Midterm Exam

October 29th, 2026 75 minutes | 100 points

Name _____

Instructions. Answer all four problems. Show your work and briefly justify each answer. Each problem has four pages, so most of this exam is space for your work! The exam is closed note. You may use a four-function calculator.

1. The mean minimizes squared error (Homework 01, Problem 4) **25 points**

Let X be a random variable and let $\mu = \mathbb{E}[X]$.

(a) Prove that, for every $a \in \mathbb{R}$, (10 points)

$$\mathbb{E}[(X-a)^2] = \mathrm{Var}(X) + (a-\mu)^2.$$

Solution. Write $X - a = (X - \mu) + (\mu - a)$. Expanding the square and taking expectations gives

$$\begin{aligned}\mathbb{E}[(X-a)^2] &= \mathbb{E}[(X-\mu)^2] + 2(\mu-a)\mathbb{E}[X-\mu] + (\mu-a)^2 \\ &= \mathrm{Var}(X) + (a-\mu)^2,\end{aligned}$$

since $\mathbb{E}[X-\mu] = \mathbb{E}[X] - \mu = 0$.

Problem 1 (continued)

(b) Conclude that μ is the unique value of a that minimizes $\mathbb{E}[(X-a)^2]$. (5 points)

Solution. The term $\mathrm{Var}(X)$ does not depend on a, while $(a-\mu)^2 \geq 0$, with equality exactly when $a = \mu$. Therefore the unique minimizer is $a = \mu$.

Problem 1 (continued)

For part (c) you may use the following, without proof.

Definition (mean and covariance of a random vector). A *random vector* $\mathbf{z} = (z_1, \dots, z_m)^\top$ is a vector whose entries are random variables. Its *mean* and *covariance matrix* are

$$\mathbb{E}[\mathbf{z}] = \begin{pmatrix} \mathbb{E}[z_1] \\ \vdots \\ \mathbb{E}[z_m] \end{pmatrix} \in \mathbb{R}^m, \qquad \mathrm{Cov}(\mathbf{z}) = \mathbb{E}\big[(\mathbf{z} - \mathbb{E}[\mathbf{z}])(\mathbf{z} - \mathbb{E}[\mathbf{z}])^\top\big] \in \mathbb{R}^{m \times m},$$

where the expectation of a random matrix is taken entry by entry. The (i, j) entry of $\mathrm{Cov}(\mathbf{z})$ is $\mathrm{Cov}(z_i, z_j)$, and its diagonal entries are the variances $\mathrm{Var}(z_i)$.

(c) Now let $\mathbf{x} = (x_1, \dots, x_n)^\top$ be a random vector in $\mathbb{R}^n$, and let $\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}]$. Prove that, (7 points)

$$\mathbb{E}\big[\|\mathbf{x} - \mathbf{a}\|_2^2\big] = \mathrm{tr}\,\big(\mathrm{Cov}(\mathbf{x})\big) + \|\mathbf{a} - \boldsymbol{\mu}\|_2^2,$$

for every $\mathbf{a} \in \mathbb{R}^n$.

Solution. Write $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)^\top$, so $\mu_j = \mathbb{E}[x_j]$. Since $\|\mathbf{x} - \mathbf{a}\|_2^2 = \sum_{j=1}^n (x_j - a_j)^2$, linearity of expectation and part (a), applied to each entry x_j, give

$$\mathbb{E}\big[\|\mathbf{x} - \mathbf{a}\|_2^2\big] = \sum_{j=1}^n \mathbb{E}[(x_j - a_j)^2] = \sum_{j=1}^n \mathrm{Var}(x_j) + \sum_{j=1}^n (a_j - \mu_j)^2 = \mathrm{tr}\,\big(\mathrm{Cov}(\mathbf{x})\big) + \|\mathbf{a} - \boldsymbol{\mu}\|_2^2,$$

since the diagonal entries of $\mathrm{Cov}(\mathbf{x})$ are the variances $\mathrm{Var}(x_j)$.

Problem 1 (continued)

(d) Conclude that $\boldsymbol{\mu}$ is the unique vector $\mathbf{a}$ that minimizes $\mathbb{E}[\|\mathbf{x}-\mathbf{a}\|_2^2]$. You may use part (c) even if you did not prove it. (3 points)

Solution. As in part (b), $\operatorname{tr}(\operatorname{Cov}(\mathbf{x}))$ does not depend on $\mathbf{a}$, while $\|\mathbf{a}-\boldsymbol{\mu}\|_2^2 \geq 0$, with equality exactly when $\mathbf{a}=\boldsymbol{\mu}$. So $\boldsymbol{\mu}$ is the unique minimizer.

2. Column space and null space from the SVD (Homework 04, Problem 1) **25 points**

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ have rank $r \geq 1$ and compact SVD

$$\mathbf{X} = \mathbf{U}_r \Sigma_r \mathbf{V}_r^\top = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \qquad \sigma_1 \geq \cdots \geq \sigma_r > 0,$$

with orthonormal $\mathbf{u}_1, \ldots, \mathbf{u}_r \in \mathbb{R}^n$ and orthonormal $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{R}^d$.

(a) Show that $\mathcal{C}(\mathbf{X}) = \text{span}\{\mathbf{u}_1, \ldots, \mathbf{u}_r\}$. *Hint:* For ($\supseteq$), recall that $\mathbf{X}\mathbf{v}_i = \sigma_i \mathbf{u}_i$. (6 points)

Solution. ($\subseteq$) For any $\boldsymbol{\theta} \in \mathbb{R}^d$,

$$\mathbf{X}\boldsymbol{\theta} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i (\mathbf{v}_i^\top \boldsymbol{\theta}) = \sum_{i=1}^{r} \sigma_i (\mathbf{v}_i^\top \boldsymbol{\theta})\, \mathbf{u}_i,$$

a linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_r$. ($\supseteq$) Since $\sigma_i > 0$, we have $\mathbf{u}_i = \mathbf{X}(\mathbf{v}_i/\sigma_i) \in \mathcal{C}(\mathbf{X})$ for each $i \leq r$. Since $\mathcal{C}(\mathbf{X})$ is a subspace, it contains every linear combination of $\mathbf{u}_1, \ldots, \mathbf{u}_r$.

Problem 2 (continued)

(b) Conclude that $\mathbf{P} = \mathbf{U}_r\mathbf{U}_r^\top$ is the orthogonal projection onto $\mathcal{C}(\mathbf{X})$. (4 points)

Solution. The $\mathbf{u}_i$ are orthonormal, hence linearly independent, and by part (a) they span $\mathcal{C}(\mathbf{X})$. So they are an orthonormal basis of $\mathcal{C}(\mathbf{X})$, and the definition of the orthogonal projection gives $\mathbf{P} = \mathbf{U}_r\mathbf{U}_r^\top$.

(c) Show that, for $\mathbf{z} \in \mathbb{R}^d$, $\mathbf{Xz} = \mathbf{0}$ if and only if $\mathbf{V}_r^\top\mathbf{z} = \mathbf{0}$. (6 points)

Solution. ($\Leftarrow$) If $\mathbf{V}_r^\top\mathbf{z} = \mathbf{0}$, then $\mathbf{Xz} = \mathbf{U}_r\Sigma_r(\mathbf{V}_r^\top\mathbf{z}) = \mathbf{0}$. ($\Rightarrow$) If $\mathbf{Xz} = \mathbf{0}$, multiply on the left by $\mathbf{U}_r^\top$ and use $\mathbf{U}_r^\top\mathbf{U}_r = \mathbf{I}_r$:

$$\mathbf{0} = \mathbf{U}_r^\top\mathbf{U}_r\Sigma_r\mathbf{V}_r^\top\mathbf{z} = \Sigma_r\mathbf{V}_r^\top\mathbf{z}.$$

Since Σ_r is diagonal with positive diagonal entries, it is invertible, so $\mathbf{V}_r^\top\mathbf{z} = \mathbf{0}$.

Problem 2 (continued)

(d) Explain why part (c) says exactly that $\mathcal{N}(\mathbf{X}) = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}^\perp$. (4 points)

Solution. The entries of $\mathbf{V}_r^\top \mathbf{z}$ are $\mathbf{v}_1^\top \mathbf{z}, \ldots, \mathbf{v}_r^\top \mathbf{z}$. So $\mathbf{V}_r^\top \mathbf{z} = \mathbf{0}$ if and only if $\mathbf{z}$ is orthogonal to each $\mathbf{v}_i$, which (by linearity of the inner product) holds if and only if $\mathbf{z}$ is orthogonal to every vector in $\text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$. Combined with part (c), $\mathbf{z} \in \mathcal{N}(\mathbf{X})$ if and only if $\mathbf{z} \in \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}^\perp$.

(e) Show that $\mathcal{C}(\mathbf{X}^\top) = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$. (5 points)

Solution. Transposing the SVD gives $\mathbf{X}^\top = \sum_{i=1}^r \sigma_i \mathbf{v}_i \mathbf{u}_i^\top$. This is a compact SVD of $\mathbf{X}^\top$, with the same positive singular values and the roles of the orthonormal sets $\{\mathbf{u}_i\}$ and $\{\mathbf{v}_i\}$ swapped. Applying part (a) to $\mathbf{X}^\top$ gives $\mathcal{C}(\mathbf{X}^\top) = \text{span}\{\mathbf{v}_1, \ldots, \mathbf{v}_r\}$. (Directly: $\mathbf{X}^\top \mathbf{w} = \sum_i \sigma_i (\mathbf{u}_i^\top \mathbf{w}) \mathbf{v}_i$, and $\mathbf{v}_i = \mathbf{X}^\top (\mathbf{u}_i / \sigma_i)$.)

Problem 2 (continued)

3. Training error always prefers $\lambda \downarrow 0$ (Homework 05, Problem 2) 25 points

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ have rank $r \geq 1$ and compact SVD $\mathbf{X} = \sum_{i=1}^{r} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, with $\sigma_1 \geq \cdots \geq \sigma_r > 0$, orthonormal $\mathbf{u}_1, \ldots, \mathbf{u}_r \in \mathbb{R}^n$, and orthonormal $\mathbf{v}_1, \ldots, \mathbf{v}_r \in \mathbb{R}^d$. Let $\mathbf{y} \in \mathbb{R}^n$ and $\lambda > 0$, and let $\widehat{\boldsymbol{\theta}}_\lambda$ denote the ridge regression estimator

$$\widehat{\boldsymbol{\theta}}_\lambda = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^d} \Big\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_2^2 \Big\}.$$

You may use the following two theorems, without proof.

Theorem (the ridge solution). For every $\lambda > 0$, the matrix $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d$ is invertible and $\widehat{\boldsymbol{\theta}}_\lambda = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}_d)^{-1} \mathbf{X}^\top \mathbf{y}$.

Theorem (ridge regression filters the singular values). For every $\lambda > 0$,

$$\widehat{\boldsymbol{\theta}}_\lambda = \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2 + \lambda} (\mathbf{u}_i^\top \mathbf{y})\, \mathbf{v}_i.$$

It is tempting to choose λ by making the training error $\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}_\lambda\|_2^2$ as small as possible. In this problem you will show that this never works. Let $\mathbf{P} = \sum_{i=1}^{r} \mathbf{u}_i \mathbf{u}_i^\top$ be the orthogonal projection onto $\mathcal{C}(\mathbf{X})$.

(a) Show that the fitted values are (5 points)

$$\mathbf{X}\widehat{\boldsymbol{\theta}}_\lambda = \sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} (\mathbf{u}_i^\top \mathbf{y})\, \mathbf{u}_i.$$

Solution. Apply $\mathbf{X}$ to the second theorem and use $\mathbf{X}\mathbf{v}_i = \sigma_i \mathbf{u}_i$:

$$\mathbf{X}\widehat{\boldsymbol{\theta}}_\lambda = \sum_{i=1}^{r} \frac{\sigma_i}{\sigma_i^2 + \lambda} (\mathbf{u}_i^\top \mathbf{y})\, \sigma_i \mathbf{u}_i = \sum_{i=1}^{r} \frac{\sigma_i^2}{\sigma_i^2 + \lambda} (\mathbf{u}_i^\top \mathbf{y})\, \mathbf{u}_i.$$

Problem 3 (continued)

(b) Show that (6 points)

$$\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}_\lambda = \sum_{i=1}^{r} \frac{\lambda}{\sigma_i^2 + \lambda} (\mathbf{u}_i^\top \mathbf{y})\, \mathbf{u}_i + (\mathbf{y} - \mathbf{P}\mathbf{y}).$$

Solution. Write $\mathbf{y} = \mathbf{P}\mathbf{y} + (\mathbf{y} - \mathbf{P}\mathbf{y}) = \sum_{i=1}^{r} (\mathbf{u}_i^\top \mathbf{y}) \mathbf{u}_i + (\mathbf{y} - \mathbf{P}\mathbf{y})$ and subtract part (a), using $1 - \frac{\sigma_i^2}{\sigma_i^2 + \lambda} = \frac{\lambda}{\sigma_i^2 + \lambda}$.

Problem 3 (continued)

(c) Conclude that (8 points)

$$\|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\theta}}_\lambda\|_2^2 = \sum_{i=1}^{r} \Big(\frac{\lambda}{\sigma_i^2 + \lambda}\Big)^2 (\mathbf{u}_i^\top \mathbf{y})^2 + \|\mathbf{y} - \mathbf{P}\mathbf{y}\|_2^2,$$

and that the training error never decreases as λ grows.

Solution. The vectors $\mathbf{u}_1, \ldots, \mathbf{u}_r$ are orthonormal, and $\mathbf{y} - \mathbf{P}\mathbf{y} \perp \mathcal{C}(\mathbf{X})$, so $\mathbf{u}_i^\top(\mathbf{y} - \mathbf{P}\mathbf{y}) = 0$ for every i. So the $r + 1$ terms in part (b) are mutually orthogonal, and the Pythagorean theorem gives the formula. Each $\lambda/(\sigma_i^2 + \lambda) = 1 - \sigma_i^2/(\sigma_i^2 + \lambda)$ is positive and increases with λ, so each term never decreases, and neither does the sum.

Problem 3 (continued)

(d) Find the limits of the training error as $\lambda \to 0^+$ and as $\lambda \to \infty$, and interpret each. (6 points)

Solution. As $\lambda \to 0^+$, every factor $\lambda/(\sigma_i^2 + \lambda) \to 0$, so the training error tends to $\|\mathbf{y} - \mathbf{P}\mathbf{y}\|_2^2$, the least squares residual. As $\lambda \to \infty$, every factor tends to 1, so the training error tends to

$$\sum_{i=1}^{r} (\mathbf{u}_i^\top \mathbf{y})^2 + \|\mathbf{y} - \mathbf{P}\mathbf{y}\|_2^2 = \|\mathbf{P}\mathbf{y}\|_2^2 + \|\mathbf{y} - \mathbf{P}\mathbf{y}\|_2^2 = \|\mathbf{y}\|_2^2,$$

the error of the model $\boldsymbol{\theta} = \mathbf{0}$. Minimizing training error therefore always picks least squares, since least squares is *defined* as the minimizer of the training error; judging λ honestly requires data not used in the fit.

4. Lloyd's algorithm finds well-separated groups (Homework 06, Problem 1) **25 points**

Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ and $[n] = \{1, \dots, n\}$. 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.

Lloyd's algorithm. Input: data $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ and initial centers $\boldsymbol{\mu}_1, \dots, \boldsymbol{\mu}_k$. Repeat:

1. *Assignment step.* Assign each point to a nearest center: $c(i) \in \arg\min_\ell \|\mathbf{x}_i - \boldsymbol{\mu}_\ell\|_2$. If the current label of i is among the nearest, keep it.
2. *Update step.* For each ℓ with $S_\ell = \{i : c(i) = \ell\}$ non-empty, set $\boldsymbol{\mu}_\ell = \bar{\mathbf{x}}_{S_\ell}$. (If S_ℓ is empty, leave $\boldsymbol{\mu}_\ell$ where it is.)

until the assignment c does not change.

Suppose the data consists of well-separated groups. Precisely, $[n]$ is split into disjoint *groups* $G_1, \dots, G_K$, and there are numbers $\delta, \Delta > 0$ such that

- *(tight groups)* $\|\mathbf{x}_i - \mathbf{x}_j\|_2 \le \delta$ whenever i and j are in the same group, and
- *(separated groups)* $\|\mathbf{x}_i - \mathbf{x}_j\|_2 \ge \Delta$ whenever i and j are in different groups.

Suppose further *that $\Delta > 2\delta$, and that Lloyd's algorithm with $k = K$ starts from K data points as centers, one from each group.* In this problem you will show that Lloyd's algorithm then finds the groups exactly. *Hint:* Remember the triangle inequality, $\|\mathbf{x} + \mathbf{y}\|_2 \le \|\mathbf{x}\|_2 + \|\mathbf{y}\|_2$.

(a) Show that every point of a group is within distance δ of the mean of that group. (6 points)

Solution. Let $i \in G_a$ and let $m = |G_a|$. Since $\bar{\mathbf{x}}_{G_a} - \mathbf{x}_i = \frac{1}{m} \sum_{j \in G_a} (\mathbf{x}_j - \mathbf{x}_i)$, the triangle inequality gives

$$\|\bar{\mathbf{x}}_{G_a} - \mathbf{x}_i\|_2 \le \frac{1}{m} \sum_{j \in G_a} \|\mathbf{x}_j - \mathbf{x}_i\|_2 \le \frac{1}{m} \cdot m\delta = \delta.$$

Problem 4 (continued)

(b) Show that the first assignment step makes the clusters exactly the groups. (7 points)

Solution. Let $\mathbf{s}_a$ be the starting center in G_a, and let $i \in G_a$. Then $\|\mathbf{x}_i - \mathbf{s}_a\|_2 \le \delta < \Delta \le \|\mathbf{x}_i - \mathbf{s}_b\|_2$ for every $b \neq a$, so $\mathbf{x}_i$ is assigned to the cluster of $\mathbf{s}_a$, with no ties. The clusters are therefore exactly the groups, and the update step moves each center to its group mean $\bar{\mathbf{x}}_{G_a}$.

Problem 4 (continued)

(c) Show that the second assignment step changes nothing, so the algorithm stops with the groups as its clusters. (7 points)

Solution. Let $i \in G_a$. By part (a), $\mathbf{x}_i$ is within δ of its own group mean. For $b \neq a$, pick any $j \in G_b$. The triangle inequality and part (a) again give

$$\Delta \leq \|\mathbf{x}_i - \mathbf{x}_j\|_2 \leq \|\mathbf{x}_i - \bar{\mathbf{x}}_{G_b}\|_2 + \|\bar{\mathbf{x}}_{G_b} - \mathbf{x}_j\|_2 \leq \|\mathbf{x}_i - \bar{\mathbf{x}}_{G_b}\|_2 + \delta,$$

so $\|\mathbf{x}_i - \bar{\mathbf{x}}_{G_b}\|_2 \geq \Delta - \delta > \delta$, using $\Delta > 2\delta$. Every point is therefore still assigned to its own group, the assignment does not change, and the algorithm stops.

Problem 4 (continued)

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

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

(d) Show that the K-means cost of the returned clusters and centers is at most $n\delta^2$. (5 points)

Solution. By parts (b) and (c), the algorithm returns the clusters $G_1, \dots, G_K$ with centers the group means $\bar{\mathbf{x}}_{G_1}, \dots, \bar{\mathbf{x}}_{G_K}$. By part (a), each term $\|\mathbf{x}_i - \bar{\mathbf{x}}_{G_a}\|_2^2$ with $i \in G_a$ is at most δ^2. There are n terms in total, so

$$\operatorname{cost}(G_1, \dots, G_K; \bar{\mathbf{x}}_{G_1}, \dots, \bar{\mathbf{x}}_{G_K}) = \sum_{a=1}^{K} \sum_{i \in G_a} \|\mathbf{x}_i - \bar{\mathbf{x}}_{G_a}\|_2^2 \le n\delta^2.$$