ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA TOÁN - TIN HỌC

BTLT TUẦN 8: KHAI THÁC DỮ LIỆU

GVHD: NGUYỄN THANH BÌNH
Sinh viên thực hiện: Nguyễn Công Hoài Nam
Mã số sinh viên: 21280099

Ngày 4 tháng 5 năm 2024

1. *Consider the 1-dimensional data set with 10 data points {1, 2, 3, . . . , 10}. Show three iterations of the k - means algorithms when k = 2, and the random seeds are initialized to {1, 2}.*

Khởi tạo trọng tâm cho hai cluster lấy từ seeds:

$$\text{centroid của } c_1 = 1$$
$$\text{centroid của } c_2 = 1$$

Tính khoảng cách của mỗi điểm dữ liệu đến centroid của các cluster và cập nhật chung vào cluster có khoảng cách nhỏ nhất

- Iteration 1

$$\min(|1-1|, |1-2|) = 0 \quad \text{thuộc } c_1$$
$$\min(|2-1|, |2-2|) = 0 \quad \text{thuộc } c_1$$

Tương tự cho toàn bộ dữ liệu

i	|i-1|	|i-2|	$i \in$
1	0	0	c_1
2	1	0	c_2
3	2	1	c_2
4	3	2	c_2
5	4	3	c_2
6	5	4	c_2
7	6	5	c_2
8	7	6	c_2
9	8	7	c_2
10	9	8	c_2

Các điểm còn lại cũng thuộc về c_2
Vì vậy

$$c_1 = \{1\}$$
$$c_2 = \{2, 3, \ldots, 10\}$$

- Iteration 2
Cập nhật lại centroid của cluster c_1, c_2

$$\text{centroid của } c_1 = \text{mean}(c_1) = 1$$
$$\text{centroid của } c_2 = \text{mean}(c_2) = \text{mean}(2, 3, \ldots, 10) = \frac{54}{9} = 6$$

Tương tự, ta tính khoảng cách và gán chúng vào các cluster

i	|i-1|	|i-6|	$i \in$
1	0	5	c_1
2	1	4	c_1
3	2	3	c_1
4	3	2	c_2
5	4	1	c_2
6	5	0	c_2
7	6	1	c_2
8	7	2	c_2
9	8	3	c_2
10	9	4	c_2

Vì vậy

$$c_1 = \{1, 2, 3\}$$
$$c_2 = \{4, 5, \ldots, 10\}$$

- Iteration 3
  Cập nhật lại centroid của cluster c_1, c_2

$$\text{centroid của } c_1 = mean(1, 2, 3) = \frac{6}{3} = 2$$
$$\text{centroid của } c_2 = mean(4, 5, \ldots, 10) = \frac{49}{7} = 7$$

Và ta cũng tính khoảng cách tới các centroid

i	|i-2|	|i-7|	$i \in$
1	1	6	c_1
2	0	5	c_1
3	1	4	c_1
4	2	3	c_1
5	3	2	c_2
6	4	1	c_2
7	5	0	c_2
8	6	1	c_2
9	7	2	c_2
10	8	3	c_2

Có được:

$$c_1 = \{1, 2, 3, 4\}$$
$$c_2 = \{5, 6, \ldots, 10\}$$

Trên đây là 3 lần lặp của thuật k-means trên bộ dữ liệu.

2. *Consider the 1-dimensional data set {1, . . . , 10}. Apply a hierarchical agglomerative approach, with the use of minimum, maximum, and group average criteria for merging. Show the first six merges.*

Phương pháp gom cụm phân cấp với các tiêu chí tối thiểu, tối đa và trung bình, 6 cụm gộp đầu tiên là:

- Gộp (1, 2)
- Gộp (3, 4)
- Gộp (5, 6)
- Gộp (7, 8)
- Gộp (9, 10)
- Gộp (1, 2, 3, 4)

Thứ tự gộp này giống nhau cho cả ba tiêu chí (tối thiểu, tối đa và trung bình), vì chúng đều sử dụng tiêu chí tìm nút có chỉ số nhỏ nhất trong tất cả các gộp cùng chất lượng. Điều này dẫn đến việc chúng ta có cùng một chuỗi gộp đầu tiên.