ENGG 5103: Techniques for data mining

Wanli Wang
2020fall-ENGG5103-HW1-Wang-Wanli-1155160517

I. K-MEANS CLUSTERING (25%)

Given the following points, present each iteration of k-means clustering algorithm until convergence. The distance is measured using Euclidean distance.

$P1 : (1, 1.5), P2 : (2, 3), P3 : (4, 5), P4 : (5, 7), P5 : (3.5, 4.5), P6 : (2.5, 5.5)$

(1) Suppose $k = 2$, initial guess is $(1, 2)$ and $(3, 4)$.

Initial Center: $(1, 2), (3, 4)$.

After one iteration, we have clustering centers $(1.5, 2.25), (3.75, 5.5)$.

The first clustering contains points $P1 : (1, 1.5), P2 : (2, 3)$.

The second clustering contains points $P3 : (4, 5), P4 : (5, 7), P5 : (3.5, 4.5), P6 : (2.5, 5.5)$.

(2) Suppose $k = 2$, initial guess is $(2, 5)$ and $(3, 6)$.

Initial Centers: $(2, 5)$ and $(3, 6)$.

After one iteration, we have clustering centers $(2.25, 3.625), (4.5, 6.0)$.

The first clustering contains points $P1 : (1, 1.5), P2 : (2, 3), P5 : (3.5, 4.5), P6 : (2.5, 5.5)$

The second Clustering contains points $P3 : (4, 5), P4 : (5, 7)$

(3) Suppose $k = 3$, initial guess is $(1, 2), (3, 4)$ and $(2, 3)$.

Initial Center: $(1, 2), (3, 4)$ and $(2, 3)$.

After one iteration, we have clustering centers (1.0, 1.5), (2.0, 3.0), (3.75, 5.5).

The first clustering contains point $P1 : (1, 1.5)$.

The second clustering contains point $P2 : (2, 3)$.

The third clustering contains points $P3 : (4, 5), P4 : (5, 7), P5 : (3.5, 4.5), P6 : (2.5, 5.5)$.

Code is as follows: https://github.com/wangwanlitype1/ENGG5103

II. CLUSTER COHESION AND SEPARATION (25%)

Given the following points, answer the following questions.

P1: (1,1.5) P2: (2, 3) P3: (4, 5) P4: (5,7) P5: (3.5, 4.5) P6: (2.5, 5.5)

(1) Suppose they are divided into two clusters. P1, P2 and P3 belong to cluster 1, P4, P5 and P6 belong to cluster 2. Please calculate the SSE, SSB and TSS.

Answer:

The P1, P2 and P3 belong to cluster 1, we therefore have corresponding $c_1 : (7/3, 9.5/3)$.

The P4, P5 and P6 belong to cluster 2, we therefore have corresponding $c_2 : (11/3, 17/3)$.

The SSE,SSB and TSS are defined as:

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} (x - c_i)^2 \tag{1}$$

$$\text{SSB} = \sum_{i=1}^{K} |C_i|(c - c_i)^2 \tag{2}$$

$$\text{TSS} = \sum_{i=1}^{K} \sum_{x \in C_i} (x - c)^2 \tag{3}$$

We can calculate c as

$$c : (0.5 * (7/3 + 11/3), 0.5 * (9.5/3 + 17/3)) = (3, 13.25/3) \tag{4}$$

$$\begin{aligned} \text{SSE} = &\begin{bmatrix} 1 - 7/3 \\ 1.5 - 9.5/3 \end{bmatrix}^2 + \begin{bmatrix} 2 - 7/3 \\ 3 - 9.5/3 \end{bmatrix}^2 + \begin{bmatrix} 4 - 7/3 \\ 5 - 9.5/3 \end{bmatrix}^2 \\ &+ \begin{bmatrix} 5 - 11/3 \\ 7 - 17/3 \end{bmatrix}^2 + \begin{bmatrix} 3.5 - 11/3 \\ 4.5 - 17/3 \end{bmatrix}^2 + \begin{bmatrix} 2.5 - 11/3 \\ 5.5 - 17/3 \end{bmatrix}^2 = \begin{bmatrix} 47/6 \\ 28/3 \end{bmatrix} \end{aligned} \tag{5}$$

$$\text{SSB} = 3\begin{bmatrix} 3-7/3 \\ 13.25/3-9.5/3 \end{bmatrix}^2 + 3\begin{bmatrix} 3-11/3 \\ 13.25/3-17/3 \end{bmatrix}^2 = \begin{bmatrix} 8/3 \\ 75/8 \end{bmatrix} \tag{6}$$

$$\begin{aligned} \text{TSS} = &\begin{bmatrix} 1-3 \\ 1.5-13.25/3 \end{bmatrix}^2 + \begin{bmatrix} 2-3 \\ 3-13.25/3 \end{bmatrix}^2 + \begin{bmatrix} 4-3 \\ 5-13.25/3 \end{bmatrix}^2 \\ &+ \begin{bmatrix} 5-3 \\ 7-13.25/3 \end{bmatrix}^2 + \begin{bmatrix} 3.5-3 \\ 4.5-13.25/3 \end{bmatrix}^2 + \begin{bmatrix} 2.5-3 \\ 5.5-13.25/3 \end{bmatrix}^2 = \begin{bmatrix} 21/2 \\ 449/24 \end{bmatrix} \end{aligned} \tag{7}$$

(2) Suppose they are divided into three clusters. P1, P2 belongs to cluster 1, P3, P4 belongs to cluster 2, P5 and P6 belongs to cluster 3. Please calculate the SSE, SSB and TSS.

Answer:

The P1, P2 belong to cluster 1, we therefore have corresponding $c_1 : (1.5, 2.25)$.

The P3, P4 belong to cluster 2, we therefore have corresponding $c_2 : (4.5, 6)$.

The P5, P6 belong to cluster 3, we therefore have corresponding $c_3 : (3, 5)$.

We can calculate c as

$$c : (1/3 * (1.5 + 4.5 + 3), 1/3 * (2.25 + 6 + 5)) = (3, 13.25/3) \tag{8}$$

$$\begin{aligned} \text{SSE} = &\begin{bmatrix} 1-1.5 \\ 1.5-2.25 \end{bmatrix}^2 + \begin{bmatrix} 2-1.5 \\ 3-2.25 \end{bmatrix}^2 \\ &+ \begin{bmatrix} 4-4.5 \\ 5-6 \end{bmatrix}^2 + \begin{bmatrix} 5-4.5 \\ 7-6 \end{bmatrix}^2 \\ &+ \begin{bmatrix} 3.5-3 \\ 4.5-5 \end{bmatrix}^2 + \begin{bmatrix} 2.5-3 \\ 5.5-5 \end{bmatrix}^2 \\ &= \begin{bmatrix} 3/2 \\ 29/8 \end{bmatrix} \end{aligned} \tag{9}$$

$$\text{SSB} = 2\begin{bmatrix} 3-1.5 \\ 13.25/3-2.25 \end{bmatrix}^2 + 2\begin{bmatrix} 3-4.5 \\ 13.25/3-6 \end{bmatrix}^2 + 2\begin{bmatrix} 3-3 \\ 13.25/3-5 \end{bmatrix}^2 = \begin{bmatrix} 9 \\ 181/12 \end{bmatrix} \tag{10}$$

$$\begin{aligned} \text{TSS} = &\begin{bmatrix} 1-3 \\ 1.5-13.25/3 \end{bmatrix}^2 + \begin{bmatrix} 2-3 \\ 3-13.25/3 \end{bmatrix}^2 + \begin{bmatrix} 4-3 \\ 5-13.25/3 \end{bmatrix}^2 \\ &+ \begin{bmatrix} 5-3 \\ 7-13.25/3 \end{bmatrix}^2 + \begin{bmatrix} 3.5-3 \\ 4.5-13.25/3 \end{bmatrix}^2 + \begin{bmatrix} 2.5-3 \\ 5.5-13.25/3 \end{bmatrix}^2 = \begin{bmatrix} 21/2 \\ 449/24 \end{bmatrix} \end{aligned} \tag{11}$$

(3) Show the relationship among SSE, SSB and TSS.

$$\text{SSE} + \text{SSB} = \text{TSS} \tag{12}$$

Equations (5) and (6) generates

$$\text{SSE} + \text{SSB} = \begin{bmatrix} 47/6 \\ 28/3 \end{bmatrix} + \begin{bmatrix} 8/3 \\ 75/8 \end{bmatrix} = \text{TSS} = \begin{bmatrix} 21/2 \\ 449/24 \end{bmatrix} \tag{13}$$

which is equal to equation (7).

Similarly, equations (9) and (10) also generates

$$\text{SSE} + \text{SSB} = \begin{bmatrix} 3/2 \\ 29/8 \end{bmatrix} + \begin{bmatrix} 9 \\ 181/12 \end{bmatrix} = \text{TSS} = \begin{bmatrix} 21/2 \\ 449/24 \end{bmatrix} \tag{14}$$

III. Hierarchical Clustering (25%)

Given the following points, answer the following questions.

$P1 : (1, 1.5), P2 : (2, 3), P3 : (4, 5), P4 : (5, 7), P5 : (3.5, 4.5), P6 : (2.5, 5.5)$

(1) Perform hierarchical clustering using complete linkage with agglomerative algorithm.

1.1) The point $P5 : (3.5, 4.5)$ is clustered into the point $P3 : (4, 5)$, namely $(P3 : (4, 5), P5 : (3.5, 4.5))$.

1.2) The point $P6 : (2.5, 5.5)$ is clustered into $(P3 : (4, 5), P5 : (3.5, 4.5))$,namely $(P3 : (4, 5), P5 : (3.5, 4.5), P6 : (2.5, 5.5))$.

1.3) The point $P2 : (2, 3)$ is clustered into $P1 : (1, 1.5)$, namely $(P1 : (1, 1.5), P2 : (2, 3))$.

1.4) The point $P4 : (5, 7)$ is clustered into $(P3 : (4, 5), P5 : (3.5, 4.5), P6 : (2.5, 5.5))$, namely $(P3 : (4, 5), P4 : (5, 7), P5 : (3.5, 4.5), P6 : (2.5, 5.5))$.

As a result, the first clusters contain points $P1 : (1, 1.5), P2 : (2, 3)$. The second clusters contain points $P3 : (4, 5), P4 : (5, 7), P5 : (3.5, 4.5), P6 : (2.5, 5.5)$.

(2) Perform hierarchical clustering using single linkage with agglomerative algorithm.

2.1) The point $P5 : (3.5, 4.5)$ is clustered into the point $P3 : (4, 5)$, namely $(P3 : (4, 5), P5 : (3.5, 4.5))$.

2.2) The point $P6 : (2.5, 5.5)$ is clustered into $(P3 : (4, 5), P5 : (3.5, 4.5))$,namely $(P3 : (4, 5), P5 : (3.5, 4.5), P6 : (2.5, 5.5))$.

2.3) The point $P2 : (2, 3)$ is clustered into $P1 : (1, 1.5)$, namely $(P1 : (1, 1.5), P2 : (2, 3))$.

2.4) The points $(P3 : (4, 5), P5 : (3.5, 4.5), P6 : (2.5, 5.5))$ are clustered into $(P1 : (1, 1.5), P2 : (2, 3))$, namely $(P1 : (1, 1.5), P2 : (2, 3), P3 : (4, 5), P5 : (3.5, 4.5), P6 : (2.5, 5.5))$.

As a result, the first clusters contain points $P1 : (1, 1.5), P2 : (2, 3), P3 : (4, 5), P5 : (3.5, 4.5), P6 : (2.5, 5.5)$. The second clusters contain points $P4 : (5, 7)$.

Code is as follows: https://github.com/wangwanlitype1/ENGG5103

IV. SOM clustering (25%)

X = 1, 1.5, 2.5, 4.8, 2.8, 3.2, 3.8, 4.7, 3.3, 3.1

Y = 2, 2.3, 2.8, 4.5, 5.1, 6.3, 6.7, 6.8, 5.7, 3.9

Suppose that a = 0.2, a(neighbor) = 0.3, the size of neighbor is 5, perform SOM clustering and show the final results.

Answer:

Clustering simulation results for the first 5 iterations are as follows:

The first iteration is as follows:

Cluster1: correponding sample index[1, 2, 5, 6, 7, 8, 9]

Cluster2: correponding sample index[3, 4, 10]

The second iteration is as follows:

Cluster1: correponding sample index[1, 2, 5, 6, 7, 8, 9]

Cluster2: correponding sample index[3, 4, 10]

The third iteration is as follows:

Cluster1: correponding sample index[1, 5, 6, 7, 9]

Cluster2: correponding sample index[2, 8]

Cluster3: correponding sample index[3, 4, 10]

The forth iteration is as follows:

Cluster1: correponding sample index[1, 5, 6, 7, 9]

Cluster2: correponding sample index[2]

Cluster3: correponding sample index[3, 4, 10]

Cluster4: correponding sample index[8]

The fifth iteration is as follows:

Cluster1: correponding sample index[1, 5, 6, 7, 9]

Cluster2: correponding sample index[2]

Cluster3: correponding sample index[3, 4, 10]

Cluster4: correponding sample index[8]

It is necessary to notice that

a): I did not adopt fixed parameters like $a = 0.2$, $a(neighbor) = 0.3$ and the size of neighbor 5 since fixed parameters are not enough for us to do clustering. For example, it is better to set parameter of learning rate as a nonlinear function of the discrete time t and topology distance l. In general, the learning rate of the weight $e(t, l)$ is set as

$$e(t, l) = \eta(t)e^{-l}. \tag{15}$$

In the following code, the learning rate $e(t,l)$ is set as

$$e(t,l)=\frac{e^{-l}}{t+2}. \tag{16}$$

This means that learning rate decreases along with increasing discrete time t and topology distance l.

b): The SOM clustering result may be not fixed and changed for each run since the weight of the SOM is initialized randomly.

c): The output of SOM clustering includes 2×2 neurons.

Code is as follows: https://github.com/wangwanlitype1/ENGG5103