National Sun Yat-sen University
Department of Applied Mathematics

Ph.D. Qualifying Exam
Statistical and Machine Learning Methods

September 3, 2026 (Thursday), 1:00–5:00 p.m.

INSTRUCTIONS:

- You have **240 minutes** to complete this exam which consists of **10 questions**.
- Please show all your steps to receive partial credit.
- When a conclusion or choice is requested, identify the relevant mechanism, assumption, diagnostic, or limitation. Concise, well-justified answers are sufficient; unsupported answers receive no credit.
- Please organize your work in a reasonably neat and coherent way.
- Please label the answer clearly.
- Common notation: i.i.d., independent and identically distributed; pdf, probability density function; RSS, residual sum of squares; CV, cross-validation; OOB, out-of-bag; PCA, principal component analysis; LDA, linear discriminant analysis. All vectors are column vectors. For a matrix $\mathbf{A}$, $\mathbf{A}^\mathsf{T}$ is its transpose, $\mathrm{tr}(\mathbf{A})$ is its trace, and $\mathbf{I}_d$ is the $d \times d$ identity matrix. The symbols E, Var and Cov denote expectation, variance and covariance. Also, $I\{A\}$ is one when the statement A is true and zero otherwise, $[u]_+ = \max(u, 0)$, $\mathrm{sign}(t)$ is 1 for $t \geq 0$ and -1 otherwise, $\mathbf{0}$ (or $\mathbf{0}_d$) denotes a zero vector, and $\|\cdot\|$ denotes the Euclidean norm. Any additional notation is defined in the question where it is used.

Question	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Points	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**10**	**100**
Score											

Question 1. Consider the linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \mathbf{X} = \begin{pmatrix} 1 & -1 \\ 1 & 0 \\ 1 & 1 \end{pmatrix}, \qquad \mathbf{y} = \begin{pmatrix} 1 \\ 2 \\ 5 \end{pmatrix},$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2)^{\mathsf{T}}$, with intercept β_1 and slope β_2. Here $\mathbf{y} \in \mathbb{R}^3$ collects the three observed responses, $\mathbf{X}$ is the displayed 3×2 design matrix whose second column holds the predictor values $-1, 0, 1$, and $\boldsymbol{\varepsilon} \in \mathbb{R}^3$ is the random-error vector. Assume $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}) = 4\mathbf{I}_3$.

(a) **[4 points]** Fit the least-squares line by minimizing $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$. Show the two normal equations that result, and report the fitted intercept and slope. Do not quote a closed-form least-squares estimator without showing how the equations are obtained.

(b) **[4 points]** Find a vector $\mathbf{c}_0 \in \mathbb{R}^3$ such that the fitted slope can be written $\hat{\beta}_2 = \mathbf{c}_0^{\mathsf{T}}\mathbf{y}$. Another slope estimator is $\widetilde{\theta} = \mathbf{c}^{\mathsf{T}}\mathbf{y}$, where $\mathbf{c} = (0, -1, 1)^{\mathsf{T}}$. Verify from $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$ that both estimators are unbiased for β_2.

(c) **[2 points]** Compute the variances of the two slope estimators. Which estimator is more precise? Your variance calculation expresses each variance through the estimator's weight vector; state which quantity in that expression differs between the two weight vectors, and explain why a larger value of that quantity makes the estimator noisier.

Question 2. Consider linear regression of a centered response vector $\mathbf{y} \in \mathbb{R}^N$ on an $N \times p$ design matrix $\mathbf{X}$ whose columns, one per predictor, are centered and standardized, with coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$. Ridge, lasso, and the elastic net estimate $\boldsymbol{\beta}$ by minimizing, respectively,

$$\begin{aligned} &\text{ridge:} && \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2\|\boldsymbol{\beta}\|^2, \\ &\text{lasso:} && \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda \textstyle\sum_{j=1}^{p} |\beta_j|, \\ &\text{elastic net:} && \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda_2\|\boldsymbol{\beta}\|^2 + 2\lambda \textstyle\sum_{j=1}^{p} |\beta_j|, \end{aligned}$$

with nonnegative tuning parameters λ_2 and λ. In the ridge and elastic-net fits compared in parts (b) and (c), the quadratic-penalty parameter λ_2 is positive. This question compares the structure and use of the three methods.

(a) **[4 points]** Suppose the columns of $\mathbf{X}$ are orthonormal, so that $\mathbf{X}^\mathsf{T}\mathbf{X} = \mathbf{I}_p$ and the least-squares coefficient of the jth predictor is $\hat{\beta}_j = \mathbf{x}_j^\mathsf{T}\mathbf{y}$, where $\mathbf{x}_j$ is the jth column of $\mathbf{X}$. In this case the lasso criterion above separates across coordinates: dropping the terms that do not involve the jth coefficient, coordinate j of the lasso problem is to minimize, over the scalar β_j,

$$q_j(\beta_j) = (\beta_j - \hat{\beta}_j)^2 + 2\lambda|\beta_j|, \qquad \lambda > 0.$$

By considering $\beta_j > 0$, $\beta_j < 0$, and $\beta_j = 0$, derive the soft-threshold rule for the minimizing coefficient. Then explain what it is about this rule that can set a coefficient exactly to zero, rather than merely shrinking it towards zero.

(b) **[3 points]** Two standardized predictors have correlation 0.98 and carry nearly interchangeable signal. Across three small perturbations of the training sample, with each method's tuning parameters held fixed, their fitted coefficient pairs are

	1	2	3
ridge	$(0.51, 0.49)$	$(0.47, 0.53)$	$(0.50, 0.50)$
lasso	$(1.00, 0)$	$(0, 0.99)$	$(0.98, 0)$
elastic net	$(0.54, 0.46)$	$(0.48, 0.52)$	$(0.52, 0.48)$

Using the table, compare the three methods on three points: whether a method sets a coefficient exactly to zero, how much each coefficient pair moves across the three perturbations, and whether the two correlated predictors end up sharing the signal or only one of them is kept. Then explain the mechanism behind the difference: why can small perturbations flip which variable the lasso selects here, while the ridge coefficients move only slightly?

(c) **[3 points]** Their tenfold cross-validation summaries are

	ridge	lasso	elastic net
error	1.01	1.00	0.99
standard error	0.06	0.07	0.06

Treat all methods within one standard error of the minimum as having essentially the same estimated prediction error. Suppose first that the goal is stable prediction and both correlated predictors may be kept: which methods are reasonable choices? Suppose instead that the reported model must keep only one member of the correlated pair: which method is then a reasonable choice? Finally, explain why this table does not show that the variable the lasso keeps is the uniquely important one.

Question 3. A centered two-variable data set has covariance matrix

$$\mathbf{S} = \frac{1}{5}\begin{pmatrix} 21 & 8 \\ 8 & 9 \end{pmatrix}.$$

The two eigenvalues are 5 and 1. A principal-component loading is required to have unit length. For a centered observation $\mathbf{x}$, define its score along a unit loading $\mathbf{v}$ by $z = \mathbf{v}^\mathsf{T}\mathbf{x}$ and its one-component reconstruction by $\hat{\mathbf{x}} = z\mathbf{v}$.

(a) **[4 points]** Find a unit loading vector for the first principal component. For $\mathbf{x} = (3, 1)^\mathsf{T}$, compute its first score, its one-component reconstruction, and the proportion of total variance explained (PVE) by the first component. The loading may point either way; just make sure the score you report uses the same direction as the loading you chose.

(b) **[4 points]** For any unit vector $\mathbf{v}$, prove that

$$\|\mathbf{x} - \mathbf{v}\mathbf{v}^\mathsf{T}\mathbf{x}\|^2 = \|\mathbf{x}\|^2 - (\mathbf{v}^\mathsf{T}\mathbf{x})^2.$$

Use this result to compute the squared reconstruction error of the displayed observation. More generally, let $\mathbf{X}_0$ be a centered random vector with covariance matrix $\mathbf{S}$. Prove that its expected one-component reconstruction error along $\mathbf{v}$ is $\mathrm{tr}(\mathbf{S}) - \mathbf{v}^\mathsf{T}\mathbf{S}\mathbf{v}$. Hence explain why, among all unit vectors $\mathbf{v}$, the leading eigenvector makes this expected reconstruction error smallest, and report the resulting minimum expected reconstruction error for the covariance matrix in this question.

(c) **[2 points]** Explain why a high proportion of predictor variance explained does not by itself guarantee good prediction of an external response variable.

Question 4. Three candidate bandwidths are considered for the same local polynomial smoother. Tenfold cross-validation evaluates the candidates, giving

candidate	training RSS	CV error (SE)	tr($\mathbf{S}$)
A	18	5.2 (0.4)	3
B	12	4.8 (0.5)	6
C	7	5.0 (0.6)	10

Here training RSS is the residual sum of squares on the training data, SE is the standard error of the reported CV error, and tr($\mathbf{S}$) is the trace of the fitted smoother matrix $\mathbf{S}$; a larger value of the trace means a more flexible fit. The one-standard-error rule uses the reported standard error of the minimum-CV candidate and chooses the least flexible candidate whose CV error is no more than that amount above the minimum CV error.

(a) **[4 points]** Which candidate does minimum CV select? Which one does the one-standard-error rule select? For the second rule, write down the numerical cutoff you used, namely the minimum CV error plus the standard error reported for that candidate, say which candidates fall on or below the cutoff, and explain why your pick is the one the rule chooses. Training RSS would point to yet another candidate: explain what training RSS measures that makes it rank the candidates differently from CV error.

(b) **[3 points]** Explain the bias–variance tradeoff as bandwidth decreases and tr($\mathbf{S}$) increases, and relate it to the table.

(c) **[3 points]** Suppose that before fitting the smoother, one uses the response values to select a predictor from a larger candidate set. Explain why this supervised selection must be repeated using only the training portion of each fold, rather than performed once on the full data before cross-validation. State how cross-validation evaluates the complete foldwise predictor-selection-and-smoothing procedure and selects a bandwidth. Then state how that procedure is rerun to obtain the final model.

Question 5. A two-term additive model is fitted by backfitting. Updating term j means applying its fixed smoother to the partial residual obtained after subtracting the current fit of the other term. Two update orders, starting from the same zero component fits, give the following traces of the residual sum of squares (RSS):

update order	sweep 1	sweep 2	converged
$1 \to 2$	1.20	0.41	0.30
$2 \to 1$	0.95	0.46	0.30

At the final reported sweep, the maximum change in either fitted component is below 10^{-4} for both orders.

(a) **[5 points]** Answer all four parts below. (i) Explain why the lower RSS of order $2 \to 1$ after one sweep does not show that this order defines a better statistical model. (ii) State what the table tells you about the intermediate iterates, about the limiting RSS, and about a reasonable rule for stopping the iteration. (iii) Does the table by itself prove that the two orders end at the same fitted vector? (iv) Does the table establish which order converges faster? For (iii) and (iv), give the reason for your answer.

(b) **[3 points]** Explain what a partial residual represents and why the second update uses the newly computed first component rather than its starting value.

(c) **[2 points]** In a standard additive model with an intercept, explain why each component is usually centered to have sample mean zero. What ambiguity does this convention remove?

Question 6. In one boosting iteration, a regression tree's terminal region is held fixed and contains the current residuals

$$(z_1, \ldots, z_5) = (-4, 0, 1, 1, 2).$$

The terminal-node update fits a common location c to these residuals by minimizing $\sum_{i=1}^{5} \rho(z_i - c)$ for a loss function ρ. For Huber loss with $\delta = 1$, define

$$\rho_1(u) = \begin{cases} u^2/2, & |u| \le 1, \\ |u| - 1/2, & |u| > 1, \end{cases} \qquad \psi_1(u) = \max(-1, \min(u, 1)).$$

Let $S(c) = \sum_i \psi_1(z_i - c)$. You may use that S is continuous and nonincreasing and that the derivative of $\sum_i \rho_1(z_i - c)$ is $-S(c)$.

(a) **[2 points]** Find the squared-error and absolute-error terminal node updates, that is, the value of c minimizing the displayed sum when $\rho(u) = u^2$ and when $\rho(u) = |u|$. No minimized objective values are required.

(b) **[4 points]** Compute $S(0)$ and $S(1)$. Then fix any c with $0 \le c \le 1$ and evaluate $\psi_1(z_i - c)$ for each of the five residuals in turn: for each one, decide whether $z_i - c$ lies inside $[-1, 1]$, so that $\psi_1(z_i - c) = z_i - c$, or outside it, so that the value is clipped to -1 or $+1$. Add the five results to obtain $S(c)$ as a function of c on this interval, solve $S(c) = 0$, and then use the properties supplied above to argue that this root is the unique global minimizer of the Huber criterion.

(c) **[4 points]** Replace the residual -4 by $-M$, where $M > 2$, leaving the other four residuals fixed. For each of the three updates (squared-error, absolute-error, Huber), say whether it keeps moving as M grows or settles at a fixed value, and give the value or expression it takes. Now suppose you want an update with two properties: one very large negative residual must not be able to drag it without limit, yet the update should still respond to the sizes of the ordinary residuals, which the absolute-error update ignores. Which of the three updates has both properties? Justify your answer using the behaviour you have just described.

Question 7. At a fixed predictor value, a fitted linear smoother and a fitted regression tree both predict 4.0 from the observed training data. Their nonparametric-bootstrap prediction summaries computed from a finite collection of refits, conditional on those observed data, are

fitting method	bootstrap mean	bootstrap SD
fixed linear smoother	4.0	0.15
regression tree	3.7	1.60

Here the bootstrap SD is the standard deviation of the refitted predictions across those refits. Let E_* and Var_* refer only to bootstrap randomness while the observed data are held fixed.

(a) **[4 points]** The ideal bagged prediction averages over the whole bootstrap distribution; here only a finite number of refits is available, so use each method's bootstrap mean from the table as the estimate of its ideal bagged prediction. Write down the resulting estimate for the smoother and the estimate for the tree. Then state which of the two methods is more unstable when the data are resampled, and for which of them averaging the refits is more worth doing. Base the comparison on the spread reported in the table, not on the fact that the two original fitted values are equal.

(b) **[3 points]** For a general fixed sample $x_1, \ldots, x_n$, let $X_1^*, \ldots, X_n^*$ be independent draws, each uniformly distributed over the observed values. Derive and simplify the conditional expectation of $\bar{X}^* = n^{-1} \sum_{j=1}^{n} X_j^*$, expressing your answer in terms of $\bar{x} = n^{-1} \sum_{i=1}^{n} x_i$. State what your result implies about the ideal bagged version of the sample mean.

(c) **[3 points]** Explain why bagging can reduce the variance of an unstable fitted rule but does not generally remove its systematic bias.

Question 8. Five trees produce the predictions below for four binary-labeled training observations. A dash means that the observation was included in that tree's bootstrap training sample, so that tree is not allowed to cast an out-of-bag (OOB) vote for that observation.

i	y_i	T_1	T_2	T_3	T_4	T_5
1	0	0	1	–	0	0
2	1	–	1	0	1	–
3	1	1	–	1	0	–
4	0	–	0	1	–	1

Use ordinary majority vote; no row has a tie.

(a) **[4 points]** For each observation, identify its available OOB votes, compute its OOB prediction, and determine whether it is misclassified. Then compute the overall OOB error rate.

(b) **[3 points]** A tree is trained using N independent bootstrap draws from N observations. Derive the probability that a fixed observation is OOB for one tree. For $N = 4$ and $B = 5$ independent trees, compute the probability that the observation receives no OOB vote at all. You may use $(175/256)^5 \approx 0.15$.

(c) **[3 points]** Explain why OOB error is a prediction-error estimate similar to cross-validation rather than training error. If in-bag trees were mistakenly allowed to vote as well, would the resulting error estimate be biased too high or too low, and why?

Question 9. Four one-dimensional training observations are

$$x = (-2, -1, 1, 2), \qquad y = (-1, -1, 1, 1).$$

A candidate support vector classifier $f_b(x) = bx$ classifies an observation to the class $\text{sign}(f_b(x))$. Define the signed margin $m_i = y_i f_b(x_i)$, the hinge loss $[1 - m_i]_+$, and the comparison criterion

$$J_C(b) = \frac{1}{2}b^2 + C\sum_{i=1}^{4}[1 - y_i f_b(x_i)]_+, \qquad C > 0,$$

which is the support-vector objective evaluated at the candidate. An observation is on the unit margin when $m_i = 1$, inside it when $m_i < 1$, and strictly outside it when $m_i > 1$.

(a) **[6 points]** For each candidate $b = 0.5$ and $b = 1$, compute all signed margins and hinge losses. Derive $J_C(b)$ as a function of C for each candidate, find the value of C at which the two criterion values are equal, and state which candidate has the smaller J_C for C below that value and which has the smaller J_C for C above it.

(b) **[4 points]** For each candidate, identify observations that are on, inside, or strictly outside the unit margin and state whether any are misclassified. Which observations could be deleted from the training sample without changing that candidate's value of J_C, and why?

Question 10. Consider the one-dimensional observations

$$0, \qquad 2, \qquad 8, \qquad 10$$

and two current centers $\mu_1 = 0$, $\mu_2 = 6$. In a two-component spherical Gaussian mixture, hold the current centers and positive mixing proportions fixed while taking the E-step limit as the common variance tends to zero, with no distance ties. You may use the fact that the E-step then approaches nearest-center assignment and the subsequent M-step replaces each center by the mean of its assigned observations.

(a) **[3 points]** Perform one K-means assignment-and-update step: assign every observation to its nearest current center and compute the two updated centers. Then verify that, with the updated centers, a further assignment step leaves every assignment unchanged, and state what this implies about the algorithm on this data set.

(b) **[4 points]** For any nonempty one-dimensional cluster C, let $n = |C|$ and $\bar{x}_C = n^{-1} \sum_{x_i \in C} x_i$. Prove that for every candidate center c,

$$\sum_{x_i \in C} (x_i - c)^2 = \sum_{x_i \in C} (x_i - \bar{x}_C)^2 + n(\bar{x}_C - c)^2,$$

and deduce that the cluster mean is the unique center minimizing the cluster's sum of squared errors, that is, the sum of squared distances from its observations to the center.

(c) **[3 points]** Explain why the sum of squared errors decreasing at every K-means step does not guarantee the globally best clustering. State how initialization affects the result and describe the role of multiple starts in selecting a reported solution.