STA 35C Statistical Data Science III

Final Exam

Instructor: Dogyoon Song

Name: _____ **Student ID:** _____

Instructions: This is a **closed-book final exam**. 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 are allowed. **You have 120 minutes** to complete the exam. The **total score is 180 points**, with **up to 8 bonus points**. Once you receive this exam problem set, please confirm that your copy includes all pages.

- Make sure to clearly write your name and ID above.
- Present answers succinctly, but include all relevant steps for full credit.
- Partial credit is possible only if your reasoning is clearly presented and traceable.
- If necessary, round numerical answers to three decimal places. You may leave fractions, logarithms, and exponentials unevaluated when appropriate.

Problem	**Score**
Problem 1	
Problem 2	
Problem 3	
Problem 4	
Problem 5	
Problem 6	
Total	

Problem 1 (30 points total).

In each subproblem, **check all** boxes in front of statements you believe are **correct**. Each subproblem is worth 3 points, and you receive the 3 points only if you check *all* correct options and no incorrect options.

In each subproblem, there may be one, two, or three correct answers, but not all or none.

(a) Training vs. test error

- ☐ Training error is computed on a held-out validation set, not on the data used to fit the model.
- ☐ Test error measures performance on new, unseen data.
- ☐ As model flexibility increases, training error usually increases.
- ☐ The model with the lowest training error always has the lowest test error.

(b) Cross-validation and bootstrap

- ☐ In k-fold CV, each fold is used once as validation data.
- ☐ LOOCV is the special case of k-fold CV with $k = n$.
- ☐ The bootstrap resamples observations with replacement to estimate uncertainty.
- ☐ Validation observations should be used to fit the model whose validation error is being evaluated.

(c) Linear regression and R^2

- ☐ In multiple regression, β_j is interpreted without holding other predictors fixed.
- ☐ Adding predictors always decreases test error.
- ☐ Standard errors quantify uncertainty in coefficient estimates.
- ☐ A large R^2 proves that the predictors cause the response.

(d) Classification thresholds

- ☐ A false positive occurs when $Y = 0$ but we predict $\hat{Y} = 1$.
- ☐ Lowering the threshold p^* usually increases the number of predicted positives.
- ☐ Lowering p^* usually decreases the false positive rate.
- ☐ Lowering p^* usually increases the true positive rate.

(e) Model selection and regularization

- ☐ Best subset selection fits all subsets of predictors of each size.
- ☐ Forward stepwise selection always finds the globally best subset.
- ☐ Ridge regression can set coefficients exactly to zero, like lasso.
- ☐ Lasso can perform variable selection by setting some coefficients to zero.

(f) Multiple testing

- ☐ FWER is the probability of making at least one false rejection.
- ☐ FDR is the probability of making at least one false discovery.
- ☐ Bonferroni is typically less conservative than Benjamini–Hochberg.
- ☐ FDR control guarantees that no false positives occur.

(g) Regression splines

- ☐ A cubic spline is a piecewise cubic polynomial.
- ☐ A cubic spline must be continuous up to its third derivative at knots.
- ☐ A natural cubic spline adds boundary constraints that make the tails linear.
- ☐ A natural cubic spline has more degrees of freedom than a cubic spline with the same knots.

(h) Smoothing splines

- ☐ With distinct x_i's, the effective degrees of freedom increase from 2 toward n as λ increases.
- ☐ When $\lambda = 0$, the smoothing spline becomes the least-squares line.
- ☐ Increasing λ usually makes the fitted curve more wiggly.
- ☐ A smoothing spline balances data fit and a curvature penalty.

(i) Principal component analysis

- ☐ PCA chooses directions that best predict a response Y.
- ☐ PCA is unsupervised and uses only X, not Y.
- ☐ The first PC direction maximizes variance of the projected data.
- ☐ Standardization never matters for PCA.

(j) Clustering

- ☐ In a dendrogram, horizontal spacing between leaves directly measures distance.
- ☐ Feature scaling has no effect on clustering results.
- ☐ K-means requires choosing K before fitting.
- ☐ Hierarchical clustering produces a dendrogram.

Problem 2 (20 points total).

(a) (7 points) Four regression models are fit to the same training data and evaluated on an independent test set:

Model	Flexibility	Training MSE	Test MSE
A	Very Low	46	48
B	Moderately low	30	28
C	Moderately high	12	36
D	Very high	4	60

Which model would you choose for prediction? Which model(s) appear to overfit? Briefly justify.

(b) (7 points) Explain the bias–variance tradeoff, including descriptions of how squared bias, variance, test error, and irreducible error typically behave as model flexibility increases.

(c) (6 points) Explain the difference between supervised and unsupervised learning. Give one example of each from the methods studied in this course, specifying the data format and the goal.

Problem 3 (40 points total + 4 bonus points).

(a) (6 points) A fitted regression model for exam score is

$$\widehat{Y} = 60 + 4X,$$

where X is hours studied. Compute $\widehat{Y}$ for $X = 8$, interpret the slope 4 in context, and briefly explain why the actual score may differ from $\widehat{Y}$.

(b) (10 points) Suppose four regression fits using predictors X_1, X_2, X_3, X_4 produce the following coefficient estimates and validation MSEs:

Fit	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	**Validation MSE**
OLS	7.8	−6.9	0.4	5.1	21
Fit A	4.2	−3.8	0.2	3.0	15
Fit B	5.0	0	0	2.0	13
Fit C	0	0	0	0	35

Using OLS as a reference, identify which of Fits A–C looks most like ridge regression and which looks most like lasso with a moderate shrinkage parameter λ. Which fit would you choose for prediction based on validation MSE? Briefly justify each answer.

(c) (12 points) Consider the fitted basis-function model

$$\widehat{Y} = 10 + 2X + 5I(X > 5) + 3(X - 8)_+,$$

where I is the indicator function and $(u)_+ = \max\{u, 0\}$. List all basis functions in this model, and write $\widehat{y}(x)$ piecewise on $0 \le x \le 5$, $5 < x \le 8$, and $8 < x \le 10$. Then draw a graph of the fitted function, and label $\widehat{y}(0)$, $\widehat{y}(5)$, $\widehat{y}(5^+)$, $\widehat{y}(8)$, and $\widehat{y}(10)$ on your graph.

(d) (12 points) Consider one-dimensional cubic splines with $K = 3$ (interior) knots.

(i) How many degrees of freedom does an ordinary cubic regression spline have?

(ii) What extra constraints does a natural cubic spline impose? What boundary behavior does it lead to, and why is this useful?

(iii) A smoothing spline estimates g by minimizing

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int \big(g''(t)\big)^2 \, dt.$$

Explain the roles of the two terms and describe the resulting regression functions in the two limits $\lambda \to 0$ and $\lambda \to \infty$, respectively.

(e*) (4 bonus points) Two students make the following claims about smoothing splines fit to data with distinct $x_1, \ldots, x_n$. Briefly explain what is wrong with each claim.

- Student A: "Since the smoothing spline has knots at all observed x_i's, it must interpolate the data."
- Student B: "Increasing λ removes knots one by one, so the smoothing spline becomes smoother because it has fewer knots."

Problem 4 (30 points total).

A bank wants to predict whether a customer will default. Let $Y = 1$ denote default and $Y = 0$ no default.

(a) (6 points) Suppose we first compute two principal component scores Z_1, Z_2 from a set of financial predictors, then fit logistic regression:

$$\log\left(\frac{\hat{p}(z)}{1-\hat{p}(z)}\right) = -1 + 2z_1 - z_2, \qquad \hat{p}(z) = \widehat{P}(Y = 1 \mid Z = z).$$

For $z_{\text{new}} = (1, 0.5)$, compute $\hat{p}(z_{\text{new}})$ and classify using $p^* = 0.5$. Write the decision boundary for $p^* = 0.5$, and state the region in the (z_1, z_2)-plane classified as default.

Hint: $e^{0.5} \approx 1.65$.

(b) (8 points) On a test set of 100 customers, the classifier gives the following confusion matrix:

	Predicted 1	Predicted 0
$Y = 1$	42	8
$Y = 0$	12	38

(i) Compute the false positive rate (FPR), false negative rate (FNR), and the total error rate.

(ii) For the same model and the same test set of 50 defaults and 50 non-defaults, suppose we obtain:

p^*	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
TP	50	48	47	45	42	38	35	25	20
FP	40	30	24	18	12	9	7	5	2

If the bank requires FPR ≤ 0.20, which threshold among these should it choose to maximize TPR?

(c) (10 points) Now suppose the bank uses only the first principal component score Z_1 as a one-dimensional risk score and fits an LDA classifier. Assume

$$Z_1 \mid (Y = 0) \sim N(-1, 1), \qquad Z_1 \mid (Y = 1) \sim N(2, 1),$$

with equal class priors. Here, $N(\mu, \sigma^2)$ represents the normal distribution with mean μ and variance σ^2.

(i) Find the LDA cutoff and classify a new customer with $z_1 = 0.8$.

(ii) Suppose Z_1 explains 90% of the variance in the original predictors. Does this high PVE alone justify discarding Z_2 when predicting default? Briefly explain, and state what evidence you would want to check before dropping Z_2.

(d) (6 points) Suppose the bank fits logistic regression using each possible subset of three candidate predictors Z_1, Z_2, Z_3. Each model is evaluated on the same held-out validation set:

Features	Train Acc.	Valid. Acc.	Valid. FNR	Features	Train Acc.	Valid. Acc.	Valid. FNR
$\emptyset$	0.60	0.62	1.00	Z_1, Z_2	0.86	0.84	0.18
Z_1	0.79	0.78	0.10	Z_1, Z_3	0.89	0.86	0.26
Z_2	0.76	0.75	0.08	Z_2, Z_3	0.84	0.81	0.16
Z_3	0.78	0.73	0.31	Z_1, Z_2, Z_3	0.96	0.80	0.34

Among the listed subsets, which would you choose if the only goal were to maximize validation accuracy? Which would you choose if the bank additionally requires FNR ≤ 0.15?

Problem 5 (20 points total).

(a) (6 points) A regression coefficient is estimated as $\hat{\beta} = 2.4$ with standard error 0.8. Compute the t-statistic for testing $H_0 : \beta = 0$. Using the rule $|t| > 2$ for approximate 5% significance, decide whether to reject H_0. Construct an approximate 95% confidence interval using 1.96 as the multiplier.

(b) (6 points) You have an estimate $\hat{\theta} = 10$ of a parameter θ. Five bootstrap estimates are

$$7, \quad 9, \quad 10, \quad 12, \quad 14.$$

Using a normal approximation, construct a 95% bootstrap confidence interval for θ, estimating the standard error by the sample standard deviation of the bootstrap estimates.

(c) (8 points) Suppose seven hypotheses have the following p-values, already sorted in increasing order:

$$H_{0,1} : 0.004, \quad H_{0,2} : 0.008, \quad H_{0,3} : 0.045, \quad H_{0,4} : 0.050, \quad H_{0,5} : 0.085, \quad H_{0,6} : 0.120, \quad H_{0,7} : 0.200.$$

At level $\alpha = 0.05$, which hypotheses are rejected by the Bonferroni method? At FDR level $q = 0.10$, which hypotheses are rejected by Benjamini–Hochberg? List the rejected hypotheses precisely.

Problem 6 (40 points total + 4 bonus points).

(a) (10 points) Consider the centered two-dimensional dataset

$$\mathcal{X} = \{(-3,-4), (-2,-1),\ (0,0),\ (2,1), (3,4)\}.$$

Compute the *directional variance* along each direction, using denominator n:

$$\mathbf{e}_1 = (1,0), \qquad \mathbf{e}_2 = (0,1), \qquad \mathbf{u} = \frac{1}{\sqrt{2}}(1,1).$$

Based on these three numbers, decide which of the three directions seems closest to the first principal component direction. Briefly justify.

(b) (10 points) PCA on a dataset produces six principal components with variances

$$26, \qquad 12, \qquad 6, \qquad 3, \qquad 2, \qquad 1.$$

Compute the proportion of variance explained by each PC. Compute the cumulative PVE for the first two PCs and determine the smallest number of PCs needed to explain at least 90% of the total variance.

(c) (10 points) Consider four points:

$$\mathbf{z}_1 = (0, 0), \quad \mathbf{z}_2 = (0, 1), \quad \mathbf{z}_3 = (4, 0), \quad \mathbf{z}_4 = (5, 1).$$

Using the K-means update-and-reassign algorithm, perform *two iterations* of K-means with $K = 2$ on $\{z_1, z_2, z_3, z_4\}$, assuming the initial cluster assignment:

$$C_1 = \{1, 3\}, \qquad C_2 = \{2, 4\}.$$

Show the centroids and updated cluster assignments at each iteration clearly.

(d) (10 points) Four observations A, B, C, D have the following pairwise distance matrix:

	A	B	C	D
A	0	1	5	6
B	1	0	4	5
C	5	4	0	2
D	6	5	2	0

Using *complete linkage*, draw a dendrogram, clearly marking the merge sequence and merge heights. To obtain two clusters, give a valid range of cut heights and identify the resulting clusters.

(e*) **(4 bonus points)** PCA keeps directions of largest variance, while clustering may depend on directions of separation between groups. Give a short example or explanation showing why reducing data to only the first PC before clustering can sometimes destroy meaningful cluster structure.