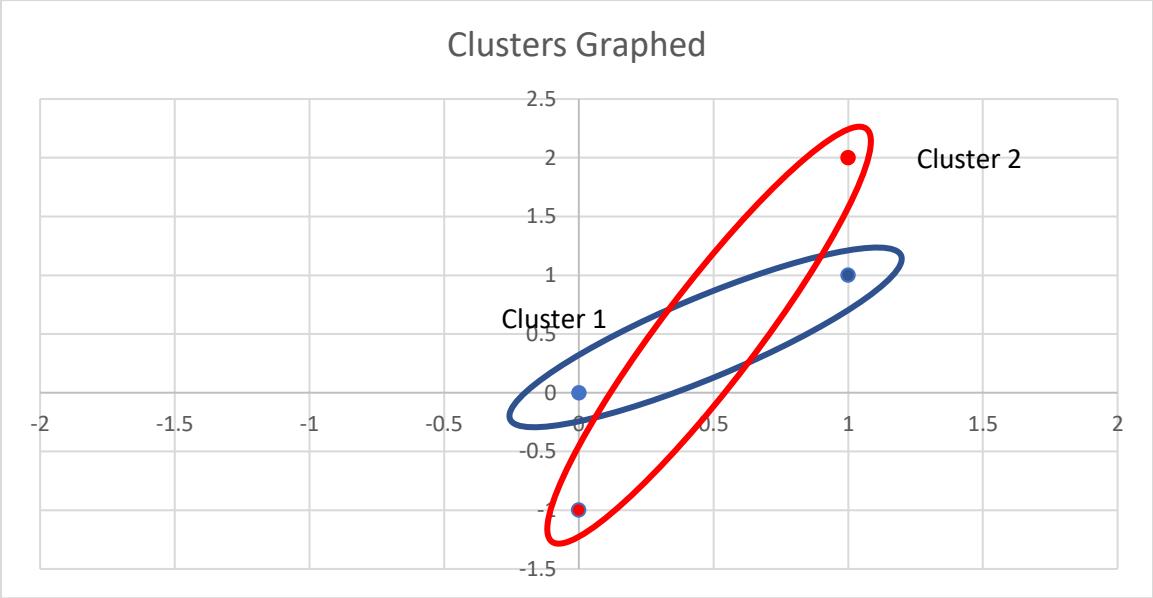
1. Theory Questions

1.

$$C_1 = \begin{pmatrix} (1,2) \\ (0,-1) \end{pmatrix}, C_2 = \begin{pmatrix} (0,0) \\ (1,1) \end{pmatrix}$$



a. The weighted average intra-cluster distance if you are using Euclidean distance?

$$\text{Euclidean } Distance: distance(A,B) = \sqrt{\sum_{i=1}^{D}(A_i - B_i)^2}$$

$$Average\ Pairwise\ Intracluster\ Distance: G_i = \frac{\sum_{x,y \in C_i} d(x,y)}{(2 * |C_i|)}$$

$$Weighted\ Average\ IntraCluster\ Distance: W_j = \frac{\sum_{i=1}^{j} |C_i| * G_i}{N}$$

Cluster 1:

$$G_1 = \frac{\sqrt{(1-0)^2 + (2--1)^2}}{(2*2)} = \frac{\sqrt{1+9}}{4} = .79056$$

Cluster 2:

$$G_2 = \frac{\sqrt{(0-1)^2 + (0-1)^2}}{(2*2)} = \frac{\sqrt{2}}{4} = .3535533$$

Weighted Average Intra-cluster distance

$$W = \frac{|C_1| * G_1 + |C_2| * G_2}{N} = \frac{2 * .79056 + 2 * .3535533}{4} = .57206$$

b. The single link similarity between the clusters if we're using cosine similarity as our similarity function?

$$similarity(A,B) = \cos(\emptyset) = \frac{A \cdot B}{\|A\| * \|B\|} = \frac{\sum_{i=1}^{D} A_i B_i}{\sqrt{\sum_{i=1}^{D} A_i^2} * \sqrt{\sum_{i=1}^{D} B_i^2}}$$

Single link means comparing the most similar: (1,2) and (1,1)

$$similarity((1,1),(1,2)) = \frac{(1*1)+(1*2)}{\sqrt{1^2+1^2} * \sqrt{1^2+2^2}} = \frac{3}{\sqrt{2}*\sqrt{5}} = .94868$$

c. The complete link similarity between the clusters if we're using cosine similarity as our similarity function?

Complete link means comparing the furthest apart: (1,1) and (0,-1)

$$similarity((1,1),(0,-1)) = \frac{(1*0)+(1*-1)}{\sqrt{1^2+1^2} * \sqrt{0^2+-1^2}} = \frac{-1}{\sqrt{2}*\sqrt{1}} = -.707106$$

d. The average link similarity between the clusters if we're using cosine similarity as our similarity function?

Average Link means comparing the average of the two points in the cluster

$$AvgC_1 = \left(\frac{1+0}{2}, \frac{2+-1}{2}\right) = (.5,.5)\ , AvgC_2 = \left(\frac{0+1}{2}, \frac{0+1}{2}\right) = (.5,.5)$$

$$similarity((.5,.5),(.5,.5)) = \frac{(.5*.5)+(.5*.5)}{\sqrt{.5^2+.5^2} * \sqrt{.5^2+.5^2}} = \frac{.5}{\sqrt{2}/2*\sqrt{2}/2} = 1$$

2. Given an average intracluster distance for clustering level j, Wj, what is the fourth derivative at j, namely W_j''''

We can break the W_j'' given in the notes into parts, and use the same pattern to get the second derivative of those parts.

$$W_j'' = \frac{(W_{j+2} - 2\,W_j + W_{j-2})}{4} = \frac{W_{j+2}}{4} - \frac{2W_j}{4} + \frac{W_{j-2}}{4}$$

$$W_j'''' = \frac{1}{4}\left(\frac{(W_{j+4} - 2\,W_{j+2} + W_j)}{4}\right) - \frac{2}{4}\left(\frac{(W_{j+2} - 2\,W_j + W_{j-2})}{4}\right) + \frac{1}{4}\left(\frac{(W_j - 2\,W_{j-2} + W_{j-4})}{4}\right) = \frac{1}{16}(W_{j+4} - 4W_{j+2} - 2W_j - 4W_{j-2} + W_{j-4})$$

3. What is the weighted average purity of the clusters created by the clustering algorithm?

$Clustering\ Algorithm: CA_1 = (1,2,3,4)\ , CA_2 = (5,6,7,8)$

$Clustering\ by\ Hand: CH_1 = (3,4)\ , CH_2 = (1,2,5,6,7,8)$

$Purity(C_i) = \frac{1}{|C_i|} * max_j\ (N_{ij})$

$Average\ Purity = \frac{1}{N}\sum_{i=1}^{k}|C_i| * Purity(C_i)$

First, calculate all the possible Nij's

$N_{11} = 2$, $N_{12} = 2$

$N_{21} = 2$, $N_{22} = 4$

Now Purity of Each Cluster

$$Purity(C_1) = \frac{1}{|C_1|} * max_j\left(N_{1j}\right) = \frac{1}{4} * 2 = \frac{1}{2}$$

$$Purity(C_2) = \frac{1}{|C_2|} * max_j\left(N_{2j}\right) = \frac{1}{4} * 4 = 1$$

Now Average Purity

$$Average\ Purity = \frac{1}{N}\sum_{i=1}^{k}|C_i| * Purity(C_i) = \frac{1}{8}\left(\left(4 * \frac{1}{2}\right) + (4 * 1)\right) = \frac{6}{8} = .75$$