Name:

University of Illinois Spring 2020

CS 446/ECE 449 Machine Learning
Homework 7: K-Means

Due on Tuesday April 7 2020, noon Central Time

1. **[16 points]** K-Means

   We are given a dataset $\mathcal{D} = \{(x)\}$ of 2d points $x \in \mathbb{R}^2$ which we are interested in partitioning into K clusters, each having a cluster center μ_k ($k \in \{1, \ldots, K\}$) via the k-Means algorithm. This algorithm optimizes the following cost function:

$$\min_{\mu_k, r} \sum_{x \in \mathcal{D}, k \in \{1,\ldots,K\}} \frac{1}{2} r_{x,k} \|x - \mu_k\|_2^2 \quad \text{s.t.} \quad \begin{cases} r_{x,k} \in \{0,1\} & \forall x \in \mathcal{D}, k \in \{1, \ldots, K\} \\ \sum_{k \in \{1,\ldots,K\}} r_{x,k} = 1 & \forall x \in \mathcal{D} \end{cases} \tag{1}$$

   (a) (1 point) What is the domain for μ_k?

   Your answer:

   (b) (3 points) Given fixed cluster centers μ_k $\forall k \in \{1, \ldots, K\}$, what is the optimal $r_{x,k}$ for the program in Eq. (1)? Provide a reason?

   Your answer:

   (c) (3 points) Given fixed $r_{x,k}$ $\forall x \in \mathcal{D}, k \in \{1, \ldots, K\}$, what are the optimal cluster centers μ_k $\forall k \in \{1, \ldots, K\}$ for the program in Eq. (1)? Reason by first computing the derivative w.r.t. μ_k.

   Your answer:

Name:

(d) (5 points) Using Pseudo-code, sketch the algorithm which alternates the aforementioned two steps. Is this algorithm guaranteed to converge? Reason? Is this algorithm guaranteed to find the global optimum? Reason?

Your answer:

(e) (4 points) Complete `A7_KMeans.py` by implementing the aforementioned two steps. For the given dataset, after how many updates does the algorithm converge, what cost function value does it converge to and what are the obtained cluster centers?

Your answer: