Name: Dinh Nguyen

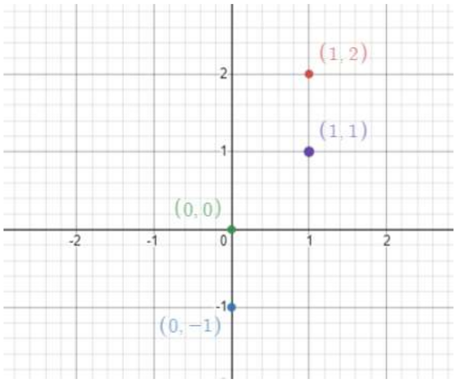
CS 383

Homework 2

Part 1. Theory

1. Given 2 clusters

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



a) Weighted Average intra-cluster distance using Euclidian distance.

$$G1 = \sum d(x,y)/2(|C1|),\ x,y \in C1$$
$$= \frac{\sqrt{(1-0)^2+(2-(-1))^2}}{2*2} = 0.79$$
$$G2 = \sum d(x,y)/2(|C2|),\ x,y \in C2$$
$$= \frac{\sqrt{(0-1)^2+(0-1)^2}}{2*2} = 0.354$$
$$Weighted\ Average\ Intra-cluster\ distance$$
$$W2 = \frac{|C1|G1+|C2|G2}{N} = \frac{2*0.79+2*0.354}{4} = 0.572$$

b) Single Link similarity between clusters using cosine similarity.
Cosine similarity for each pair of points across 2 clusters:

$$sim((1,2),(1,1)) = \frac{1*1+2*1}{\sqrt{1^2+2^2}*\sqrt{1^2+1^2}} = 0.95$$
$$sim((0,-1),(1,1)) = \frac{0*1+(-1)*1}{\sqrt{0^2+(-1)^2}*\sqrt{1^2+1^2}} = -0.71$$
$$sim(Ci,Cj) = \max sim(x,y)\ , x \in Ci, y \in Cj = 0.95$$

It is invalid to measure the cosine similarity with point (0,0).
Thus, single link similarity between the clusters is 0.95.

c) Complete link similarity between the clusters if we are using cosine similarity

$$sim((1,2),(1,1)) = \frac{1*1+2*1}{\sqrt{1^2+2^2*\sqrt{1^2+1^2}}} = 0.95$$

$$sim((0,-1),(1,1)) = \frac{0*1+(-1)*1}{\sqrt{0^2+(-1)^2*\sqrt{1^2+1^2}}} = -0.71$$

$$sim(Ci,Cj) = \min sim(x,y)\ ,x \in Ci, y \in Cj = -0.71$$

The complete link similarity is the similarity of the furthest points between the clusters, so it will have the minimum among the similarity values of the pairs. Thus, the complete link similarity of the clusters is -0.71.

d) The average link similarity between the clusters using cosine similarity

$$sim((1,2),(1,1)) = \frac{1*1+2*1}{\sqrt{1^2+2^2*\sqrt{1^2+1^2}}} = 0.95$$

$$sim((0,-1),(1,1)) = \frac{0*1+(-1)*1}{\sqrt{0^2+(-1)^2*\sqrt{1^2+1^2}}} = -0.71$$

$$sim(Ci,Cj) = \frac{1}{|C1||C2|}\sum sim(x,y)\ ,x \in Ci, y \in Cj$$

$$= \frac{0.95+-0.71}{2*2} = 0.06$$

Average link similarity is the average pair-wise similarity between clusters. Thus, the average link similarity of the clusters is 0.06.

2. Fourth derivative at j, W_j''''

$$W_j' = \frac{W_{j+1}-W_{j-1}}{2}$$

$$W_j'' = \frac{W_{j+1}'-W_{j-1}'}{2} = \frac{W_{j+2}-W_j-W_j+W_{j-2}}{4} = \frac{W_{j+2}-2W_j+W_{j-2}}{4}$$

$$W_j''' = \frac{W_{j+2}'-2W_j'+W_{j-2}'}{4} = \frac{W_{j+3}-W_{j+1}-2(W_{j+1}-W_{j-1})+W_{j-1}-W_{j-3}}{8} = \frac{W_{j+3}-3W_{j+1}+3W_{j-1}-W_{j-3}}{8}$$

$$W_j'''' = \frac{W_{j+3}'-3W_{j+1}'+3W_{j-1}'-W_{j-3}'}{8} = \frac{W_{j+4}-W_{j+2}-3(W_{j+2}-W_j)+3(W_j-W_{j-2})-(W_{j-2}-W_{j-4})}{16}$$

$$= \frac{W_{j+4}-4W_{j+2}+6W_j-4W_{j-2}+W_{j-4}}{16}$$

3. Clustering using algorithm.

C1 = {1,2,3,4}; C2 = {5, 6, 7, 8}

Now, given the hand labeled clustering as C1 = {3,4} and C2={1,2,5,6,7,8}, we can color the elements from the hand labeled one with red color (C1) and blue color for C2. Below would be the updated clusters.

C1 = {1,2,3,4}; C2 = {5, 6, 7, 8}

$$Purity\ C1 = \frac{maxNij}{|C1|} = \frac{max(2,2)}{4} = 0.5$$

$$Purity\ C2 = \frac{maxNij}{|C2|} = \frac{max(4,0)}{4} = 1$$

$$Weighed\ Average\ Purity = \frac{1}{N}\sum |Ci| Purity(Ci) = \frac{1}{8}(4*0.5+4*1) = 0.75$$

Weighed Average purity of the clusters is 75%