Vrund Patel

CS383-001

1 Theory

1. (3pts each = 12pts) Given two clusters:

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

   what is:

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

$$\boldsymbol{G_i} = \frac{\sum_{x,y \in C_i} \boldsymbol{d(x,y)}}{(\boldsymbol{2|C_i})}$$

$$\boldsymbol{G_1} = \frac{\sqrt{(1-0)^2 + (2-(-1))^2}}{(2*2)} = \frac{\sqrt{1+9}}{4} = \frac{\sqrt{\mathbf{10}}}{\mathbf{4}}$$

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

$$\boldsymbol{W_j} = \sum_{\boldsymbol{i=1}}^{\boldsymbol{j}} \frac{\boldsymbol{|C_i|}}{\boldsymbol{N}} \boldsymbol{G_i}$$

$$\boldsymbol{W_2} = \frac{2}{4} * \frac{\sqrt{10}}{4} + \frac{2}{4} * \frac{\sqrt{2}}{4} = \frac{2\sqrt{10}}{16} + \frac{2\sqrt{2}}{16} = \frac{2\sqrt{10} + 2\sqrt{2}}{16} \mathbf{.5721}$$

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

$$sim(C_i, C_j) = max_{x \in C_i, y \in C_j} sim(x, y)$$

$$\cos\theta_{C_{1,1}C_{2,1}} = \frac{A \cdot B}{||A||\,||B||} = \frac{1*0 + 2*0}{X} = \mathbf{0}$$

$$\cos\theta_{C_{1,1}C_{2,2}} = \frac{A \cdot B}{||A||\,||B||} = \frac{1*1 + 2*1}{\sqrt{1^2 + 2^2} \cdot \sqrt{1^2 + 1^2}} = \frac{3}{\sqrt{5} \cdot \sqrt{2}} = \frac{3}{\sqrt{10}} \cong \mathbf{0.949}$$

$$\cos\theta_{C_{1,2}C_{2,1}} = \frac{A \cdot B}{||A||\,||B||} = \frac{0*0 + -1*0}{X} = \mathbf{0}$$

$$\cos\theta_{C_{1,2}C_{2,2}} = \frac{A \cdot B}{||A||\,||B||} = \frac{0*1 + -1*1}{\sqrt{0^2+1^2} \cdot \sqrt{-1^2+1^2}} = \frac{-1}{\sqrt{1} \cdot \sqrt{2}} = \frac{-1}{\sqrt{2}} \cong \mathbf{-0.707}$$

$$\boldsymbol{sim(C_1, C_2)} = \frac{\mathbf{3}}{\sqrt{\mathbf{10}}} \cong \mathbf{0.949}$$

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

$$sim(C_i, C_j) = min_{x \in C_i, y \in C_j} sim(x, y)$$

$$\cos\theta_{C_{1,1}C_{2,1}} = \frac{A \cdot B}{||A||\,||B||} = \frac{1*0 + 2*0}{X} = \mathbf{0}$$

$$\cos\theta_{C_{1,1}C_{2,2}} = \frac{A \cdot B}{||A||\,||B||} = \frac{1*1 + 2*1}{\sqrt{1^2 + 2^2} \cdot \sqrt{1^2 + 1^2}} = \frac{3}{\sqrt{5} \cdot \sqrt{2}} = \frac{3}{\sqrt{10}} \cong \mathbf{0.949}$$

$$\cos\theta_{C_{1,2}C_{2,1}} = \frac{A \cdot B}{||A||\,||B||} = \frac{0*0 + -1*0}{X} = \mathbf{0}$$

$$\cos\theta_{C_{1,2}C_{2,2}} = \frac{A \cdot B}{||A||\,||B||} = \frac{0*1 + -1*1}{\sqrt{0^2 + 1^2} \cdot \sqrt{-1^2 + 1^2}} = \frac{-1}{\sqrt{1} \cdot \sqrt{2}} = \frac{-1}{\sqrt{2}} \cong \mathbf{-0.707}$$

$$\boldsymbol{sim(C_1, C_2) = \frac{-1}{\sqrt{2}} \cong -0.707}$$

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

$$sim(C_i, C_j) = \frac{1}{|C_i|\,|C_j|} \sum_{x \in C_i} \sum_{y \in C_j} sim(x, y)$$

$$\frac{1}{|C_i|\,|C_j|} * (\cos\theta_{C_{1,1}C_{2,1}} + \cos\theta_{C_{1,1}C_{2,2}} + \cos\theta_{C_{1,2}C_{2,1}} + \cos\theta_{C_{1,2}C_{2,2}}) =$$

$$\frac{1}{2 * 2} * (0 + 0.949 + 0 + (-0.707)) =$$

$$\frac{1}{4} * (0.242) = \mathbf{.0605}$$

2. (10pts) Given an average intracluster distance for clustering level j, W_j , what is the fourth derivative at j, namely W_j''''?

$$W'_j = \frac{(W_{j+1} - W_{j-1})}{2}$$

$$W''_j = \frac{(W'_{j+1} - W'_{j-1})}{2} = \frac{\left(\frac{(W_{j+2} - W_j)}{2} - \frac{(W_j - W_{j-2})}{2}\right)}{2} = \frac{(W_{j+2} - 2W_j + W_{j-2})}{4}$$

$$W'''_j = \frac{(W'_{j+2} - 2W'_j + W'_{j-2})}{4} = \frac{\left(\frac{W_{j+3} - W_{j+1}}{2}\right) - 2\left(\frac{W_{j+1} - W_{j-1}}{2}\right) + \left(\frac{W_{j-1} - W_{j-3}}{2}\right)}{4} = \frac{W_{j+3} - 3W_{j+1} + 3W_{j-1} - W_{j-3}}{8}$$

$$\boldsymbol{W''''_j} = \frac{W'_{j+3} - 3W'_{j+1} + 3W'_{j-1} - W'_{j-3}}{8} = \frac{\left(\frac{W_{j+4} - W_{j+2}}{2}\right) - 3\left(\frac{W_{j+2} - W_j}{2}\right) + 3\left(\frac{W_j - W_{j-2}}{2}\right) - \left(\frac{W_{j-2} - W_{j-4}}{2}\right)}{8} =$$

$$\boldsymbol{W''''_j} = \frac{W_{j+4} - 4W_{j+2} + 6W_j - 4W_{j-2} + W_{j-4}}{16}$$

3. (8pts) Given the output of your clustering algorithm as $C_1 = \{1, 2, 3, 4\}$, $C_2 = \{5, 6, 7, 8\}$, and a hand labeled clustering of $C_1 = \{3, 4\}$, $C_2 = \{1, 2, 5, 6, 7, 8\}$, what is the weighted average purity of the clusters created by the clustering algorithm?

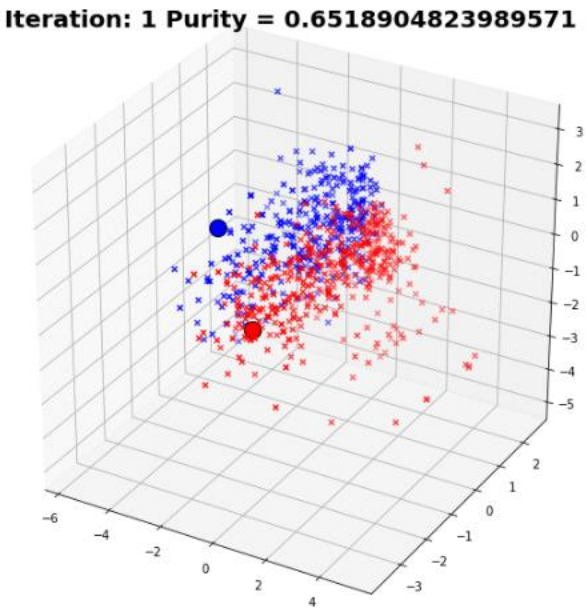
$$Cluster\ I: Purity = \frac{1}{4} * (\max(2, 2)) = \frac{2}{4}$$

$$Cluster\ II: Purity = \frac{1}{4} * (\max(0, 4)) = \frac{4}{4} = 1$$

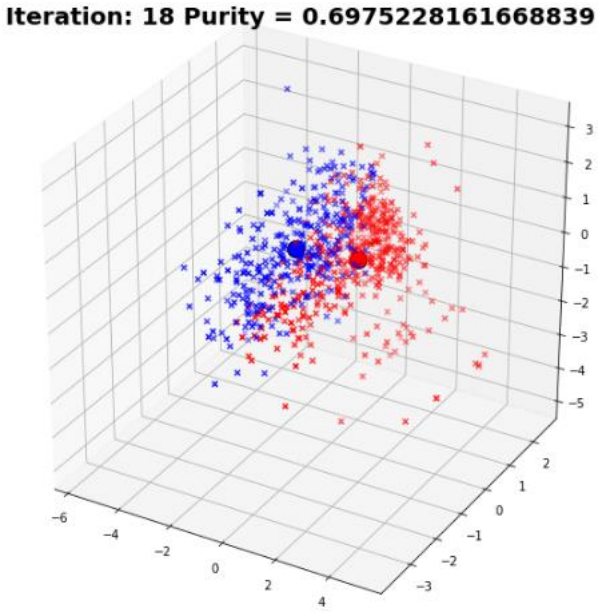
$$\boldsymbol{Total\ Purity} = \frac{1}{8} * \left(4 * \frac{2}{4} + 4 * \frac{4}{4}\right) = \frac{1}{8} * 6 = \frac{6}{8} = 0.75 = \mathbf{75\%}$$

2 Clustering

Visualization of initial clustering for k=2:



Visualization of terminal clustering for k=2:



3 Extra Credit

- K_2.mp4 – opencv video attached of K-means where k=2
- K_3.mp4 – opencv video attached of K-means where k=3

- K_4.mp4 – opencv video attached of K-means where k=4