Homework 4
MATH 123 - Spring 2023
Tufts University, Department of Mathematics
Due: February 21, 2023

Question 1

Let $\{x_i\}_{i=1}^n \subset \mathbb{R}^D$. Let $F : \mathbb{R}^D \to [0, \infty)$ be

$$F(y) = \sum_{i=1}^{n} \|x_i - y\|_2^2.$$

Prove that F is minimized for $y = \frac{1}{n}\sum_{i=1}^n x_i$, i.e. at the mean of $\{x_i\}_{i=1}^n$.

Question 2

Suppose $x_1, \ldots, x_n \subset \mathbb{R}^D$ are data, and an *outlier* x^o is added with the property that for $\delta > 0$ fixed, $\|x_i - x^o\|_2 > \delta$ for all $i = 1, \ldots, n$. Suppose we run K-means on this data with $K = 2$.

(a) Argue that as $\delta \to +\infty$, one of the clusters learned by K-means will consist only of x^o.
(b) This *lack of robustness to outliers* is sometimes considered a defect of K-means. Suggest some changes to the K-means algorithm to improve its robustness to outliers.
(c) Instead of thinking of the lack of robustness to outliers as a defect, can you think of any reasons it may be a virtue?

Question 3

K-means is often combined with a *feature extraction* step in which the data to be clustered is first transformed to a more convenient form. As the course progresses, we will consider some *data-dependent* feature extraction methods, but for now, let us consider a very particular feature extraction method: converting Cartesian to polar coordinates in $\mathbb{R}^2$.

(a) Load the data in 'CircularK_Means.m', and run K-means with $K = 2$, displaying your labels as colors on the plotted data. In terms of the K-means functional, why does this method produce the "incorrect" clusters it does?
(b) Convert the data to polar coordinates and run K-means again to show the data can be correctly labeled in this case.
(c) Explain what about the polar coordinate representation is convenient for this data.

Question 4

(a) In MATLAB, create a dataset in which single linkage and complete linkage hierarchical clustering differ substantially. Demonstrate this by computing the dendrograms using the built-in 'linkage' function in MATLAB, and arguing that they capture different structure in the data.
(b) Argue why the two linkage methods differ on this data in terms of their mathematical formulation.