ECE 6524 – Deep Learning
No. HW-01-6524

Homework Assignment #1: *k*-means Clustering

1. The *k*-means problem and its cost function

Mathematically, *k*-means problem can be stated as follows: the input is a set *S* of data points and the goal is to choose *k* representatives of S. the distortion on a point $x \in S$ is then the distance to its closest representative. The overall goal is to make sure that *every* point in *S* has low distortion; that is to minimize the *maximum* distortion in *S*. in most applications, we are more interested in minimizing the *typical* (i.e., *average*) distortion. The most popular formulation of this is the *k*-means cost function, which assumes that points lie in Euclidean space.

k-MEANS CLUSTERING
Input: Finite set $S \subset \mathbb{R}^d$; integer k.
Output: $T \subset \mathbb{R}^d$ with $|T| = k$.
Goal: Minimize $\text{cost}(T) = \sum_{x \in S} \min_{z \in T} \|x - z\|^2$.

The partition/clustering induces an optimal clustering of the data set, $S = \cup_{z \in T} C_z$, where

$$C_z = \{x \in S : \text{the closest representative of } x \text{ is } z\}.$$

Thus, the *k*-means cost function can be written as

$$\text{cost}(T) = \sum_{z \in T} \sum_{x \in C_z} \|x - z\|^2.$$

The cost function can be further formulated as following:

$$\operatorname{cost}(C_1, \ldots, C_k; z_1, \ldots, z_k) = \sum_{j=1}^{k} \sum_{x \in C_j} \|x - z_j\|^2.$$

Prove the following statement:

For each cluster C_j and its representative z_j (j = 1, 2, …, k), the cost function can be written as:

$$\operatorname{cost}(C_j; z_j) = \operatorname{cost}\left(C_j; \operatorname{mean}(C_j)\right) + |C_j| ||z_j - \operatorname{mean}(C_j)||^2.$$

2. The convergence of the *k*-means algorithm

The *k*-means algorithm can be described as follows:

initialize centers $z_1, \ldots, z_k \in \mathbb{R}^d$ and clusters $C_1, \ldots, C_k$ in any way
repeat until there is no further change in cost:
  for each j: $C_j \leftarrow \{x \in S$ whose closest center is $z_j\}$
  for each j: $z_j \leftarrow \operatorname{mean}(C_j)$

Prove the convergence property of the K-means algorithm; that is:

During the course of the K-means algorithm, the cost monotonically decreasing.

Note that you shall give the detailed, intermediate steps for the proofs.