Daniel Chirita (dc3316) **AFDS - Homework 7** April 30, 2019

Problem 1

> *Prove that optimal 2-means (k-means with k = 2) clustering in two dimensions can be done in polynomial time.*

Assuming a linear discriminant function, in 2 dimensions a 2-means decision boundary is the perpendicular bisector of the line segment joining the two cluster kernels. Observe that the optimal solution is contained in the set of all possible 2-segmentations of the set of points. The total number of such 2-segmentations is $\leq$ nC2 which is polynomial in n. QED.

Problem 2

> *In the standard version of the k-means clustering problem, the centers are not required to be data points themselves. Consider a variant where we need the cluster centers to be picked from among the given data points. We will show that the cost of the optimal solution of this variant is at most twice the cost of the optimal solution of the standard version. It suffices to show that in each cluster of the optimal solution to the standard version, we can replace the centroid by one of the data points without increasing the cost of the cluster by more than a factor 2. This can be shown as follows.*

1. *Prove that for any x_i and x_j, $||x_i - x_j||^2 = ||x_i - c||^2 + ||x_j - c||^2 - 2(x_i - c) \cdot (x_j - c)$*
2. *Use the above to show that if we replace the centroid by a data point $x_i \in X$ then the new cost of the cluster i.e. $\Sigma_{x \in X} ||x - x_i||^2$, is equal to $\Sigma_{x \in X} ||x - c||^2 + m \cdot ||x_i - c||^2$*
3. *Show by averaging that for some $x_i \in X$, the above cost is at most $2 \cdot \Sigma_{x \in X} ||x - c||^2$.*

Solution:

1. Observe the following:

$$\begin{aligned} ||x_i - x_j||^2 &= ||x_i - x_j - c + c||^2 \\ &= ||(x_i - c) - (x_j - c)||^2 \\ &= ||x_i - c||^2 - 2(x_i - c) \cdot (x_j - c) + ||x_j - c||^2 \\ &= ||x_i - c||^2 + ||x_j - c||^2 - 2(x_i - c) \cdot (x_j - c) \end{aligned} \tag{1}$$

2. The original cost is $C_c = \Sigma_{x \in X}(x - c)^2$. The new cost is:

$$\begin{aligned} C_{x_i} &= \Sigma_{x \in X} \ ||x - c||^2 + ||x_i - c||^2 - 2(x - c) \cdot (x_i - c) \\ &= \Sigma_{x \in X} \ ||x - c||^2 + \Sigma_{x \in X} ||x_i - c||^2 - 2\Sigma_{x \in X}(x - c) \cdot (x_i - c) \end{aligned} \tag{2}$$

Observe that since c is the centroid of X and vector $(x_i - c)$ is invariant in X:

$$\begin{aligned} 2\Sigma_{x\in X}(x-c)\cdot(x_i-c) &= 2(x_i-c)\cdot[\Sigma_{x\in X}(x-c)] \\ &= 2(x_i-c)\cdot[0] \\ &= 0 \end{aligned} \tag{3}$$

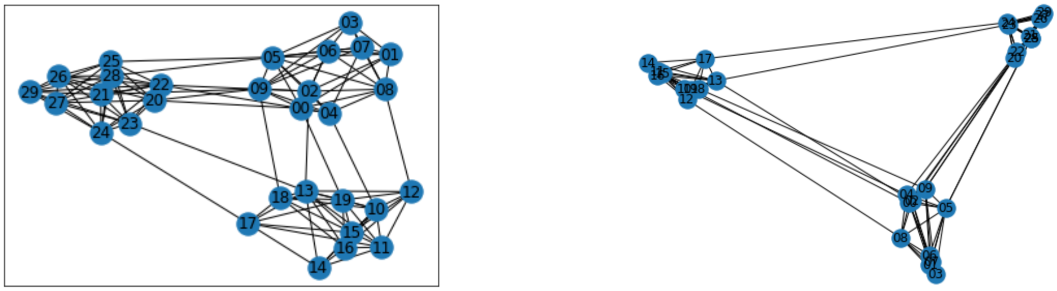
The total cost for $X = [m]$ then becomes:

$$C_{x_i} = \Sigma_{x\in X}\ ||x-c||^2 + m\cdot||x_i-c||^2 \tag{4}$$

3. Observe that $||x_i - c||^2$ is the smallest of all $||x-c||^2$ in X, since we picked the point closest to c as the new cluster kernel. Since this is the case, $\Sigma^m_{x\in X}\ ||x-c||^2 \geq m\cdot||x_i-c||^2$. Therefore, $C_{x_i} \leq 2C_c$. QED.

Problem 3

Construct a random graph with three clusters as follows. Take - three groups of vertices with $n = 10$ *vertices each. In each group add each of the n choose 2 edges with probability 0.8 independently and uniformly at random. Then, between any two groups add each of the* n^2 *possible edges with probability 0.05 independently and uniformly at random. Using the networkx library, draw the graph with the positions of the vertices chosen according to the eigenvectors corresponding to the second and third smallest eigenvalues of the Laplacian of the graph.*



(a) Original graph. Plotted with *networkx.draw()*.

(b) Spectral gap representation using the second and third lowest eigenvalues of the Laplacian of the original graph. Plotted with *networkx.draw_spectral()*.

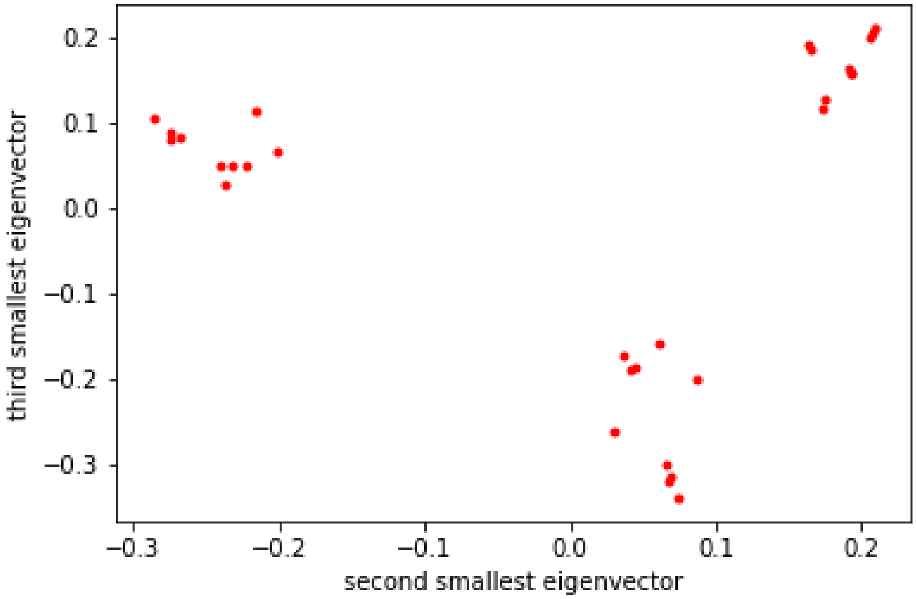


Figure 2: Spectral gap representation using the second and third lowest eigenvalues of the Laplacian of the original graph. Plotted with *pyplot* directly from the eigenvectors.

Problem 4

> *Let $G = (V, E)$ be an unweighted graph and let L be the Laplacian of G. Let us assume without loss of generality that $V = [n]$ for some n. Define for any subset $A \subseteq V$, a characteristic vector $1_A \in R^n$ whose i^{th} component is 1 if $i \in A$ and 0 otherwise.*
> *Suppose that G has k connected components and let $A_1, ..., A_k \subseteq V$ denote the vertex sets of the connected components of G. Then, prove that the subspace of the eigenvectors of L with eigenvalue 0 is spanned by the vectors $1_{A_1}, ..., 1_{A_k}$.*
> *(Hint: Recall from the lecture that for any $x \in R^n, x^T Lx = \Sigma_{(i,j)\in E}(x_i - x_j)^2$. Note that if x is an eigenvector of L with eigenvalue 0 then, $x^T Lx = 0$. When is $\Sigma_{(i,j)\in E}(x_i - x_j)^2$ equal to 0?)*

We know that $Lx = 0$, since x is an eigenvector of L with eigenvalue 0. Then,

$$x^T Lx = \frac{1}{2}\Sigma_{(i,j)\in E}(x_i - x_j)^2 = 0 \tag{5}$$

Which can only be the case when $\forall i, j \in E, x_i = x_j$, meaning that $x = \alpha[1, 1, ..., 1]^T$ for some constant α. QED.