DS 5230 Unsupervised Machine Learning and Data Mining
Homework 3 By Yanchi Li

Problem 1 a. First we plot those point in a coordinate plane to see how those points distributed.

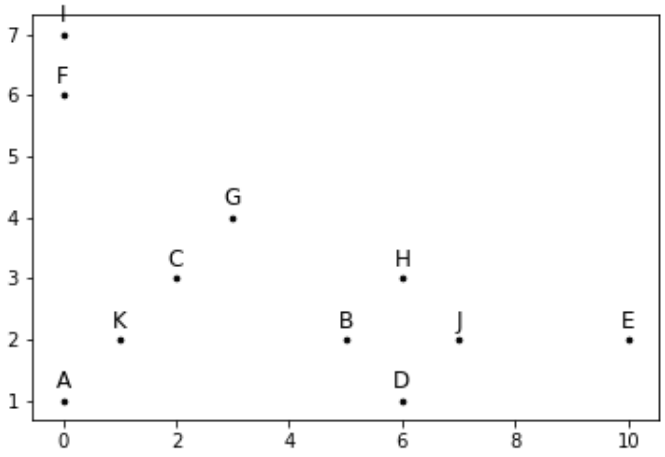

Figure 1: Raw Data

Then we are going to split them into clusters using DBSCAN algorithm with $\epsilon = \sqrt{2}$ and $Min\_Pts = 3$.

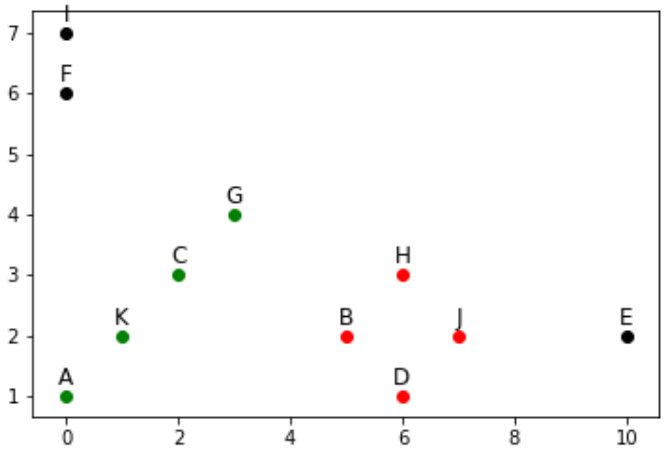

Figure 2: Clustering Result

Here we can see there are two clusters in the plot, point A, K, C, G are assigned to the green cluster and point B, H, D, J are assigned to the red cluster and there are also three points (I, F, E) considered as noise points.

b. As clearly shown in Figure 2, point A, K, C, G are density connected since the diagonal distance of a 1-by-1 cell is exactly $\sqrt{2}$. Also, there are four points nearby, where four is larger than the given Min_Pts so as a result they are assigned to the same cluster. Similarly, point B, H, D, J are also density connected and they are in the same cluster as well.
Though the number of their neighbors does not add up to three point F and I are also density connected because the distance between them is just 1.

c. As shown in Figure 2, point F, I and E are considered noise points (which are colored in black), because they does not form a neighbor with number of points larger than the given Min_Pts.

Problem 2 a.

$$\begin{aligned} m_1 &= (\frac{5+8+7}{3}, \frac{6+7+3}{3}) \\ &= (6.6, 5.3) \end{aligned}$$

$$\begin{aligned} m_2 &= (\frac{6+4+9+3+8}{5}, \frac{5+5+2+5+4}{5}) \\ &= (6, 4.2) \end{aligned}$$

b.

$$\begin{aligned} m &= (\frac{6.6 \times 3 + 6 \times 5}{8}, \frac{5.3 \times 3 + 4.2 \times 5}{8}) \\ &= (6.25, 4.625) \end{aligned}$$

c.

$$\begin{aligned} S_1 &= \sum_n I[z_n = 1](x_n - \mu_1)(x_n - \mu_1)^T \\ &= \begin{bmatrix} -1.67 \\ 0.67 \end{bmatrix} \begin{bmatrix} -1.67 & 0.67 \end{bmatrix} + \begin{bmatrix} 1.33 \\ 1.67 \end{bmatrix} \begin{bmatrix} 1.33 & 1.67 \end{bmatrix} \\ &\quad + \begin{bmatrix} 0.33 \\ -2.33 \end{bmatrix} \begin{bmatrix} 0.33 & -2.33 \end{bmatrix} \\ &= \begin{bmatrix} 4.67 & 0.67 \\ 0.67 & 8.67 \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
S_2 &= \sum_n I[z_n = 2](x_n - \mu_2)(x_n - \mu_2)^T \\
&= \begin{bmatrix} 0 \\ 0.8 \end{bmatrix} \begin{bmatrix} 0 & 0.8 \end{bmatrix} + \begin{bmatrix} -2 \\ 0.8 \end{bmatrix} \begin{bmatrix} -2 & 0.8 \end{bmatrix} + \begin{bmatrix} 3 \\ -2.2 \end{bmatrix} \begin{bmatrix} 3 & -2.2 \end{bmatrix} \\
&\quad + \begin{bmatrix} -3 \\ 0.8 \end{bmatrix} \begin{bmatrix} -3 & 0.8 \end{bmatrix} + \begin{bmatrix} 2 \\ -0.2 \end{bmatrix} \begin{bmatrix} 2 & -0.2 \end{bmatrix} \\
&= \begin{bmatrix} 26 & -11 \\ -11 & 6.8 \end{bmatrix}
\end{aligned}
$$

d.

$$
\begin{aligned}
S^W &= S_1 + S_2 \\
&= \begin{bmatrix} 30.67 & -10.67 \\ -10.67 & 15.47 \end{bmatrix}
\end{aligned}
$$

e.

$$
\begin{aligned}
S^B &= \sum_k N_k(\mu_k - \mu)(\mu_k - \mu)^T \\
&= 3 \times \begin{bmatrix} 6.6 - 6.25 \\ 5.3 - 4.625 \end{bmatrix} \begin{bmatrix} 6.6 - 6.25 & 5.3 - 4.625 \end{bmatrix} \\
&\quad + 5 \times \begin{bmatrix} 6 - 6.25 \\ 4.2 - 4.625 \end{bmatrix} \begin{bmatrix} 6 - 6.25 & 4.2 - 4.625 \end{bmatrix} \\
&= \begin{bmatrix} 0.83 & 1.42 \\ 1.42 & 2.41 \end{bmatrix}
\end{aligned}
$$

f.

$$
\begin{aligned}
SC &= \frac{tr(S_B)}{tr(S_W)} \\
&= \frac{0.83 + 2.41}{30.67 + 15.47} \\
&= 0.07
\end{aligned}
$$

Problem 3 To show the difference between the three distances, I tried to simplify the question into one merge pass. So here I just present one step of the clustering process.

I have drawn three clusters A, B and C as shown in Figure 3. I use different colors to indicate different linkage where red representing the maximum distance (complete linkage), green representing minimum distance (single linkage) and blue (may not be clear enough from the picture) for the average distance.

For minimum distance, it's definitely that we are going to merge A and B since the distance here is really small. For maximum distance, B and C is merged in this step. And at last we are going to merge A and B when we apply the average distance (just the distance between the centroids).

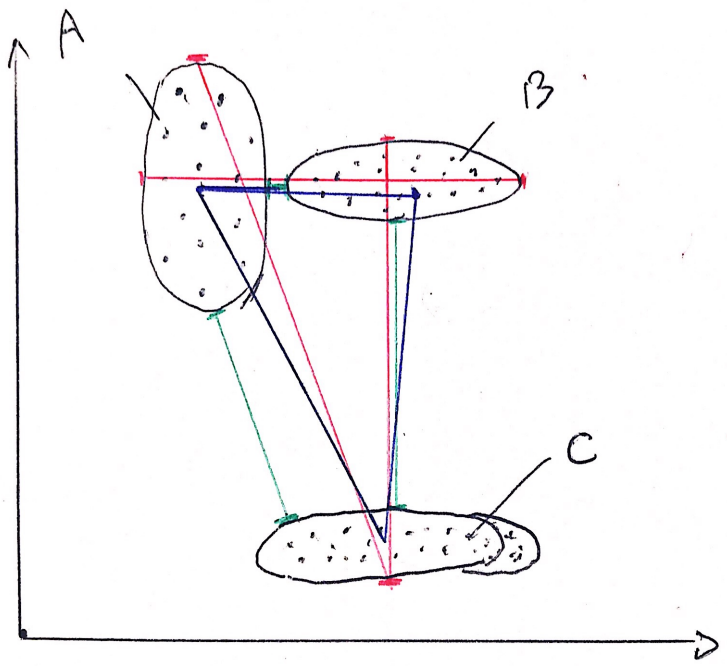

Figure 3: Agglomerative Clustering Example 2

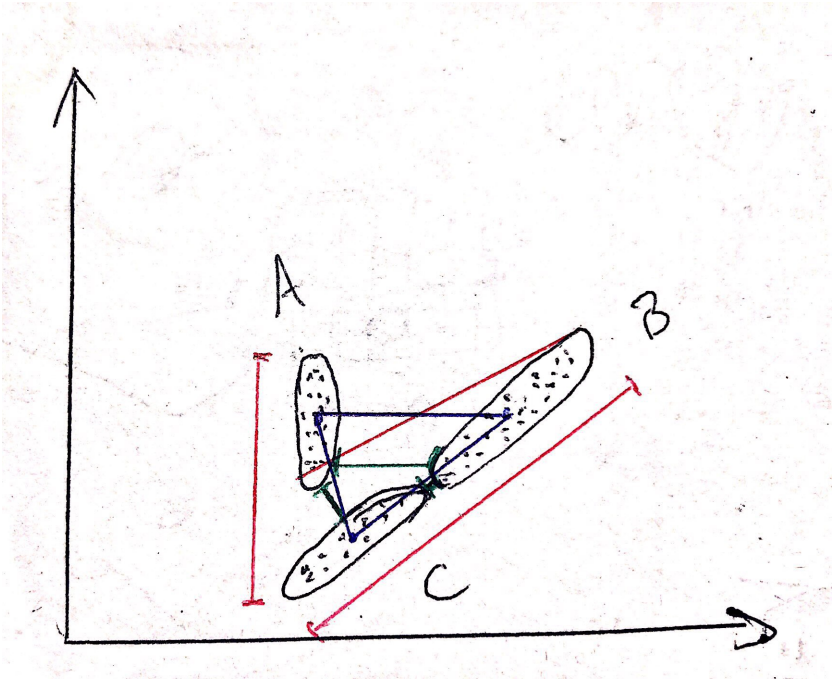

Figure 4: Aggplomerative Clustering Example 2

As the example 1 does not show the difference between minimum and average distance, I came up with a second example which is shown in Figure 4.

When we apply average distance, we could merge A and C since the blue line (the triangle between three centroids) between A and C is the smallest. And clearly, B and C are merged if we use maximum or minimum distance.