STA 35C Statistical Data Science III

Practice final exam A

Instructor: Dogyoon Song

Name: _____ **Student ID:** _____

Instructions: This practice final exam is **closed-book**, except for the permitted note sheet described below. You may bring a pen or pencil, three letter-sized sheets of *hand-written* notes (both sides), and a *non-graphing* calculator. No other materials (e.g., textbooks) are allowed. For a timed attempt, allow 120 minutes. Problems 1–6 total **180 points**; Problem 7 offers *up to 10 bonus points*. Once you receive this exam, please confirm that you have all 10 pages.

- Make sure to clearly write your name and student ID above.
- For Problem 1, give the requested answers. Include a one-sentence explanation in parts (d)–(f) and (j)–(l). For Problems 2–7, show the relevant steps for partial credit.
- If necessary, round numerical answers to three decimal places; fractions, logarithms, and exponentials may be left unevaluated.
- Complete Problems 1–6 before attempting the optional bonus questions in Problem 7.

Problem	**Score**
Problem 1	
Problem 2	
Problem 3	
Problem 4	
Problem 5	
Problem 6	
Problem 7 (bonus)	
Total	

Problem 1 (60 points). Basic concepts

(a) (5 points) Forty observations are split into five equal-sized folds for cross-validation of one candidate model. **Circle all correct statements** about this CV run.

A. Five model fits produce all the held-out predictions.
B. Each fit uses 32 training observations.
C. Each observation is used for validation four times.
D. Each observation belongs to the training set for four of the fits.

(b) (5 points) Four hypotheses H_1, H_2, H_3, H_4 have p-values $0.031, 0.009, 0.043, 0.016$, respectively. **Circle all hypotheses** rejected by Holm's step-down procedure at familywise significance level $\alpha = 0.05$.

A. H_1
B. H_2
C. H_3
D. H_4

(c) (5 points) The four pairwise distances from points in cluster A to points in cluster B are $2, 4, 5, 7$. **Circle all correct statements** about the distance between these clusters.

A. Under single linkage it is 2.
B. Under complete linkage it is 4.5.
C. Under single linkage it is 5.
D. Under complete linkage it is 7.

(d) (5 points) At a fixed x, the fitted prediction is unbiased for $\mathbb{E}[Y \mid X = x]$. Averaging over both the training sample and an independent new response at x, the expected squared prediction error must equal the irreducible error.

True **False**

Reason:

(e) (5 points) A zero lasso slope removes that term from the fitted regression equation. The corresponding population slope could still be nonzero.

True **False**

Reason:

(f) (5 points) In a smoothing-spline objective, increasing the curvature-penalty multiplier favors curves with more curvature when training RSS is the same.

True **False**

Reason:

(g) (5 points) For a line with an intercept, $\bar{x} = 4$, $\bar{y} = 10$, $S_{xx} = 8$, and $S_{xy} = 12$ (centered sums). Find the least-squares line and its prediction at $x = 6$.

(h) (5 points) A logistic model has slope $\log 2$ in the log-odds equation. The fitted class-1 odds at $x = 1$ are 3. What are the fitted odds at $x = 2$?

(i) (5 points) An observation has centered coordinates $(2, 1)$. Find its PC score along the unit loading vector $(3/5, 4/5)$.

(j) (5 points) Exam score is modeled by $\widehat{y} = 50 + 3h + 2q$, with h study hours and q completed quizzes. Interpret the coefficient of h in one sentence.

(k) (5 points) For centered data, describe in one sentence the quantity PCA maximizes to find PC1 and the constraint on its direction.

(l) (5 points) Two K-means runs on the same data use the same K. Explain in one sentence why different initial centers can give different final partitions.

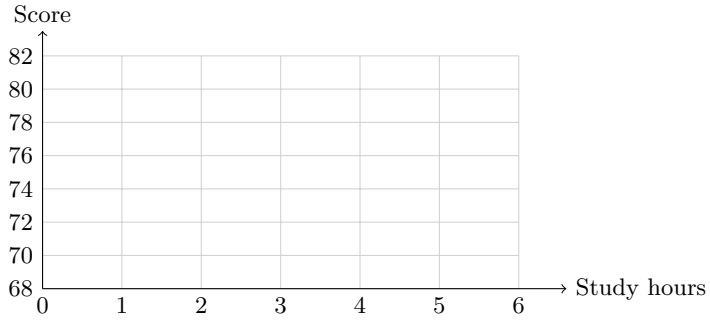
Problem 2 (20 points). Regression and classification

The two parts use separate settings.

(a) (10 points) The fitted exam scores for students studying x hours are $70 + 2x$ in Section A and $73 + x$ in Section B.

(i) Give each section's fitted slope in score points per study hour.

(ii) Draw Section A's fitted line for $0 \le x \le 6$. Mark and label the predicted score at $x = 4$.



(b) (10 points) At a factory, 20% of parts are defective. An alarm sounds for 80% of defective parts and 10% of nondefective parts. A sounding alarm leads to a defective classification if the posterior defect probability is at least 0.60.

(i) Find the defect probability for a part that triggers the alarm. How is it classified?

(ii) A second factory has the same alarm rates within each class. What is the smallest defect prevalence for which a sounding alarm would lead to a defective classification there?

Problem 3 (20 points). Model selection and the bootstrap

(a) (8 points) A fixed model's absolute prediction errors on five i.i.d. test cases are $1, 2, 3, 4, 8$. Each row gives the sorted errors from an independent bootstrap resample of five test cases drawn with replacement.

Resample	Absolute errors				
1	1	1	2	4	8
2	1	2	3	3	8
3	2	3	4	4	8

(i) Give the original median error and the three resample medians.

(ii) Let s be the sample standard deviation of the three resample medians (use divisor 2). Compute s. Which estimates the standard error of the original sample median, s or $s/\sqrt{3}$? Explain briefly.

(b) (12 points) The training cases are independent draws from one population. The table gives training RSS for least-squares models with an intercept. Model ij uses X_i, X_j.

Model	1	2	3	4	12	13	14	23	24	34
Training RSS	38	42	60	84	26	35	30	10	34	52

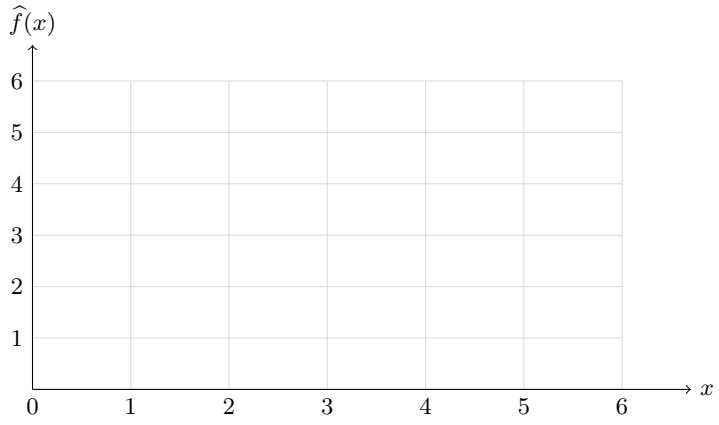
(i) Which pair does forward selection reach from the intercept-only model? Which pair does best subset selection choose? Identify the pairs eligible at step 2 of forward selection to explain the difference.

(ii) An analyst bootstraps only the test cases, holding the selected model fixed. Can this measure variability in predictions arising from selection and fitting on new training samples? For a case bootstrap that studies this variability at a fixed input, what should the analyst resample and repeat?

Problem 4 (20 points). Regression splines and smoothing splines

(a) (10 points) A fitted regression spline is $\widehat{f}(x) = 2 + x - 2(x-3)_+$, where $(u)_+ = \max(u, 0)$.

(i) Sketch the fit for $0 \le x \le 6$ and label its values at $x = 0, 3, 6$.



(ii) Consider the alternative fit $\widehat{g}(x) = 2 + x - 2\,\mathbf{1}\{x > 3\}$. Give each fit's slopes below and above 3. Which fit has a jump at 3? Give its size and direction.

(b) (10 points) A smoothing spline minimizes $\text{RSS} + \lambda J(f)$, where $J(f) = \int [f''(x)]^2\, dx$. Two readings at $x = 1$ give responses 2 and 4. Single readings at $x = 2, 3$ give 7 and 6, respectively.

(i) As $\lambda \downarrow 0$, which fitted value does the spline approach at each input? What is the smallest possible training RSS?

(ii) An analyst keeps only the three input-specific means. What weights should their squared residuals receive to reproduce the original fit for every λ, using the same penalty $\lambda J(f)$? Justify your answer by expanding $(2-t)^2 + (4-t)^2$ for $t = f(1)$.

Problem 5 (25 points). Principal components and clustering

The three parts use separate data sets.

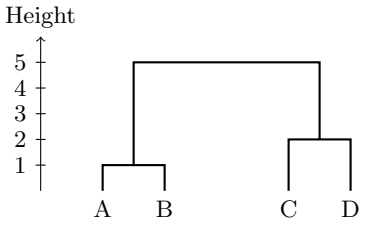
(a) (5 points) The three PC directions u, v, w are listed out of order. The sample standard deviations of their projected scores are $2, 1, 3$, respectively. Give the three directional variances, identify PC1, and compute the proportion of total variance explained by the first two PCs together.

(b) (10 points) In one dimension the current K-means clusters are $\{0, 4, 8\}$ and $\{10, 11\}$.

(i) Compute the two centroids. Reassign each observation to its nearest centroid and give the new clusters.

(ii) Do the original centroids minimize the sum of squared distances for the new memberships? Justify your answer using the updated centroid of the group containing 8.

(c) (10 points) Use the dendrogram below. List the clusters at a cut height of 1.5. Between which two successive merge heights are there exactly two clusters, with A and D in different clusters?



Problem 6 (35 points). Calibrating a parcel scale

A shipping center uses a scale's displayed reading x to predict a parcel's certified mass y (both in kg). The calibration data are independent parcels with readings between 4 and 18 kg. The goal is to predict certified mass for new parcels from the same population.

(a) (5 points) At displayed readings of 6, 10, and 14 kg, the sample mean certified masses are 5.8, 9.4, and 14.0 kg, respectively. Compute the increase in mean certified mass over each successive 4-kg interval. Over which interval is the increase larger?

(b) (5 points) One candidate fit is $\widehat{y} = 0.4 + 0.9x + 0.3(x - 10)_+$, where $(u)_+ = \max(u, 0)$.

(i) Compute the fitted mass at a reading of $x = 16$ kg.

(ii) Above 10 kg, a worker expects each additional kilogram in the reading to add 0.3 kg to fitted mass. Is the worker correct? Interpret 0.3 using the slopes on both sides of the knot.

(c) (10 points) Three models are compared on the same five equal-sized validation folds. Training MSE is computed after fitting each model to all calibration data.

Model	Training MSE	CV MSE
Straight line	1.0	1.1
One-knot spline	0.5	?
More flexible spline	0.3	0.9

The one-knot spline has fold MSEs $0.4, 0.5, 0.6, 0.7, 0.8$.

(i) Compute the missing CV MSE.

(ii) Which model does CV select, and which does training MSE select? Explain why the winning CV MSE can be optimistic and how to assess the selected model independently.

(d) (15 points) After model selection, the center refits its model on independent calibration data. Each rule releases a parcel if the upper endpoint of its designated interval is at most 13.5 kg.

Rule	Designated 95% interval	Interval at $x = 13$ kg
Center	Prediction interval for one parcel's mass	$[12.0, 14.0]$ kg
Manager	Confidence interval for mean mass	$[12.8, 13.2]$ kg

Citing the model's lowest CV MSE, the manager claims that at most 5% of the parcels her rule releases will exceed 13.5 kg.

(i) For the parcel with reading 13 kg, give the decision under each rule.

(ii) Do the mean confidence interval and the CV MSE justify the manager's 5% claim? Explain what each measures.

(iii) A new random audit of 200 independent parcels finds 6 overweight among the 80 released by the manager's fixed rule and 18 overweight among those held. Estimate the overweight fraction among releases. A one-sided test of the claim against a rate above 5% gives $p = 0.211$. At level 0.05, does the audit establish that the population rate among releases exceeds the limit, that it meets the limit, or neither? Explain.

Problem 7 (bonus). Two optional questions

The two parts are independent from each other.

(a) (5 points) Two true null hypotheses each have probability 0.05 of being rejected. Without assuming independence, find the smallest and largest possible probabilities of at least one false rejection. Describe a relationship between the rejection events that attains each bound.

(b) (5 points) Four observations are the corners $(\pm 1, \pm 1)$ of a square. Multiply every horizontal coordinate by $s > 0$. For $K = 2$, compare the left/right partition with the upper/lower partition. For which s does each have smaller within-cluster sum of squares? Give the two costs and any tie. If each rescaled coordinate is then standardized to unit sample variance, how does your comparison change?