STA 35C Statistical Data Science III

Final Exam Solution

Instructor: Dogyoon Song

Problem 1

Correct boxes:

Part	Correct statements
(a)	2
(b)	1, 2, 3
(c)	3
(d)	1, 2, 4
(e)	1, 4
(f)	1
(g)	1, 3
(h)	4
(i)	2, 3
(j)	3, 4

Problem 2

(a) For prediction, choose Model B because it has the smallest test MSE:

$$28 < 48, \qquad 28 < 36, \qquad 28 < 60.$$

Models C and D appear to overfit: they have smaller training MSEs than Model B, but larger test MSEs. Model D appears to overfit most severely.

(b) As model flexibility increases, squared bias typically decreases because the model class can represent more complex relationships. Variance typically increases because the fitted model becomes more sensitive to the particular training sample. Test error often decreases at first and then increases, producing a U-shape. Irreducible error is noise in the data-generating process and does not change with model flexibility.

(c) In supervised learning, we observe both predictors and responses, (X, Y), and aim to predict or explain Y from X. Examples include linear regression and logistic regression.

In unsupervised learning, we observe only features X, with no response Y, and aim to discover structure in X. Examples include PCA and clustering.

Problem 3

(a) At $X = 8$,

$$\widehat{Y} = 60 + 4 \times 8 = 92.$$

The slope 4 means that, according to the fitted model, each additional hour studied is associated with an increase of 4 points in predicted exam score on average.

The actual score may differ because of random error, individual differences, unobserved variables, or because the linear model is only an approximation.

(b) Fit A looks most like ridge regression: compared with OLS, all coefficients are shrunk toward zero but none are exactly zero.

Fit B looks most like lasso with moderate shrinkage: some coefficients are exactly zero, so it performs variable selection.

Based on validation MSE, choose Fit B, since it has the smallest validation MSE:

$$13 < 15, \qquad 13 < 21, \qquad 13 < 35.$$

(c) The basis functions are

$$1, \qquad X, \qquad I(X > 5), \qquad (X - 8)_+.$$

The fitted function is

$$\widehat{y}(x) = \begin{cases} 10 + 2x, & 0 \le x \le 5, \\ 15 + 2x, & 5 < x \le 8, \\ 15 + 2x + 3(x - 8) = 5x - 9, & 8 < x \le 10. \end{cases}$$

The requested values are

$$\widehat{y}(0) = 10, \qquad \widehat{y}(5) = 20, \qquad \widehat{y}(5^+) = 25, \qquad \widehat{y}(8) = 31, \qquad \widehat{y}(10) = 41.$$

The graph is a line with slope 2 from 0 to 5, jumps upward by 5 just after $x = 5$, continues with slope 2 until $x = 8$, and then has slope 5 after $x = 8$. It is discontinuous at $x = 5$ and continuous at $x = 8$.

(d)(i) An ordinary cubic ($d = 3$) regression spline with $K = 3$ interior knots has

$$K + d + 1 = 3 + 3 + 1 = 7$$

degrees of freedom.

(ii) A natural cubic spline imposes additional boundary constraints, making the fitted function linear beyond the boundary knots. Equivalently, for cubic splines, the second derivative is constrained to be zero in the tails. This reduces unstable or erratic boundary behavior.

(iii) A smoothing spline minimizes

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

The first term measures data fit. The second term penalizes curvature or wiggliness. The parameter λ controls the tradeoff. When $\lambda \to 0$, the fit approaches interpolation. When $\lambda \to \infty$, the curvature penalty dominates, and the fit approaches the least-squares line.

(e*) Student A is wrong because having knots at all observed x_i's does not by itself imply interpolation. For $\lambda > 0$, the curvature penalty prevents the smoothing spline from necessarily passing through every point. Interpolation occurs in the limit $\lambda = 0$.

Student B is wrong because increasing λ does not remove knots. The smoothing-spline solution remains a natural cubic spline with knots at the observed x_i's, but the curvature penalty reduces the effective degrees of freedom and makes the fitted curve smoother.

Problem 4

(a) At $z_{\text{new}} = (1, 0.5)$, the log-odds are

$$-1 + 2(1) - 0.5 = 0.5.$$

Therefore

$$\hat{p}(z_{\text{new}}) = \frac{e^{0.5}}{1 + e^{0.5}} \approx \frac{1.65}{2.65} \approx 0.623.$$

Since $0.623 > 0.5$, classify the customer as default:

$$\hat{Y} = 1.$$

The decision boundary for $p^* = 0.5$ is

$$-1 + 2z_1 - z_2 = 0, \qquad \text{or} \qquad z_2 = 2z_1 - 1.$$

The default region is

$$-1 + 2z_1 - z_2 \geq 0, \qquad \text{equivalently} \qquad z_2 \leq 2z_1 - 1.$$

(b)(i) From the confusion matrix,

$$TP = 42, \quad FN = 8, \quad FP = 12, \quad TN = 38.$$

Therefore

$$\text{FPR} = \frac{FP}{FP + TN} = \frac{12}{50} = 0.24,$$

$$\text{FNR} = \frac{FN}{TP + FN} = \frac{8}{50} = 0.16,$$

and

$$\text{Error rate} = \frac{FP + FN}{100} = \frac{12 + 8}{100} = 0.20.$$

(ii) The requirement $\text{FPR} \leq 0.20$ means

$$\frac{FP}{50} \leq 0.20 \quad \Longleftrightarrow \quad FP \leq 10.$$

The thresholds satisfying this are $p^* = 0.6, 0.7, 0.8, 0.9$. Their TPs are $38, 35, 25, 20$, respectively. To maximize TPR, choose

$$p^* = 0.6.$$

(c)(i) Since the class priors and variances are equal, the LDA cutoff is the midpoint:

$$z^* = \frac{-1 + 2}{2} = 0.5.$$

Since $z_1 = 0.8 > 0.5$, classify the customer as $Y = 1$, default.

(ii) No. A high PVE means Z_1 explains a large fraction of variation in the predictors X, but PCA is unsupervised and does not use Y. The lower-variance direction Z_2 could still be important for predicting default. We would want to check validation/test performance, such as error rate, FNR, AUC, or another relevant classification metric, with and without Z_2.

(d) If the only goal is to maximize validation accuracy, choose

$$\{Z_1, Z_3\},$$

since it has the largest validation accuracy, 0.86.

Under the constraint FNR ≤ 0.15, the eligible models are only $\{Z_1\}$ and $\{Z_2\}$. Among these, the larger validation accuracy is achieved by $\{Z_1\}$, with validation accuracy 0.78 and FNR 0.10. Thus choose

$$\{Z_1\}.$$

Problem 5

(a) The test statistic is

$$t = \frac{\hat{\beta} - 0}{SE(\hat{\beta})} = \frac{2.4}{0.8} = 3.$$

Since $|t| = 3 > 2$, reject $H_0 : \beta = 0$.

An approximate 95% confidence interval is

$$2.4 \pm 1.96(0.8) = 2.4 \pm 1.568.$$

Thus,

$$[0.832,\ 3.968].$$

(b) The bootstrap estimates are

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

Their mean is

$$\bar{\theta}^* = \frac{7 + 9 + 10 + 12 + 14}{5} = \frac{52}{5} = 10.4.$$

The bootstrap standard error is the sample standard deviation:

$$\begin{aligned}\widehat{\mathrm{SE}}_{\mathrm{boot}} &= \sqrt{\frac{(7-10.4)^2 + (9-10.4)^2 + (10-10.4)^2 + (12-10.4)^2 + (14-10.4)^2}{4}} \\ &= \sqrt{\frac{11.56 + 1.96 + 0.16 + 2.56 + 12.96}{4}} \\ &= \sqrt{7.3} \approx 2.702.\end{aligned}$$

A normal-approximation 95% confidence interval is

$$\hat{\theta} \pm 1.96\, \widehat{\mathrm{SE}}_{\mathrm{boot}} = 10 \pm 1.96(2.702).$$

Therefore,

$$[4.704,\ 15.296].$$

(c) There are $m = 7$ tests.

Bonferroni. The Bonferroni threshold is

$$\frac{\alpha}{m} = \frac{0.05}{7} \approx 0.00714.$$

Only 0.004 is below this threshold, so Bonferroni rejects

$$H_{0,1}.$$

Benjamini–Hochberg. At $q = 0.10$, the thresholds jq/m are

j	1	2	3	4	5	6	7
jq/m	0.0143	0.0286	0.0429	0.0571	0.0714	0.0857	0.1000

Compare:

j	1	2	3	4	5	6	7
$p_{(j)}$	0.004	0.008	0.045	0.050	0.085	0.120	0.200
jq/m	0.0143	0.0286	0.0429	0.0571	0.0714	0.0857	0.1000

The largest j such that $p_{(j)} \le jq/m$ is $j = 4$, since

$$0.050 \le 0.0571.$$

Therefore, BH rejects

$$H_{0,1},\ H_{0,2},\ H_{0,3},\ H_{0,4}.$$

Problem 6

(a) Since the data are centered, the directional variance along a unit vector $\mathbf{u}$ is

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{u}^\top x_i)^2.$$

Along $\mathbf{e}_1 = (1, 0)$,

$$\frac{(-3)^2 + (-2)^2 + 0^2 + 2^2 + 3^2}{5} = \frac{26}{5} = 5.2.$$

Along $\mathbf{e}_2 = (0, 1)$,

$$\frac{(-4)^2 + (-1)^2 + 0^2 + 1^2 + 4^2}{5} = \frac{34}{5} = 6.8.$$

Along $\mathbf{u} = \frac{1}{\sqrt{2}}(1, 1)$, the projected scores are

$$-\frac{7}{\sqrt{2}}, \quad -\frac{3}{\sqrt{2}}, \quad 0, \quad \frac{3}{\sqrt{2}}, \quad \frac{7}{\sqrt{2}}.$$

Thus the directional variance is

$$\frac{1}{5}\left[\frac{49}{2} + \frac{9}{2} + 0 + \frac{9}{2} + \frac{49}{2}\right] = \frac{58}{5} = 11.6.$$

Among the three directions, $\mathbf{u} = \frac{1}{\sqrt{2}}(1, 1)$ has the largest directional variance, so it seems closest to the first PC direction among the three candidates.

(b) The total variance is

$$26 + 12 + 6 + 3 + 2 + 1 = 50.$$

Thus the PVEs are

$$\frac{26}{50}, \quad \frac{12}{50}, \quad \frac{6}{50}, \quad \frac{3}{50}, \quad \frac{2}{50}, \quad \frac{1}{50}.$$

Numerically:

$$0.52, \quad 0.24, \quad 0.12, \quad 0.06, \quad 0.04, \quad 0.02.$$

The cumulative PVE for the first two PCs is

$$\frac{26 + 12}{50} = \frac{38}{50} \approx 0.76.$$

The cumulative PVE values are

$$\frac{26}{50} = 0.52, \quad \frac{38}{50} = 0.76, \quad \frac{44}{50} = 0.88, \quad \frac{47}{50} = 0.94.$$

Thus the smallest number of PCs needed to explain at least 90% of total variance is

$$4.$$

(c) Initially,

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

The initial centroids are

$$\bar{z}_1 = \frac{(0,0) + (4,0)}{2} = (2, 0), \qquad \bar{z}_2 = \frac{(0,1) + (5,1)}{2} = (2.5, 1).$$

Reassigning points to the nearest centroid gives

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

For the second iteration, the centroids are

$$\bar{z}_1 = \frac{(0,0) + (0,1)}{2} = (0, 0.5), \qquad \bar{z}_2 = \frac{(4,0) + (5,1)}{2} = (4.5, 0.5).$$

Reassigning again leaves the clusters unchanged:

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

Thus the algorithm has converged after the second iteration.

(d) With complete linkage, the distance between two clusters is the maximum pairwise distance between points in the two clusters.

The first merge is A with B, since

$$d(A, B) = 1$$

is the smallest distance. Merge height: 1.

The second merge is C with D, since

$$d(C, D) = 2$$

is the next smallest distance. Merge height: 2.

We then have clusters $\{A, B\}$ and $\{C, D\}$. Their complete-linkage distance is

$$\max\{d(A, C), d(A, D), d(B, C), d(B, D)\} = \max\{5, 6, 4, 5\} = 6.$$

Thus the final merge occurs at height 6.

Cutting the dendrogram to obtain two clusters gives

$$\{A, B\}, \qquad \{C, D\}.$$

(e*) PCA preserves directions of largest variance, not necessarily directions that separate clusters. For example, suppose two clusters are long horizontal bands with large variation in the x-direction but are separated vertically in the y-direction. The first PC may point horizontally because that is where the overall variance is largest. If we keep only the first PC, the vertical separation between the two clusters may disappear, making the clusters indistinguishable.

Thus, reducing to only the first PC before clustering can destroy meaningful cluster structure when cluster separation occurs in a lower-variance direction.