STA 35C Statistical Data Science III

Midterm exam 1 solution

Instructor: Dogyoon Song

Problem 1

(a) **Scenario (i).** The response is

$$Y = \text{whether the applicant defaults.}$$

The predictors are

$$X = \text{income, balance, and credit history.}$$

This is a **classification** problem because Y is categorical. The primary goal is **prediction**, because the bank wants to predict default status for a new applicant.

Scenario (ii). The response is

$$Y = \text{exam score.}$$

The predictors are

$$X = \text{study hours and lecture attendance.}$$

This is a **regression** problem because Y is quantitative. The primary goal is **inference**, because the researcher wants to understand how the predictors are associated with exam scores.

(b) The event $\{X = 1\}$ consists of the outcomes with exactly one Head:

$$\{X = 1\} = \{HT, TH\}.$$

Since the two coin tosses are fair, all four outcomes are equally likely. Therefore,

$$P(X = 1) = \frac{2}{4} = \frac{1}{2}.$$

(c) The fitted model is

$$\widehat{Y} = 62 + 4D + 3X.$$

The coefficient 4 means that, students who attended more than half of the lectures are predicted to score 4 points higher on average than students who did not attend, holding study hours fixed.

(*Note* that this is an association from the fitted regression model, not necessarily a causal effect.)

(d) We are given

$$\hat{p}(x_{\text{new}}) = 0.32.$$

If $p^* = 0.5$, then

$$0.32 < 0.5,$$

so the predicted class is 0.

If $p^* = 0.2$, then

$$0.32 \geq 0.2,$$

so the predicted class is 1.

Problem 2

(a) Given

$$X \mid A \sim \text{Binomial}\left(3, \frac{1}{3}\right).$$

First,

$$P(X = 0 \mid A) = \binom{3}{0}\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^3 = \frac{8}{27}.$$

Hence

$$P(X \geq 1 \mid A) = 1 - P(X = 0 \mid A) = 1 - \frac{8}{27} = \frac{19}{27}.$$

For a binomial random variable,

$$\mathbb{E}[X \mid A] = np = 3 \cdot \frac{1}{3} = 1,$$

and

$$\text{Var}(X \mid A) = np(1-p) = 3 \cdot \frac{1}{3} \cdot \frac{2}{3} = \frac{2}{3}.$$

Therefore,

$$P(X = 0 \mid A) = \frac{8}{27}, \qquad P(X \geq 1 \mid A) = \frac{19}{27}, \qquad \mathbb{E}[X \mid A] = 1, \qquad \text{Var}(X \mid A) = \frac{2}{3}.$$

(b) We are given

$$\mathbb{E}[X] = \frac{2}{3}, \qquad \text{Var}(X) = \frac{5}{9},$$

and

$$\mathbb{E}[Y] = 4, \qquad \text{Var}(Y) = 5, \qquad \text{Corr}(X, Y) = 0.3 = \frac{3}{10}.$$

The deployment loss score is

$$W = 9X + 2Y + 5.$$

First,

$$\mathbb{E}[W] = 9\mathbb{E}[X] + 2\mathbb{E}[Y] + 5 = 9 \cdot \frac{2}{3} + 2(4) + 5 = 6 + 8 + 5 = 19.$$

Next,

$$\text{Cov}(X, Y) = \text{Corr}(X, Y)\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)} = \frac{3}{10}\sqrt{\frac{5}{9}}\sqrt{5} = \frac{3}{10} \cdot \frac{5}{3} = \frac{1}{2}.$$

Since constants do not affect variance,

$$\begin{aligned}\text{Var}(W) &= \text{Var}(9X + 2Y) = 81\text{Var}(X) + 4\text{Var}(Y) + 2(9)(2)\text{Cov}(X, Y) \\ &= 81 \cdot \frac{5}{9} + 4(5) + 36 \cdot \frac{1}{2} = 45 + 20 + 18 = 83.\end{aligned}$$

Therefore,

$$\mathbb{E}[W] = 19, \qquad \text{Var}(W) = 83.$$

(c) Model A and Model B are equally likely:

$$P(A) = P(B) = \frac{1}{2}.$$

For Model B,

$$X \mid B \sim \text{Binomial}\left(3, \frac{1}{6}\right),$$

so

$$P(X = 0 \mid B) = \left(\frac{5}{6}\right)^3 = \frac{125}{216}.$$

Thus,

$$P(B \cap \{X = 0\}) = P(B)P(X = 0 \mid B) = \frac{1}{2} \cdot \frac{125}{216} = \frac{125}{432}.$$

By the law of total probability,

$$P(X = 0) = P(A)P(X = 0 \mid A) + P(B)P(X = 0 \mid B).$$

Therefore,

$$P(X = 0) = \frac{1}{2} \cdot \frac{8}{27} + \frac{1}{2} \cdot \frac{125}{216} = \frac{4}{27} + \frac{125}{432} = \frac{189}{432} = \frac{7}{16}, \qquad \text{because } P(X = 0 \mid A) = \frac{8}{27} \text{ from part (a).}$$

Finally,

$$P(B \mid X = 0) = \frac{P(B \cap \{X = 0\})}{P(X = 0)} = \frac{125/432}{7/16} = \frac{125}{432} \cdot \frac{16}{7} = \frac{125}{189}.$$

Thus,

$$P(B \cap \{X = 0\}) = \frac{125}{432}, \qquad P(X = 0) = \frac{7}{16}, \qquad P(B \mid X = 0) = \frac{125}{189}.$$

Interpretation: given that none of the three audited predictions was incorrect, the posterior probability that the batch came from Model B is 125/189, which is larger than 1/2 because Model B has a lower incorrect-prediction probability than Model A.

Problem 3

(a) We are given

$$\bar{x} = 5, \qquad \bar{y} = 76,$$

and

$$\sum_i (x_i - \bar{x})^2 = 40, \qquad \sum_i (x_i - \bar{x})(y_i - \bar{y}) = 120.$$

The least squares slope is

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} = \frac{120}{40} = 3.$$

The intercept is

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 76 - 3(5) = 61.$$

Thus,

$$\hat{Y} = 61 + 3X.$$

For a student who studied 8 hours,

$$\hat{Y} = 61 + 3(8) = 85.$$

Therefore,

$$\hat{\beta}_1 = 3, \qquad \hat{\beta}_0 = 61, \qquad \hat{Y}(8) = 85.$$

(b) The coefficient table is

Coefficient	**Estimate**	**Std. Error**	t**-statistic**
X	2.4	0.6	?
D	5.0	3.5	?

For X and D, their t-statistics are

$$t_X = \frac{2.4}{0.6} = 4, \qquad \text{and} \qquad t_D = \frac{5.0}{3.5} = \frac{10}{7} \approx 1.429.$$

Using $|t| \geq 2$ as the approximate 5% significance rule, X is statistically significant because

$$|t_X| = 4 \geq 2.$$

The attendance variable D is not statistically significant by this rule because

$$|t_D| \approx 1.429 < 2.$$

Interpretation of 2.4: holding lecture attendance fixed, each additional study hour is associated with a predicted increase of 2.4 points in exam score. This effect is statistically significant by the rule above.

Interpretation of 5.0: holding study hours fixed, students who attended lectures are predicted to score 5.0 points higher than students who did not attend lectures. However, this coefficient is not statistically significant, so the data do not provide strong evidence of an attendance effect with study hours controlled.

(c) The model comparison table is

Model	**Predictors**	**Training** R^2	**Test MSE**
A	X	0.58	3.7
B	X, D	0.64	3.2
C	X, D, X^2	0.66	4.5

For prediction, we should choose the model with the smallest test MSE, which is Model B:

$$3.2 < 3.7 < 4.5.$$

Model C can have the largest training R^2 and the largest test MSE because training R^2 measures fit on the training data. A more flexible model can fit the training data better while generalizing worse to new data. This is an example of overfitting.

(d*) The fitted model is

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \beta_3 (D \times X) + \varepsilon,$$

with

$$\hat{\beta}_0 = 35, \qquad \hat{\beta}_1 = 5, \qquad \hat{\beta}_2 = 1, \qquad \hat{\beta}_3 = -0.5.$$

For $D = 0$,

$$\widehat{Y} = 35 + X.$$

For $D = 1$,

$$\widehat{Y} = 35 + 5 + X - 0.5X = 40 + 0.5X.$$

The estimated attendance effect at study hours x is

$$(40 + 0.5x) - (35 + x) = 5 - 0.5x.$$

Set this equal to zero:

$$5 - 0.5x = 0.$$

Therefore,

$$x = 10.$$

Thus, the estimated attendance effect is zero at $x = 10$ study hours.

The coefficient $\hat{\beta}_3 = -0.5$ means that the estimated attendance effect decreases by 0.5 points for each additional study hour. Equivalently, the slope with respect to study hours is 0.5 points lower for students who attended lectures.

Problem 4

(a) The fitted logistic regression model is

$$\log\left(\frac{\hat{p}(x)}{1 - \hat{p}(x)}\right) = -2 + x_1 + x_2.$$

For

$$x_{\text{new}} = (1, 2),$$

the fitted log-odds are

$$-2 + 1 + 2 = 1.$$

Therefore,

$$\hat{p}(x_{\text{new}}) = \frac{e^1}{1 + e^1} = \frac{e}{1 + e} \approx \frac{2.718}{3.718} \approx 0.731.$$

Since

$$0.731 \geq 0.5,$$

we predict

$$\hat{y}_{\text{new}} = 1.$$

For $p^* = 0.5$, the decision boundary is where the fitted log-odds equal 0:

$$-2 + x_1 + x_2 = 0, \qquad \text{or equivalently,} \qquad x_1 + x_2 = 2.$$

The model classifies an applicant as default when

$$-2 + x_1 + x_2 \geq 0,$$

or equivalently,

$$x_1 + x_2 \geq 2.$$

(b) The confusion matrix is

	Predicted 1	Predicted 0
$Y = 1$	18	7
$Y = 0$	10	65

Thus,

$$TP = 18, \qquad FN = 7, \qquad FP = 10, \qquad TN = 65.$$

The error rate is

$$\frac{FP + FN}{TP + TN + FP + FN} = \frac{10 + 7}{100} = 0.17.$$

The true positive rate is

$$\text{TPR} = \frac{TP}{TP + FN} = \frac{18}{18 + 7} = \frac{18}{25} = 0.72.$$

The false positive rate is

$$\text{FPR} = \frac{FP}{FP + TN} = \frac{10}{10 + 65} = \frac{10}{75} = \frac{2}{15} \approx 0.133.$$

The false negative rate is

$$\text{FNR} = \frac{FN}{TP + FN} = \frac{7}{25} = 0.28.$$

If the threshold is lowered from $p^* = 0.5$ to $p^* = 0.2$, more applicants will generally be predicted as class 1. Therefore, FPR generally increases and FNR generally decreases.

We cannot determine in general whether the overall error rate increases or decreases, because that depends on the tradeoff between the increase in false positives and the decrease in false negatives.

(c) There are six points from class 0 with y-values $0, 1, 2, 2, 3, 4$, and two points from class 1 with y-values $4, 6$. Thus,

$$\hat{\pi}_0 = \frac{6}{8} = \frac{3}{4}, \qquad \text{and} \qquad \hat{\pi}_1 = \frac{2}{8} = \frac{1}{4}.$$

The class means are

$$\hat{\mu}_0 = \frac{0 + 1 + 2 + 2 + 3 + 4}{6} = \frac{12}{6} = 2, \qquad \text{and} \qquad \hat{\mu}_1 = \frac{4 + 6}{2} = 5.$$

We are given the pooled-variance estimate $\hat{\sigma}^2 = 2$.

Therefore, the discriminant functions are

$$\delta_0(z) = z\frac{2}{2} - \frac{2^2}{2 \cdot 2} + \log\left(\frac{3}{4}\right) = z - 1 + \log\left(\frac{3}{4}\right),$$
$$\delta_1(z) = z\frac{5}{2} - \frac{5^2}{2 \cdot 2} + \log\left(\frac{1}{4}\right) = 2.5z - \frac{25}{4} + \log\left(\frac{1}{4}\right).$$

The cutoff is obtained by solving $\delta_1(z) = \delta_0(z)$. Compute the difference:

$$\begin{aligned}\delta_1(z) - \delta_0(z) &= 2.5z - \frac{25}{4} + \log\left(\frac{1}{4}\right) - \left[z - 1 + \log\left(\frac{3}{4}\right)\right] \\ &= 1.5z - \frac{21}{4} + \log\left(\frac{1/4}{3/4}\right) \\ &= 1.5z - \frac{21}{4} - \log 3.\end{aligned}$$

Therefore, the cutoff satisfies

$$1.5z^* - \frac{21}{4} - \log 3 = 0, \qquad \text{or equivalently,} \qquad z^* = \frac{2}{3}\left(\frac{21}{4} + \log 3\right) = \frac{6.349}{1.5} \approx 4.233.$$

Since class 1 has the larger mean, the rule predicts class 1 for

$$z \geq 4.233.$$

For $z = 4.0 < 4.233$, the LDA classifier predicts class 0, meaning no default.

(d*) There are 25 actual defaults and 75 actual non-defaults. Thus,

$$\text{TPR} = \frac{TP}{25}, \qquad \text{FPR} = \frac{FP}{75}.$$

The ROC points are:

p^*	1.0	0.8	0.6	0.4	0.2	0
TP	0	10	16	20	23	25
FP	0	2	5	12	30	75
TPR	0	0.4	0.64	0.8	0.92	1
FPR	0	2/75	5/75	12/75	30/75	1

Numerically,

p^*	1.0	0.8	0.6	0.4	0.2	0
(FPR, TPR)	$(0, 0)$	$(0.027, 0.400)$	$(0.067, 0.640)$	$(0.160, 0.800)$	$(0.400, 0.920)$	$(1, 1)$

The bank requires

$$\text{FPR} \leq 0.20.$$

The thresholds satisfying this constraint are

$$p^* = 1.0,\ 0.8,\ 0.6,\ 0.4.$$

Among these, the largest TPR is achieved at

$$p^* = 0.4,$$

with

$$\text{FPR} = 0.160, \qquad \text{TPR} = 0.800.$$

Therefore, among the listed thresholds, the bank should choose

$$p^* = 0.4.$$