Assignment 7

CS-UH-2218: Algorithmic Foundations of Data Science

Assignments are to be submitted in groups of two or three. *Upload the solutions on NYU classes as one PDF file for theoretical assignments and separate source code files for each programming assignment. Submit only one copy per group. Clearly mention the participant names on each file you submit.*

Problem 1 (10 points).
Prove that optimal 2-means (k-means with $k = 2$) clustering in two dimensions can be done in polynomial time.

Hint: Show that the two clusters defined by any two centers can be separated by a line. In how many distinct ways can you split a set of n points in the plane into two parts using a line?

Problem 2 (10 points).
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.

Let $X = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_m\}$ be the data points in one cluster and let $\mathbf{c}$ be their centroid. Note that the cost of this cluster is $\sum_{\mathbf{x} \in X} \|\mathbf{x} - \mathbf{c}\|^2$.

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

To show that the above bound is tight, give a simple example where replacing the cluster center by any data point does increase the cost by a factor of two.

Problem 3 (10 points).
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 $\binom{n}{2}$ 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.

Problem 4 (10 points)**.**
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* $\mathbb{1}_A \in \mathbb{R}^n$ whose i^{th} component is 1 if $i \in A$ and 0 otherwise.

Suppose that G has k connected components and let $A_1, \cdots, A_k \subseteq V$ 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 $\mathbb{1}_{A_1}, \cdots, \mathbb{1}_{A_k}$.

Hint: Recall from the lecture that for any $x \in \mathbb{R}^n$, $x^T L\, x = \sum_{\{i,j\} \in E} (x_i - x_j)^2$. Note that if x is an eigenvector of L with eigenvalue 0 then, $x^T L\, x = 0$. When is $\sum_{\{i,j\} \in E} (x_i - x_j)^2$ equal to 0?