Bài tập Lý thuyết môn Khai phá Dữ liệu

Họ và tên: Huỳnh Nguyễn Thế Dân

MSSV: 21110256

Lớp: 21TTH1

Questions

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}.

5. 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.

Answer

*For exercise 1:

First Iteration

- **Initial Centroids:**
  - Cluster 1: Centroid $m_1 = 1$
  - Cluster 2: Centroid $m_2 = 2$
- **Cluster Assignment:**
  - Data {1}: closest to cluster 1 (distance 0)
  - Data {2}: closest to cluster 2 (distance 0)
  - Data {3}: closest to cluster 2 (distance 1)
  - Data {4, 5, 6, 7, 8, 9, 10}: closest to cluster 2
- **Updated Centroids:**
  - Cluster 1: $m_1 = \frac{1}{1} = 1$
  - Cluster 2: $m_2 = \frac{2+3+4+5+6+7+8+9+10}{9} = 6$

Second Iteration

- **Initial Centroids:**
  - Cluster 1: $m_1 = 1$
  - Cluster 2: $m_2 = 6$
- **Cluster Assignment:**
  - Data {1, 2, 3}: closest to cluster 1
  - Data {4, 5, 6, 7, 8, 9, 10}: closest to cluster 2
- **Updated Centroids:**
  - Cluster 1: $m_1 = \frac{1+2+3}{3} = 2$
  - Cluster 2: $m_2 = \frac{4+5+6+7+8+9+10}{7} = 7$

Third Iteration

- **Initial Centroids:**
  - Cluster 1: $m_1 = 2$
  - Cluster 2: $m_2 = 7$
- **Cluster Assignment:**
  - Data {1, 2, 3, 4}: closest to cluster 1
  - Data {5, 6, 7, 8, 9, 10}: closest to cluster 2
- **Updated Centroids:**
  - Cluster 1: $m_1 = \frac{1+2+3+4}{4} = 2.5$
  - Cluster 2: $m_2 = \frac{5+6+7+8+9+10}{6} = 7.5$

---

*For exercise 5:

Let's apply a hierarchical agglomerative clustering approach to the 1-dimensional data set {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}. We'll consider three different criteria for merging clusters: minimum (single linkage), maximum (complete linkage), and group average.

Minimum (Single Linkage)

Start with each point as its own cluster.

- Merge 1: The closest clusters are {1} and {2}, distance 1.
  - Merged cluster: {1, 2}
- Merge 2: The closest clusters are {3} and {1, 2}, distance 1.
  - Merged cluster: {1, 2, 3}
- Merge 3: The closest clusters are {4} and {1, 2, 3}, distance 1.
  - Merged cluster: {1, 2, 3, 4}
- Merge 4: The closest clusters are {5} and {1, 2, 3, 4}, distance 1.

    - Merged cluster: {1, 2, 3, 4, 5}
- Merge 5: The closest clusters are {6} and {1, 2, 3, 4, 5}, distance 1.
    - Merged cluster: {1, 2, 3, 4, 5, 6}
- Merge 6: The closest clusters are {7} and {1, 2, 3, 4, 5, 6}, distance 1.
    - Merged cluster: {1, 2, 3, 4, 5, 6, 7}

Maximum (Complete Linkage)

Start with each point as its own cluster.

- Merge 1: The closest clusters are {1} and {2}, distance 1.
    - Merged cluster: {1, 2}
- Merge 2: The closest clusters are {3} and {4}, distance 1.
    - Merged cluster: {3, 4}
- Merge 3: The closest clusters are {5} and {6}, distance 1.
    - Merged cluster: {5, 6}
- Merge 4: The closest clusters are {7} and {8}, distance 1.
    - Merged cluster: {7, 8}
- Merge 5: The closest clusters are {9} and {10}, distance 1.
    - Merged cluster: {9, 10}
- Merge 6: The closest clusters are {1, 2} and {3, 4}, distance 2.
    - Merged cluster: {1, 2, 3, 4}

Group Average

Start with each point as its own cluster.

- Merge 1: The closest clusters are {1} and {2}, distance 1.
    - Merged cluster: {1, 2}
- Merge 2: The closest clusters are {3} and {1, 2}, distance 1.5.
    - Merged cluster: {1, 2, 3}
- Merge 3: The closest clusters are {4} and {1, 2, 3}, distance 2.
    - Merged cluster: {1, 2, 3, 4}
- Merge 4: The closest clusters are {5} and {1, 2, 3, 4}, distance 2.5.
    - Merged cluster: {1, 2, 3, 4, 5}
- Merge 5: The closest clusters are {6} and {1, 2, 3, 4, 5}, distance 3.
    - Merged cluster: {1, 2, 3, 4, 5, 6}
- Merge 6: The closest clusters are {7} and {1, 2, 3, 4, 5, 6}, distance 3.5.
    - Merged cluster: {1, 2, 3, 4, 5, 6, 7}

End.