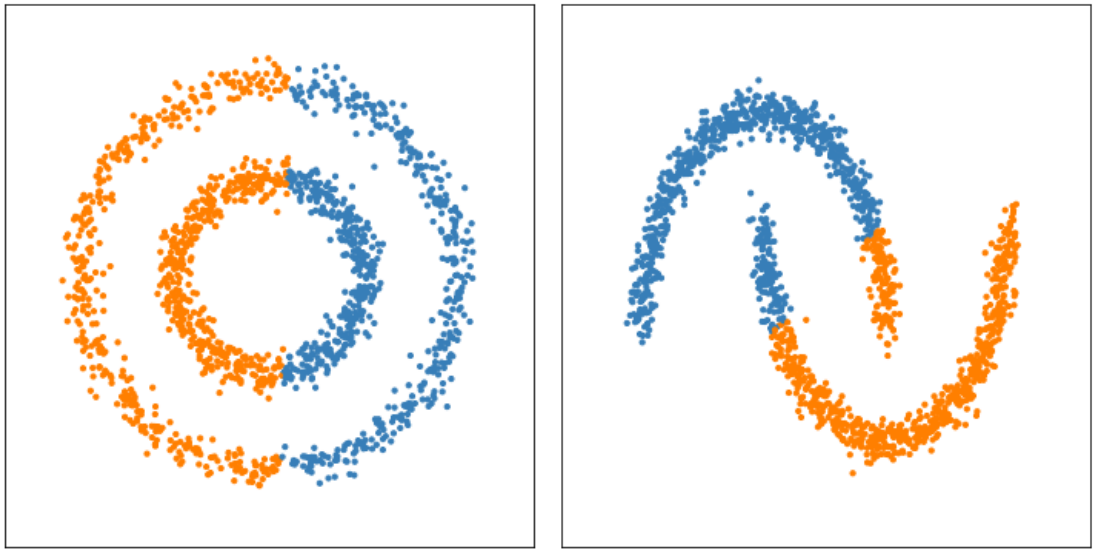
▷ Первый центр выберем случайно (равномерно по всем наблюдениям).

▷ Каждый следующий будем выбирать с вероятностью, обратно пропорциональной квадрату расстояния данной точки до ближайшего уже выбранного центра.

- Этот способ инициализации называется k-means++.

DBSCAN

- В некоторых ситуациях (кластеры сложной формы, зашумлённые данные) иерархическая кластеризация и алгоритм k средних работают плохо.



- Не будем фиксировать число кластеров, посмотрим на локальную структуру данных.

DBSCAN

- Введём два параметра: радиус ε и минимальная плотность m.
- Для каждой точки x_i рассмотрим её окрестность $U_\varepsilon(x_i) = \{x_j \in U \mid i \neq j,\ \rho(x_i, x_j) \leq \varepsilon\}$.
- Подразделим все точки на:
  - центральные: $|U_\varepsilon(x_i)| \geq m$
  - периферийные: $1 \leq |U_\varepsilon(x_i)| < m$
  - шумовые: $|U_\varepsilon(x_i)| = 0$
- На всех центральных точках построим граф: две точки в окрестности друг друга соединяются ребром.
- Одна связная компонента графа = один кластер.
- Каждую периферийную точку относим к тому же кластеру, что и ближайшая центральная.
- Шумовые точки не кластеризуем.

DBSCAN

- Это алгоритм DBSCAN.
- ⊕ На практике обычно даёт хорошие результаты.
- ⊖ Не параметризуется число кластеров.
- Можно использовать для фильтрации шума в данных, т. е. сами кластеры отбрасываются.

Оценивание качества кластеризации

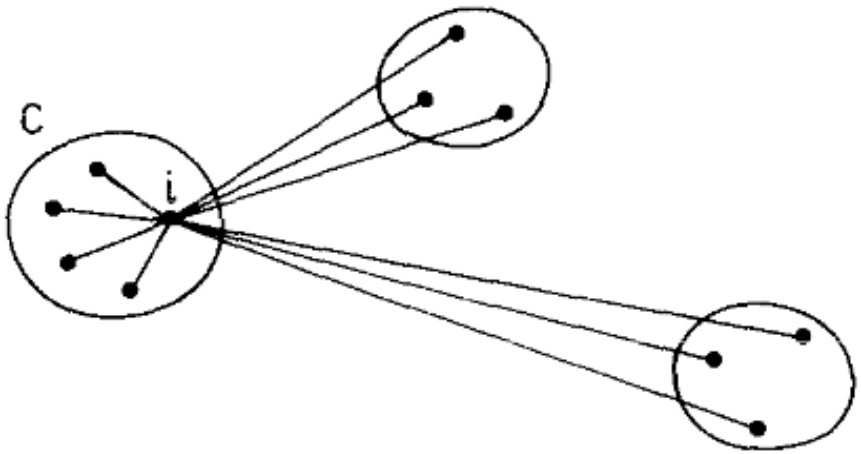
- Для подбора параметров нужно уметь оценивать, насколько кластеризация хороша.
- Геометрические характеристики: насколько плотные кластеры, далеко ли отстоят друг от друга...
- Если есть размеченные данные, можно сравнить с эталонной кластеризацией.

Силуэт

- Пусть выборка $x_1, \ldots, x_N$ разбита на кластеры.
- Для точки x_i, отнесённой к кластеру C, рассмотрим две величины:

  a_i – среднее расстояние от x_i до других точек в C;

  b_i – минимум, по всем остальным кластерам C', среднего расстояния от x_i до точек в C'.



Силуэт

- Силуэтом (silhouette) x_i называется разность $b_i - a_i$, нормированная до отрезка $[-1, 1]$.
- Усредняем по всей выборке.
- Интуитивно, тем ближе к 1,
  - чем ближе расположены точки внутри кластеров;
  - чем дальше кластеры друг от друга.

Adjusted Rand index

- Пусть для выборки $x_1, \ldots, x_N$ есть эталонная кластеризация $\mathcal{C}$, а мы построили кластеризацию $\mathcal{D}$.
- Насколько наша кластеризация близка к эталону?
  - ▷ Количество и нумерация кластеров могут различаться.
- Рассмотрим всевозможные пары точек x_i, x_j, их всего $\binom{N}{2}$.
- Обозначим:
  - a – # пар, которые в $\mathcal{C}$ попадают в один кластер и в $\mathcal{D}$ тоже;
  - b – # пар, которые в $\mathcal{C}$ попадают в разные кластеры и в $\mathcal{D}$ тоже;
  - c – # пар, которые в $\mathcal{C}$ попадают в один кластер, а в $\mathcal{D}$ в разные;
  - d – # пар, которые в $\mathcal{C}$ попадают в разные кластеры, а в $\mathcal{D}$ в один.

Adjusted Rand index

- Индексом Рэнда называется величина

$$\frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{N}{2}}.$$

- Заключён между 0 и 1, равен 1 для разбиений, которые различаются только нумерацией кластеров.
- «Исправленный» индекс Рэнда: вычитается математическое ожидание суммы $a+b$ по всевозможным парам разбиений, однотипных с $\mathcal{C}$ и $\mathcal{D}$ по численности кластеров.