Машинное обучение: Кластеризация

Н. Поваров, И. Куралёнок

СПб,
2019

Что будет сегодня про кластеризацию

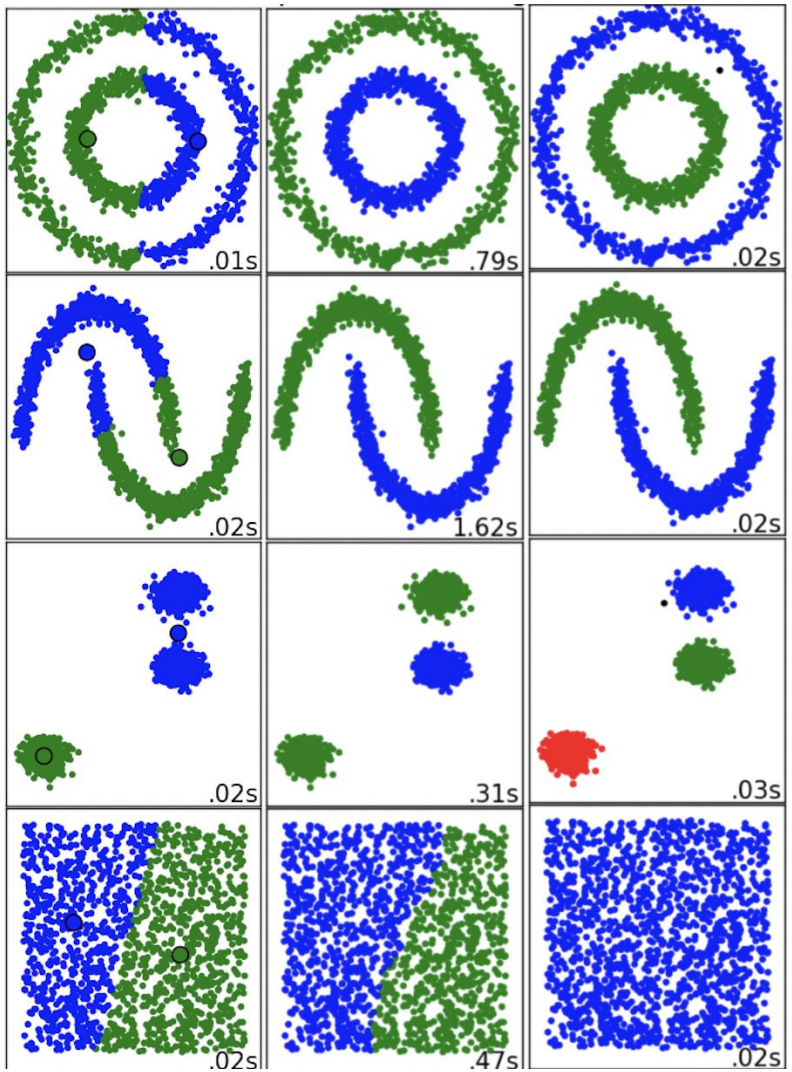
1. Что это и зачем
2. Описание нескольких алгоритмов

Что такое кластеризация

Кластерный анализ (англ. cluster analysis):

задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Картинка



Другая картинка

Виды кластеризации

1. Centroid-based
2. Connectivity-based
3. Distribution-based
4. Constraint-based

Виды кластеризации

1. Centroid-based
2. Connectivity-based
3. Distribution-based
4. Constraint-based

Другие классификации

1. Чёткие/нечёткие
2. Плоские/иерархические

Другие классификации

1. **Чёткие**/нечёткие
2. Плоские/иерархические

Другие классификации

1. **Чёткие**/нечёткие
2. **Плоские**/иерархические

Centroid-based

FOREL (centroid-based**)**

1. Случайно выбираем точку x и объявляем центром масс
2. Ищем соседей этого центра масс ближе, чем R
3. Вычисляем их центр масс
4. Повторяем шаги 2-3, пока новый центр масс не совпадет с прежним
5. Помечаем все точки внутри сферы радиуса R вокруг текущего центра масс как очередной кластер и выкидываем их из выборки
6. Повторяем шаги 1-5, пока не будет кластеризована вся выборка

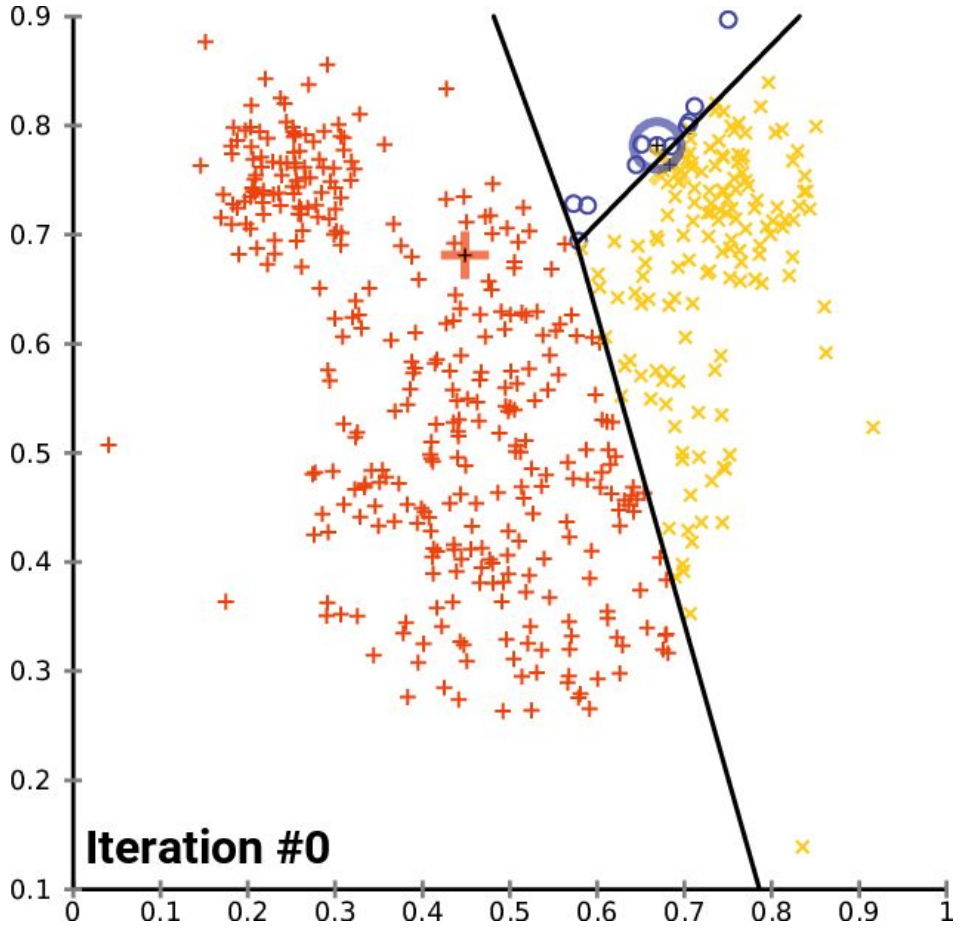
Факты о FOREL

1. Один параметр!
2. Результат зависит от рандома
3. Детище советских учёных

k-means (centroid-based)

1. Выбираем k точек и объявляем их центрами масс
2. Для каждой точки из выборки ищем ближайший центр масс и относим её к кластеру этого центра масс
3. Вычисляем новые центры масс
4. Повторяем шаги 2-3, пока не сойдётся

k-means (centroid-based)



Факты о k-means

1. Один параметр!
2. Нет гарантий сходимости
3. Можно использовать medians
4. А можно использовать medoids

Правило локтя для k-means

1. Вычисляем сумму квадратов расстояний от точек до центров
2. Рисуем график
3. Выбираем k

Правило локтя для k-means

Connectivity-based

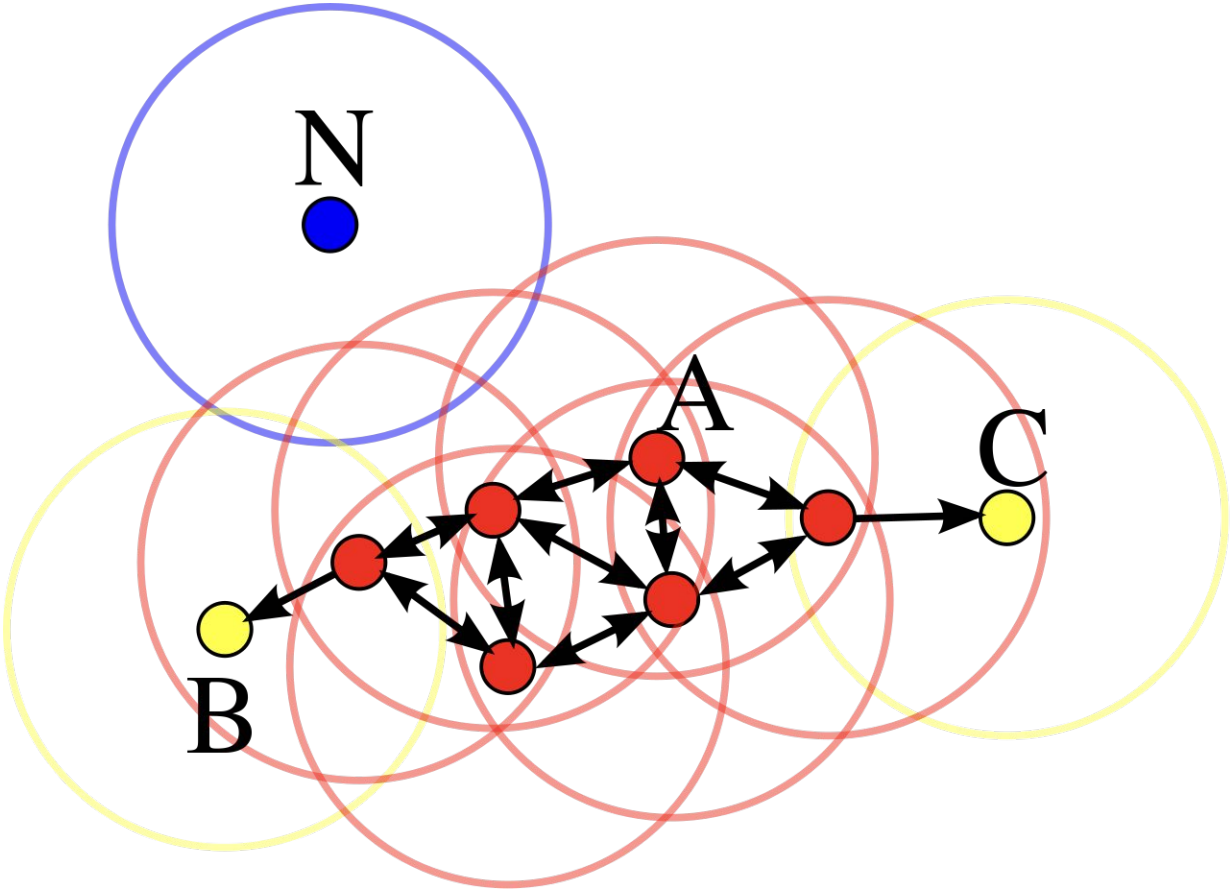
Односвязный (connectivity-based)

1. Выбираем случайную точку x и кладём её в кластер C_i
2. Ищем соседей x ближе, чем ε и добавляем их в C_i
3. Если соседи кончились, то берём следующую точку из C_i и проделываем п. 2
4. Если в C_i больше не добавить точек, то переходим к п. 1

Факты про односвязный алгоритм

-
- Один параметр!
- Сложность $O(n^3)$
- Можно сделать хитрее, чем односвязный

DBSCAN (connectivity-based)



DBSCAN (connectivity-based**)**

1. Выбираем случайную непосещённую точку
2. Если у неё больше, чем m соседей в радиусе ε, то кладём её в S и создаём новый кластер, иначе помечаем как шум
3. Для каждой точки из S:
    1. Если эту точку не посещали, то ищем всех соседей в радиусе ε
    2. Если таких соседей больше m, то добавляем их всех в S
    3. Если точка не в каком либо кластере, то добавляем в текущий кластер

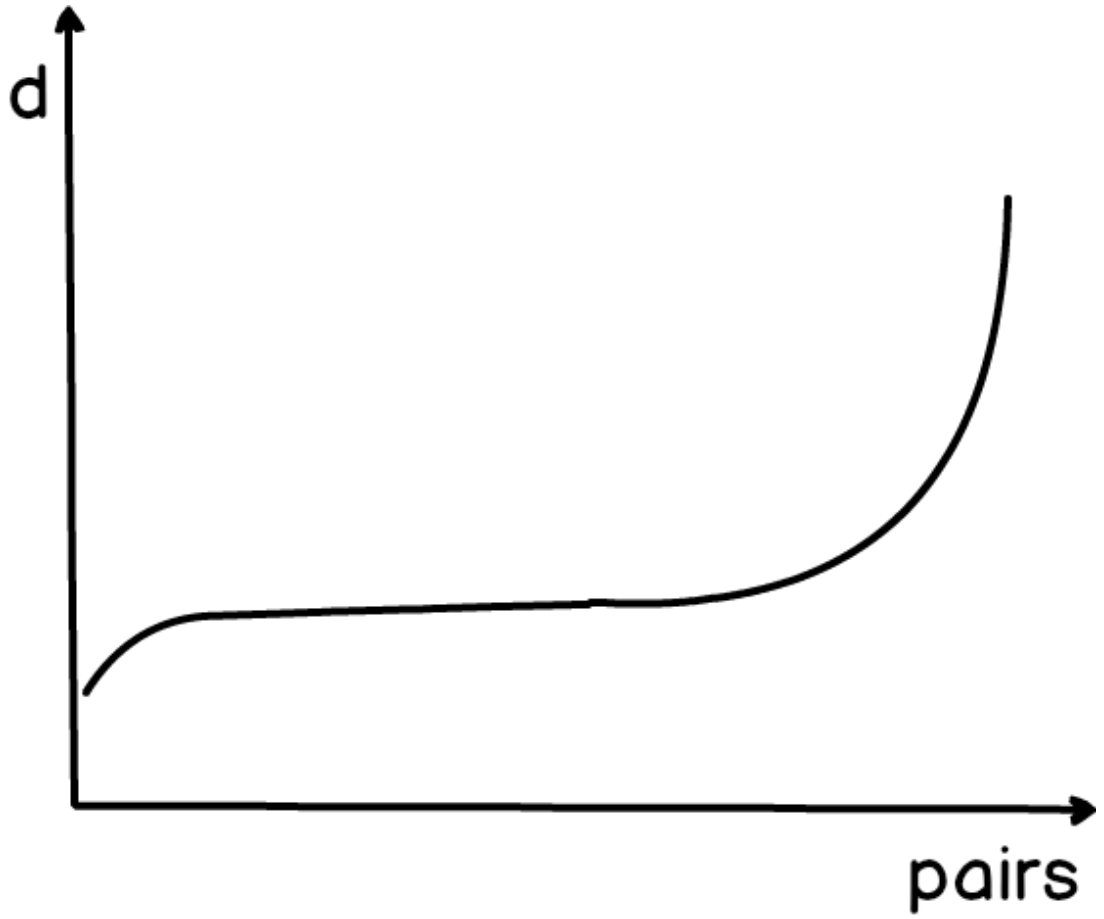
Факты о DBSCAN

1. Два параметра ☹
2. Сложность наивной реализации $O(n^2)$
3. Пограничные точки не детерминировано разбиваются
4. Можно ускорить с помощью kd-tree
5. Тут есть проклятье размерности
6. В 2014 на KDD алгоритм получил «test of time award»

«Правило локтя» для DBSCAN

1. Выбираем m
2. Для каждой точки считаем среднее расстояние до m ближайших соседей
3. Сортируем и рисуем график
4. На графике выбираем ε

«Правило локтя» для DBSCAN



Интересные примеры DBSCAN

- Convex-hull & DBSCAN clustering to predict future weather, 2015
- Modelling website user behaviors by combining the EM and DBSCAN algorithms, 2016
- Real-Time Superpixel Segmentation by DBSCAN Clustering Algorithm, 2016

Метрики качества кластеризации

- Внешние
- Внутренние

Внешние метрики

- Rand measure = Accuracy
- F-мера
- Jaccard
- …

Внутренние метрики

- Вводят внутрикластерное расстояние
- Вводят межкластерное расстояние
- Первое хотят минимизировать, второе максимизировать

Внутренние метрики

- Dunn Index
- Силуэт
- Davies-Bouldin index
- …

Dunn index

•

$$D = \frac{\min\limits_{c_k \in C} \left\{ \min\limits_{c_l \in C \setminus c_k} \{\delta(c_k, c_l)\} \right\}}{\max\limits_{c_k \in C} \{\Delta(c_k)\}}$$

$$\delta(c_k, c_l) = \min_{x_i \in c_k, x_j \in c_l} \|x_i - x_j\|$$

$$\Delta(c_k) = \max_{x_i, x_j \in c_k} \|x_i - x_j\|$$

Silhouette

- $$Sil = \frac{1}{N} \sum_{c_k \in C} \sum_{x_i \in c_k} \frac{b(x_i, c_k) - a(x_i, c_k)}{\max\{a(x_i, c_k), b(x_i, c_k)\}}$$

$$a(x_i, c_k) = \frac{1}{|c_k|} \sum_{x_j \in c_k} \|x_i - x_j\|$$

$$b(x_i, c_k) = \min_{c_l \in C \setminus c_k} \left\{ \frac{1}{|c_l|} \sum_{x_j \in c_l} \|x_i - x_j\| \right\}$$

Davies-Bouldin index

•

$$DB = \frac{1}{K} \sum_{c_k \in C} \max_{c_l \in C \setminus c_k} \left\{ \frac{S(c_k) + S(c_l)}{\|\bar{c}_k - \bar{c}_l\|} \right\}$$

$$\mathrm{S}(c_k) = \frac{1}{|c_k|} \sum_{x_i \in c_k} \|x_i - \bar{c}_k\|$$

Итого

- Есть разные методы кластеризации
- Возможно, что от них есть польза
- Их качество пытаются измерять