CSE881 HW8

Nan Cao, A52871775
Nov 12th, 2016

Problem 1

(a)

	Cluster assignment of data points								Centroid Locations		
Iter	0.1	0.2	0.8	0.9	1.0	1.3	1.8	1.9	A	B	C
0	-	-	-	-	-	-	-	-	0	0.3	2
1	1	2	2	2	2	3	3	3	0.1	0.725	1.667
2	1	1	2	2	2	3	3	3	0.15	0.9	1.667
3	1	1	2	2	2	3	3	3	0.15	0.9	1.667

Table 1-a

(b)

$$SSE = 0.2317$$

(c)

If we put 1.0 in the first group:

	Cluster assignment of data points								Centroid Locations	
Iter	0.1	0.2	0.8	0.9	1.0	1.3	1.8	1.9	A	B
0	-	-	-	-	-	-	-	-	0	2
1	1	1	1	1	1	2	2	2	0.6	1.667
2	1	1	1	1	1	2	2	2	0.6	1.667

Table 1-c-1

$$SSE_1 = 0.7, \quad SSE_2 = 0.2067 \; Choose \; A$$

	Cluster assignment of data points								Centroid Locations		
Iter	0.1	0.2	0.8	0.9	1.0	1.3	1.8	1.9	A	A_2	B
0	-	-	-	-	-	-	-	-	0	-	2
1	1	1	1	1	1	2	2	2	0.6	-	1.667
2	1	1	1	1	1	2	2	2	0.6	-	1.667
2	-	-	-	-	-	3	3	3	0.1	1	-
3	1	1	2	2	2	3	3	3	0.15	0.9	1.667
4	1	1	2	2	2	3	3	3	0.15	0.9	1.667

Table 1-c-2

$$SSE = 0.2317$$

If we put 1.0 in the first group:

	Cluster assignment of data points								Centroid Locations	
Iter	0.1	0.2	0.8	0.9	1.0	1.3	1.8	1.9	A	B
0	-	-	-	-	-	-	-	-	0	2
1	1	1	1	1	2	2	2	2	0.6	1.667
2	1	1	1	1	2	2	2	2	0.6	1.667

Table 1-c-3

$$SSE_1 = 0.5, \quad SSE_2 = 0.54\ Choose\ B$$

	Cluster assignment of data points								Centroid Locations		
IteR	0.1	0.2	0.8	0.9	1.0	1.3	1.8	1.9	A	B	B_2
0	-	-	-	-	-	-	-	-	0	1.0	-
1	1	1	1	1	2	2	2	2	0.5	1.5	-
2	1	1	1	1	2	2	2	2	0.5	1.5	-
-	1	1	1	1	-	-	-	-	-	1	1.9
3	1	1	1	1	2	2	3	3	0.5	1.15	1.85
4	1	1	1	1	2	2	3	3	0.5	1.15	1.85

Table 1-c-4

$$SSE = 0.55$$

(d)

The firt method is better, the SSE in (a) is 0.2317, and the SSE in (c) is 0.2317 and 0.55. We need to do more job when there're data points on the border. And if we group the border points in the wrong group, it may result in a higher SSE.

Problem 2

$$
\begin{aligned}
SSE =& \sum_{i\in C_1+C_2}(x_i-\mu)^2 \\
SSE_2 =& \sum_{i\in C_1}(x_i-\mu_1)^2+\sum_{i\in C_2}(x_i-\mu_2)^2 \\
SSE_2 - SSE =& (\sum_{i\in C_1}(x_i-\mu_1)^2+\sum_{i\in C_2}(x_i-\mu_2)^2) - \sum_{i\in C_1+C_2}(x_i-\mu)^2 \\
=& (\sum_{i\in C_1}(x_i^2-2x_i\mu_1+\mu_1^2)+\sum_{i\in C_2}(x_i^2-2x_i\mu_2+\mu_2^2)) - \sum_{i\in C_1+C_2}(x_i^2-2x_i\mu+\mu^2) \\
=& (\sum_{i\in C_1}x_i^2-\sum_{i\in C_1}2x_i\mu_1+N_1\mu_1^2)+(\sum_{i\in C_2}x_i^2-\sum_{i\in C_2}2x_i\mu_2+N_2\mu_2^2) \\
& -(\sum_{i\in C_1+C_2}x_i^2-\sum_{i\in C_1+C_2}2x_i\mu+N\mu^2) \\
=& (\sum_{i\in C_1+C_2}x_i^2-\sum_{i\in C_1}x_i^2-\sum_{i\in C_2}x_i^2)+(N\mu^2-(N1\mu_1^2+N_2\mu_2^2)) \\
=& N\mu^2-(N_1\mu_1^2+N_2\mu_2^2) \\
=& (N_1+N_2)(\frac{N_1\mu_1+N_2\mu_2}{N_1+N_2})^2-(N1\mu_1^2+N_2\mu_2^2) \\
=& \frac{(N_1\mu_1+N_2\mu_2)^2}{N_1+N_2}-(N1\mu_1^2+N_2\mu_2^2) \\
=& \frac{(N_1^2\mu_1^2-2N_1N_2\mu_1\mu_2+N_2^2\mu_2^2)-((N_1(N_1+N_2)\mu_1^2+N_2(N_1+N_2)\mu_2^2))}{N1+N_2} \\
=& -\frac{N_1N_2(\mu_1-\mu_2)^2}{N}\le 0
\end{aligned}
$$

SSE is non-increasing when the data is split into 2 clusters.

Problem 3

(b)

I prefer correlation, because prices of different stocks in same category may under different scales but change in the similar way. If we use Euclidean distance, stocks in same category may have a greater distance than that of stocks from the different categories.

(d)

$$Confusion\ Matrix = \begin{bmatrix} 22 & 0 & 12 & 0 & 0 & 0 & 4 & 0 & 17 & 1 \\ 17 & 0 & 7 & 0 & 1 & 0 & 0 & 0 & 21 & 0 \\ 4 & 0 & 26 & 1 & 2 & 0 & 4 & 0 & 28 & 0 \\ 1 & 0 & 36 & 0 & 2 & 1 & 3 & 0 & 14 & 0 \\ 2 & 0 & 10 & 0 & 0 & 0 & 0 & 0 & 16 & 0 \\ 6 & 1 & 28 & 0 & 0 & 0 & 5 & 1 & 34 & 2 \\ 13 & 0 & 7 & 0 & 0 & 0 & 1 & 0 & 8 & 0 \\ 10 & 0 & 9 & 0 & 0 & 0 & 0 & 0 & 13 & 0 \\ 13 & 0 & 7 & 0 & 0 & 0 & 2 & 0 & 15 & 0 \\ 0 & 0 & 3 & 0 & 0 & 0 & 0 & 0 & 2 & 0 \end{bmatrix}$$

$$Correct\ Rate = \frac{Trace}{Sum} = 0.1481$$

(f)

$$Confusion\ Matrix = \begin{bmatrix} 5 & 4 & 11 & 3 & 1 & 3 & 3 & 10 & 12 & 4 \\ 7 & 10 & 5 & 3 & 2 & 1 & 8 & 5 & 2 & 3 \\ 12 & 13 & 6 & 14 & 4 & 3 & 7 & 5 & 1 & 0 \\ 6 & 6 & 12 & 13 & 6 & 3 & 3 & 4 & 3 & 1 \\ 5 & 10 & 0 & 0 & 0 & 1 & 6 & 2 & 4 & 0 \\ 26 & 10 & 16 & 0 & 1 & 2 & 8 & 3 & 9 & 2 \\ 4 & 3 & 3 & 5 & 1 & 2 & 0 & 2 & 5 & 4 \\ 4 & 8 & 6 & 3 & 0 & 3 & 1 & 5 & 2 & 0 \\ 4 & 1 & 3 & 7 & 0 & 1 & 3 & 2 & 8 & 8 \\ 1 & 0 & 2 & 0 & 0 & 0 & 2 & 0 & 0 & 0 \end{bmatrix}$$

$$Correct\ Rate = \frac{Trace}{Sum} = 0.1143$$

Euclidean distance is better for the data set.