CS 383 Assignment 2

Kevin Tayah (kst46@drexel.edu)

February 13, 2021

1. Theory Questions

1. Given these two clusters:

$$C_1 = (1,2),(0,-1), C_2 = (0,0),(1,1)$$

(a) The weighted average intra-cluster distance using the Euclidean distance can be calculated as such:
Based on the function, $W_j = \frac{\sum_{i=1}^{j} |C_i| G_i}{N}$ where $j = 2$ due to there being 2 clusters, $G_i = \frac{\sum_{x,y \in C_i} d(x,y)}{2|C_i|}$ for a cluster i and using a particular distance function, in this case Euclidean, $d(A,B) = \sqrt{\sum_{i=1}^{D} (A_i - B_i)^2}$. We can do the following calculations:

$$G_1 = \frac{\sqrt{1^2 + (-2)^2}}{4} = \frac{\sqrt{5}}{4}$$
$$G_2 = \frac{\sqrt{(-1)^2 + (-1)^2}}{4} = \frac{\sqrt{2}}{4}$$
$$W_2 = \frac{(\frac{\sqrt{5}}{4} \cdot 2) + (\frac{\sqrt{2}}{4} \cdot 2)}{4} = \frac{\sqrt{5} + \sqrt{2}}{8} = 0.45628519248$$

Our weighted average intra-cluster distance using Euclidean distance is equal to 0.45628519248.

(b) The single link similarity between clusters can be calculated based upon this algorithm:
$sim(C_i, C_j) = max_{x \in C_i, y \in C_i}(sim(x,y))$ where sim is our similarity function of choice; in this case we are using the cosine similarity, $\frac{C_i \cdot C_j}{\|C_i\| \|C_j\|}$. Putting this all together, if we were to define C_{lk} where l is the cluster number and kth is the element in that cluster, we can calculate the single link similarity as so:

$$sim(C_1, C_2) = max(\frac{C_{11} \cdot C_{21}}{\|C_{11}\| \|C_{21}\|}, \frac{C_{12} \cdot C_{22}}{\|C_{12}\| \|C_{22}\|})$$
$$sim(C_1, C_2) = max(0, \frac{-1}{\sqrt{2}})$$
$$sim(C_1, C_2) = 0$$

Our single link similarity between clusters is 0.

(c) The complete link similarity between clusters can be calculated as such: $sim(C_i, C_j) = min_{x \in C_i, y \in C_i}(sim(x, y))$ and since we already know $sim(C_{11}, C_{21})$ and $sim(C_{12}, C_{22})$ since we calculated it above. We know the answer comes out to be:

$$sim(C_1, C_2) = min(0, \frac{-1}{\sqrt{2}})$$
$$sim(C_1, C_2) = \frac{-1}{\sqrt{2}}$$

Our complete link similarity between clusters is $\frac{-1}{\sqrt{2}})$
.

(d) The average link similarity between the clusters can be calculated as such: $sim(C_i, C_j) = \frac{1}{|C_i||C_j|} \sum_{x \in C_i} \sum_{y \in C_j} sim(x, y)$. If we were to apply this to our clusters, it can be evaluated as such:

$$sim(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{x \in C_1} \sum_{y \in C_2}$$
$$sim(C_1, C_j) = \frac{1}{2 \cdot 2}(sim(C_{11}, C_{21}), sim(C_{12}, C_{21}), sim(C_{11}, C_{22}), sim(C_{12}, C_{22}))$$
$$sim(C_1, C_j) = 0.25(\frac{3}{\sqrt{10}} - \frac{1}{\sqrt{2}})$$
$$sim(C_1, C_j) = 0.06039$$

2. Fourth derivative at j of W_j, given an average intracluster distance is calculated as such:

$$W_j = \frac{\sum_{i=1}^{j} |C_i| G_i}{N}$$
$$W_j^{'} = \frac{W_{j+1} - W_{j-1}}{2}$$
$$W_j^{''} = \frac{W_{j+2} - 2W_j + W_{j-2}}{4}$$
$$W_j^{'''} = \frac{W_{j+2}^{'} - 2W_j^{'} + W_{j-2}^{'}}{4}$$
$$W_j^{'''} = \frac{\frac{W_{j+3} - W_{j+1}}{2} - W_{j+1} - W_{j-1} + \frac{W_{j-1} - W_{j-3}}{2}}{4}$$
$$W_j^{'''} = \frac{W_{j+3} - W_{j+1} - 4(W_{j+1} - W_{j-1}) + W_{j-1} - W_{j-3}}{8}$$
$$W_j^{''''} = \frac{W_{j+3}^{'} - W_{j+1}^{'} - 4(W_{j+1}^{'} - W_{j-1}^{'}) + W_{j-1}^{'} - W_{j-3}^{'}}{8}$$
$$W_j^{''''} = \frac{\frac{W_{j+4} - W_{j+2}}{2} - \frac{W_{j+2} - W_j}{2} - 4(\frac{W_{j+2} - W_j}{2} - \frac{W_j - W_{j-2}}{2}) + \frac{W_j - W_{j-2}}{2} - \frac{W_{j-2} - W_{j-4}}{2}}{8}$$
$$W_j^{''''} = \frac{W_{j+4} - W_{j+2} - W_{j+2} + W_j + W_j - W_{j-2} - W_{j-2} + W_{j-4}}{16} - \frac{W_{j+2} - W_j - W_j + W_{j-2}}{4}$$

3. Given a clustering of $C_1 = \{1, 2, 3, 4\}, C_2 = \{5, 6, 7, 8\}$ and the hand labeled clustering of $C_1 = \{3, 4\}, C_2 = \{1, 2, 5, 6, 7, 8\}$. The weighted average purity of the clusters created by the clustering algorithm can be calculated using the following equation:
   Average Purity $= \frac{1}{N} \sum_{i=1}^{k} |C_i| Purity(C_i)$
   and Purity $= \frac{1}{|C_i|} \max_j N_{ij}$

   Cluster 1: $N_{11} = \frac{1}{4} \max(2, 2) = \frac{1}{2}$

   Cluster 2: $N_{22} = \frac{1}{4} \max(4, 2) = 1$

   Average Purity $= \frac{1}{8}(2 * \frac{1}{2} + 6 * 1) = \frac{7}{8} = 0.875 = 87.5\%$