Quiz 07

October 22nd, 2026 15 minutes | 5 points

Name _____

Instructions. Answer all three parts. Show your work and briefly justify each answer. Parts (a) and (b) are worth 2 points each and part (c) is worth 1 point. Use the space provided.

1. More clusters never hurt the cost (Homework 06, Problem 2) 5 points

Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ and $[n] = \{1, \dots, n\}$. For $1 \le k \le n$, a *k-partition* is a split of $[n]$ into disjoint non-empty sets $S_1, \dots, S_k$ whose union is $[n]$. For a non-empty $S \subseteq [n]$, write $\bar{\mathbf{x}}_S = \frac{1}{|S|} \sum_{i \in S} \mathbf{x}_i$ for the mean of the points in S, and $\bar{\mathbf{x}} = \bar{\mathbf{x}}_{[n]}$ for the overall mean.

Definition. For a k-partition $S_1, \dots, S_k$ and *centers* $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k \in \mathbb{R}^d$, the *$k$-means cost* is

$$\operatorname{cost}(S_1, \dots, S_k; \boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k) = \sum_{\ell=1}^{k} \sum_{i \in S_\ell} \|\mathbf{x}_i - \boldsymbol{\mu}_\ell\|_2^2.$$

We write opt_k for the minimum of the cost over all k-partitions and all choices of centers.

(a) Show that $\operatorname{opt}_n = 0$. (2 points)

Solution. Every set in an n-partition is non-empty, and n disjoint non-empty subsets of $[n]$ must each have exactly one element. So the only n-partition puts every point in its own cluster, $S_i = \{i\}$ (up to relabeling). With centers $\boldsymbol{\mu}_i = \mathbf{x}_i$, the cost is $\sum_{i=1}^{n} \|\mathbf{x}_i - \mathbf{x}_i\|_2^2 = 0$. The cost is a sum of squared norms, so it is never negative. Therefore $\operatorname{opt}_n = 0$.

(b) Let $1 \le k < n$. Starting from a k-partition and centers whose cost is opt_k, build a $(k+1)$-partition and centers whose cost is at most opt_k. Conclude that $\mathrm{opt}_{k+1} \le \mathrm{opt}_k$. *Hint:* since $k < n$, some cluster has at least two points. (2 points)

Solution. Let $S_1, \dots, S_k$ and $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k$ have cost opt_k. The k clusters hold $n > k$ points in total, so some cluster S_ℓ has at least two points. Pick $i \in S_\ell$ and move $\mathbf{x}_i$ into a new cluster $S_{k+1} = \{i\}$ with center $\boldsymbol{\mu}_{k+1} = \mathbf{x}_i$, keeping every other cluster and center. Both $S_\ell \setminus \{i\}$ and $\{i\}$ are non-empty, so this is a $(k+1)$-partition. Only the term for $\mathbf{x}_i$ changes: it was $\|\mathbf{x}_i - \boldsymbol{\mu}_\ell\|_2^2$ and is now $\|\mathbf{x}_i - \mathbf{x}_i\|_2^2 = 0$. So the new cost is

$$\mathrm{opt}_k - \|\mathbf{x}_i - \boldsymbol{\mu}_\ell\|_2^2 \le \mathrm{opt}_k \,.$$

Since opt_{k+1} is the minimum over all $(k+1)$-partitions and centers, it is at most this cost, so $\mathrm{opt}_{k+1} \le \mathrm{opt}_k$.

Lemma (sum of squares around a point). For every non-empty $S \subseteq [n]$ and every $\boldsymbol{\mu} \in \mathbb{R}^d$,

$$\sum_{i \in S} \|\mathbf{x}_i - \boldsymbol{\mu}\|_2^2 = \sum_{i \in S} \|\mathbf{x}_i - \bar{\mathbf{x}}_S\|_2^2 + |S| \, \|\bar{\mathbf{x}}_S - \boldsymbol{\mu}\|_2^2.$$

(c) Show that the best single center is the mean of the whole dataset. That is, show that $\boldsymbol{\mu}_1 = \bar{\mathbf{x}}$ is the unique center achieving opt_1, and (1 point)

$$\mathrm{opt}_1 = \sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|_2^2.$$

Solution. The only 1-partition is $S_1 = [n]$, so opt_1 is the minimum over $\boldsymbol{\mu} \in \mathbb{R}^d$ of $\sum_{i=1}^n \|\mathbf{x}_i - \boldsymbol{\mu}\|_2^2$. By the lemma with $S = [n]$,

$$\sum_{i=1}^{n} \|\mathbf{x}_i - \boldsymbol{\mu}\|_2^2 = \sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|_2^2 + n\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_2^2 \ge \sum_{i=1}^{n} \|\mathbf{x}_i - \bar{\mathbf{x}}\|_2^2,$$

with equality exactly when $\|\bar{\mathbf{x}} - \boldsymbol{\mu}\|_2 = 0$, that is, $\boldsymbol{\mu} = \bar{\mathbf{x}}$.