# Homework 6

ESE 4020/5420

Due December 4, 2024 at 11:59pm

**Problem 1.** *k-means is suboptimal.* Consider the following 4 points on the 2-D plane (think of $t$ as a positive number larger than 3):

$$\{(-1,-t),(+1,-t),(-1,t),(+1,t)\}.$$

We want to use $k$-means to cluster these points into 2 clusters.

(a) What are the optimal centers for the 2-Means problem? (Here, by optimal we mean the two centers that minimize the 2-Means objective over the 4 data points.)

(b) Let us now see what the K-Means algorithm would give us (here, $K = 2$). Find an initial set of centers so that if the K-Means algorithm starts with those centers then it will not find the optimal centers found in part (a).

**Problem 2.** K-means clustering can be viewed as an optimization problem that attempts to minimize some objective function. For the given objectives, determine the update rule for the centroid, $c_k$ of the $k$-th cluster $C_k$ . In other word, find the optimal $c_k$ that minimizes the objective function. The data $x$ contains $p$ features.

(a) Show that setting the objective to the sum of the squared Euclidean distances of points from the center of their clusters,

$$\sum_{k=1}^{K}\sum_{x\in C_k}\sum_{i=1}^{p}(c_{ki}-x_i)^2$$

results in an update rule where the optimal centroid is the mean of the points in the cluster. [Hint: Look at the derivatives of the objective with respect to each $c_{ki}$, and set them to zero. Solving these equations should determine $c_k$ for each $k$.]

(b) Show that setting the objective to the sum of the Manhattan distances of points from the center of their clusters,

$$\sum_{k=1}^{K}\sum_{x\in C_k}\sum_{i=1}^{p}|c_{ki}-x_i|$$

results in an update rule where the optimal centroid is the median of the points in the cluster.

**Problem 3.** Load the Labeled Faces in the Wild dataset from sklearn. You can load this data as follows:

```
from sklearn.datasets import fetch_lfw_people
faces = fetch_lfw_people(min_faces_per_person=60)
```

For this exercise, we will use PCA on image data, in particular pictures of faces, to extract features.

(a) Perform PCA on the dataset to find the first 150 components. Since this is a large dataset, you should use randomized PCA instead, which can also be found on sklearn. Show the eigenfaces associated with the first 1 through 25 principal components.

(b) Using the first 150 components you found, reconstruct a few faces of your choice and compare them with the original input images.