GEOMETRIC DATA ANALYSIS: PROBLEM SET 1

1. Problems

For these problems, I would like you to begin by writing your own implementation of the k-means clustering algorithm, k-medians clustering algorithm, single-linkage clustering, and spectral clustering. For the time being, you can assume (for the k-medians and single-linkage clustering) that the finite metric space of data (X, ∂_X) is a subset of Euclidean space $\mathbb{R}^n$ and inherits the subspace metric. As a word to the wise, in subsequent problem sets, you may need to use other metrics — please structure your code accordingly. For spectral clustering, we will experiment with different ways of creating the graph that is input, so please structure your code accordingly. If not otherwise specified, use $w_{ij} = e^{-\partial_X(x_i,x_j)^2/2\sigma^2}$.

Next, write code to sample from distributions constructed as mixtures of spherical Gaussians in Euclidean space; you will use this to explore the performance and behavior of your clustering code. This code should take as parameters the dimension, the number of Gaussians, their centers and widths, and the number of samples to generate.

(1) Explore what happens when you use k-means and single-linkage clustering to cluster points sampled from spherical Gaussians as the distance between the centers of the spherical Gaussians varies. Discuss your findings.
(2) This problem is about the performance of k-means clustering in the presence of noise.
   (a) The easiest kind of noise to analyze and reason about is Gaussian noise centered on the data. Explore the behavior of k-means clustering in the presence of varying amounts of Gaussian noise centered at the data points. Discuss.
   (b) A more challenging noise model involves arbitrary or adversarially placed noise. Assume that we are given $X \subseteq \mathbb{R}^n$ and suppose you are allowed to add ℓ points at any location in $\mathbb{R}^n$. How much can you cause the clustering to change under k-means and single-linkage?
(3) Can you make a hierarchical version of k-means or k-medians as k varies?
(4) Use your k-means and single-linkage clustering algorithm to cluster the included data set (which consists of points in $\mathbb{R}^3$). How many clusters are there? Discuss.
(5) Give an example to show that k-means clustering does not have the property that shrinking distances within clusters and expanding them between clusters necessarily results in the same clustering.
(6) Please write code to sample from two concentric annuli in $\mathbb{R}^2$; an annulus centered at the origin is specified by the equation $r_1 \leq x^2 + y^2 \leq r_2$, where r_1 and r_2 are the inner and outer radii. Cluster with k-means and spectral clustering, with $k = 2$. Discuss the results.

(7) Compare spectral clustering with the Gaussian weighted graph (as above) and with the m-nearest neighbors graph; try the annuli sxamples and various mixture of Gaussian models, with and without noise. Discuss the results.
(8) **(Extra credit)** Please write code to simulate the shuffling random walk and experimentally verify how many riffle shuffles it takes to bring a deck of cards close to random. (Hint: compute the empirical distribution of the top card after one shuffle, then two, then three, and so on.)
(9) **(Extra credit)** Explain the statement and proof of the Rayleigh-Ritz theorem.
(10) **(Extra credit)** Use the LP relaxation discussed in class to perform k-medians. Explore the performance on mixtures of Gaussians and compare to the iterative algorithm. Which does a better job as the centers of the Gaussians get closer together?