CSCI 4360/6360 Data Science II
Department of Computer Science
University of Georgia

Assignment 4: I Heard You Like Graphs

Aditya Shinde

Out October 5, 2017

Questions

Spectral Clustering **[40pts]**

1

$\Theta = 1.5$. Since the farthest points in each cluster are less than 1.5 units apart but the nearest points between both clusters are more than 1.5 units apart.

2

For the best clustering and getting a real solution, the graph should be disconnected. It should have k connected components. Even if it does not, spectral clustering will work but the first eigen vector will be all 1. In this case, it is possible to have 2 connected components by selecting $\Theta = 1.5$. That will connect all the points in the same cluster but exclude connections across different clusters. So if all elements in the adjacency matrix are symetric and at the most 1, the corrosponding eigen vectors will be unit vectors. For instance, in the given diagram, there are two different connected components.
One of the components has m_1 points and the other has m_2 points. The A matrix will

have the shape $(m_1 + m_2, m_1 + m_2)$

$$\begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & 1 & \dots & 1 \\ 1 & 1 & 1 & 1_{(m_1,m_1)} & \dots & 1 \\ \vdots & \vdots & \vdots & \vdots & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1_{(m_1+m_2,m_1+m_2)} \end{bmatrix}$$

The eigen vectors for these will be

$$\begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1_{m_1} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \; and \; \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0_{m_1} \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

The other eigen values will be zero.

5

The PCA of X is done in terms of its covariance matrix $Q = X^T X$

$$Q = U\Sigma V^T$$

Here U are the principle components of X. The SVD of Y is written as,

$$Y = U\Sigma V^T$$

$$Y = X^T$$

So,

$$(X^T)^T = \left(U\Sigma V^T\right)^T$$

$$X = V\Sigma U^T$$

$$X^T X = V\Sigma U^T U\Sigma V^T$$

U is an orthonormal matrix. So $U^T U = I$

$$X^T X = V\Sigma^2 V^T$$

This corrosponds to the PCA equation above. With the difference being that the eigen values are the squares of the singular values.

Hierarchical Clustering **[20pts]**

1

$$\Delta(X,Y) = \sum_{i\in X\cup Y} ||\vec{x}_i - \vec{\mu}_{X\cup Y}||^2 - \sum_{i\in X} ||\vec{x}_i - \vec{\mu}_X||^2 - \sum_{i\in Y} ||\vec{x}_i - \vec{\mu}_Y||^2$$

$$= \sum_{i\in X\cup Y} ||x_i||^2 - 2\sum_{i\in X\cup Y} ||x_i||\mu_{X\cup Y} + \sum_{i\in X\cup Y} \mu_{X\cup Y}^2$$
$$- \left(\sum_{i\in X} ||x_i||^2 - 2\sum_{i\in X} ||x_i||\mu_X + \sum_{i\in X} \mu_X^2\right)$$
$$- \left(\sum_{i\in Y} ||x_i||^2 - 2\sum_{i\in Y} ||x_i||\mu_Y + \sum_{i\in Y} \mu_Y^2\right)$$

Since the norm is squared and the summation is over groups of points,

$$\sum_{i\in X\cup Y} ||x_i||^2 = \sum_{i\in X} ||x_i||^2 + \sum_{i\in Y} ||x_i||^2$$

And so if we put this in the previous equation,

$$\Delta(X,Y) = 2\sum_{i\in X} ||x_i||\mu_X - \sum_{i\in X} \mu_X^2 + 2\sum_{i\in Y} ||x_i||\mu_Y - \sum_{i\in Y} \mu_Y^2 - 2\sum_{i\in X\cup Y} ||x_i||\mu_{X\cup Y} + \sum_{i\in X\cup Y} \mu_{X\cup Y}^2$$

Now the global mean can be written in terms of μ_X and μ_Y as,

$$\mu_{X\cup Y} = \frac{(n_X\mu_X + n_Y\mu_Y)}{n_X + n_Y}$$

$$= 2\sum_{i\in X} ||x_i||\mu_X - n_X\mu_X^2 + 2\sum_{i\in Y} ||x_i||\mu_Y - n_Y\mu_Y^2 - 2\sum_{i\in X\cup Y} ||x_i||\mu_{X\cup Y} + (n_X+n_Y)\left(\frac{(n_X\mu_X + n_Y\mu_Y)}{n_X + n_Y}\right)^2$$

$$= \left(\frac{(n_X\mu_X + n_Y\mu_Y)^2}{n_X + n_Y}\right) - n_X\mu_X^2 - n_Y\mu_Y^2 - 2\sum_{i\in X\cup Y} ||x_i||\mu_{X\cup Y} + 2\sum_{i\in Y} ||x_i||\mu_Y + 2\sum_{i\in X} ||x_i||\mu_X$$

$$= \left(\frac{(n_X\mu_X + n_Y\mu_Y)^2}{n_X + n_Y}\right) - n_X\mu_X^2 - n_Y\mu_Y^2 + 2\left(-\frac{(n_X\mu X + n_Y\mu Y)^2}{n_X + n_Y} + n_X\mu_X^2 + n_Y\mu_Y^2\right)$$

$$= -\frac{n_X n_Y(\mu_X - \mu_Y)^2}{n_X + n_Y} + 2\left(-\frac{(n_X\mu X + n_Y\mu Y)^2}{n_X + n_Y} + n_X\mu_X^2 + n_Y\mu_Y^2\right)$$

$$= -\frac{n_X n_Y (\mu_X - \mu_Y)^2}{n_X + n_Y} + 2\left(\frac{n_X n_Y (\mu_X - \mu_Y)^2}{n_X + n_Y}\right)$$

$$= \frac{n_X n_Y (\mu_X - \mu_Y)^2}{n_X + n_Y}$$

2

No. Ward's metric will not always merge the clusters with closer centers. Ward's metric computes the total cost of merging the clusters. So even if the pair in P_2 is closer than the pair in P_1, if the number of points in the clusters in P_1 is greater than that of P_2, agglomerative clustering will merge P_1.

$$\Delta(X,Y) = \sum_{i \in X \cup Y} ||\vec{x}_i - \vec{\mu}_{X \cup Y}||^2 - \sum_{i \in X} ||\vec{x}_i - \vec{\mu}_X||^2 - \sum_{i \in Y} ||\vec{x}_i - \vec{\mu}_Y||^2$$

So, if,

$$n_{X \cup Y \in P_2} >> n_{X \cup Y \in P_1}$$

Then, it is possible to have a certain n for which,

$$\Delta(X,Y)_{X \cup Y \in P_2} > \Delta(X,Y)_{X \cup Y \in P_1}$$