STA 35C Statistical Data Science III

Midterm exam 1 solution

Instructor: Dogyoon Song

Problem 1: Solution (20 points + 2 bonus)

(a) $\mathbb{E}[X]$ and $\mathrm{Var}(X)$ for $X \sim \mathrm{Binomial}(2, \frac{1}{3})$.

- $\mathbb{E}[X] = n\,p = 2 \times \frac{1}{3} = \frac{2}{3}$.
- $\mathrm{Var}(X) = n\,p\,(1-p) = 2 \times \frac{1}{3} \times \frac{2}{3} = \frac{4}{9}$.

Method 2: Alternatively, $X = X_1 + X_2$ where $X_1, X_2 \sim \mathrm{Bernoulli}(\frac{1}{3})$ i.i.d. Since $\mathbb{E}[X_1] = \frac{1}{3}$ and $\mathrm{Var}(X_1) = \mathbb{E}[X_1^2] - \mathbb{E}[X_1]^2 = \frac{1}{3} - \frac{1}{9} = \frac{2}{9}$, it follows that

$$\mathbb{E}[X] = \mathbb{E}[X_1] + \mathbb{E}[X_2] = \tfrac{2}{3}, \quad \text{and} \quad \mathrm{Var}(X) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2) = \tfrac{4}{9}.$$

Method 3: Otherwise, we can directly use the PMF to obtain

$$\mathbb{E}[X] = \sum_{x=0}^{2} x\; p_X(x) = 0 \cdot \binom{2}{0} \cdot \left(\frac{2}{3}\right)^2 + 1 \cdot \binom{2}{1} \cdot \frac{1}{3} \cdot \frac{2}{3} + 2 \cdot \binom{2}{2} \cdot \left(\frac{1}{3}\right)^2 = \frac{2}{3}$$

Similarly, we can compute $\mathbb{E}[X^2] = \frac{8}{9}$, and thus, $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2 = \frac{4}{9}$.

(b) $\mathbb{E}[W]$ and $\mathrm{Var}(W)$ for $W = X + 2\,Y + 2$.
Given: $\mathbb{E}[X] = \frac{2}{3}$, $\mathrm{Var}(X) = \frac{4}{9}$, $\mathbb{E}[Y] = 9$, $\mathrm{Var}(Y) = 9$, $\mathrm{corr}(X, Y) = 0.3$.

- $\mathbb{E}[W] = \mathbb{E}[X] + 2\,\mathbb{E}[Y] + 2 = \frac{2}{3} + 2 \times 9 + 2 = \frac{2}{3} + 18 + 2 = \frac{62}{3}$.
- $\mathrm{Cov}(X, Y) = \rho \sqrt{\mathrm{Var}(X)}\,\sqrt{\mathrm{Var}(Y)} = 0.3 \times \sqrt{\frac{4}{9}} \times \sqrt{9} = 0.3 \times \frac{2}{3} \times 3 = 0.6$.

$\therefore \mathrm{Var}(W) = \mathrm{Var}(X) + 4\,\mathrm{Var}(Y) + 4\,\mathrm{Cov}(X, Y) = \frac{4}{9} + 4 \times 9 + 4 \times 0.6 = \frac{4}{9} + 36 + 2.4 = 38.4 + 0.444\ldots \approx 38.84$.

(c) Bayesian Update: Factories A vs. B.

(i) (5 points) *Probability a randomly chosen box is from A* and *has exactly one defective ($X = 1$):*
Each box is A or B with prob. $\frac{1}{2}$. If a box is from A, $p = \frac{1}{3}$. Then

$$\Pr(X = 1 \mid A) = \binom{2}{1}\left(\frac{1}{3}\right)^1\left(\frac{2}{3}\right)^1 = 2 \times \frac{1}{3} \times \frac{2}{3} = \frac{4}{9}.$$

Thus

$$\Pr(\text{A and } X = 1) = \Pr(A) \times \Pr(X = 1 \mid A) = \frac{1}{2} \times \frac{4}{9} = \frac{4}{18} = \frac{2}{9} \approx 0.2222.$$

(ii) (5 points) *Posterior* $\Pr(A \mid X = 1)$ *if* $p(B) = \frac{1}{2}, p = \frac{1}{10}$ *in B:*

$$\Pr(X = 1 \mid B) = \binom{2}{1}\left(\frac{1}{10}\right)^1\left(\frac{9}{10}\right)^1 = 2 \times 0.1 \times 0.9 = 0.18.$$

$$\Pr(\text{B and } X = 1) = \tfrac{1}{2} \times 0.18 = 0.09.$$

Therefore,

$$\Pr(A \mid X = 1) = \frac{\Pr(A,\ X = 1)}{\Pr(A,\ X = 1) + \Pr(B,\ X = 1)} = \frac{\frac{2}{9}}{\frac{2}{9} + 0.09} = \frac{0.2222}{0.2222 + 0.09} = \frac{0.2222}{0.3122} \approx 0.712.$$

(iii*) (*2 bonus points) Now with four factories (A,B,C,D) of unknown priors. If $X = 1$,

- Factory C ($p = 1$) always yields $X = 2$. So $\Pr(X = 1 \mid C) = 0$.
- Factory D ($p = 0$) always yields $X = 0$. So $\Pr(X = 1 \mid D) = 0$.

Hence the posterior of D is 0 if $X = 1$, regardless of the prior.

Problem 2: Solution (25 points)

(a) Four Scenarios (12 points).

(i) Nutritionist (3 pts)

- X = (age, weight, exercise), Y = daily protein intake.
- *Regression* problem (continuous Y).
- Primarily *prediction* to forecast intake.

(ii) Market Analyst (3 pts)

- X = browsing habits, Y = phone plan {A,B,C}.
- *Classification* problem (categorical Y).
- Goal is *prediction* for the new user.

(iii) Admissions Officer (3 pts)

- X = homework grades, Y = final exam score (numeric).
- *Regression* problem.
- Focus on *inference*: which assignments matter most.

(iv) Real Estate Agent (3 pts)

- X = (location, bedrooms, area, building age), Y = monthly rent.
- *Regression* problem.
- Goal is *inference*: find the factor(s) significantly affecting rent.

(b) Model Comparison (13 points).

(i) (5 points) Evaluate predictive performance via a metric, such as MSE $= \frac{1}{n}\sum_{i=1}^{n}\big(y_i - f(x_i)\big)^2$. A lower MSE indicates better prediction. (Ideally we want to check on a test set too, but we cannot.)

(ii) (4 points) If Model (2) outperforms Model (1) *only* on training data, it might overfit. We might still prefer Model (1) for interpretability or simpler structure.

(iii) (4 points) If Model (2) also excels on test data, it likely generalizes well. Model (1) might be chosen for inference/interpretability or practical concerns (e.g., simpler to explain or cheaper to implement).

Problem 3: Solution (40 points + 2 bonus)

We have Y = puzzle-solving time (minutes), X_1 = indicator for ≥ 2 yrs experience, X_2 = memory score.

(a) Prediction (15 points).

(i) (5 points)

$$\hat{Y}_A = 11.2 - 5 \times (1) = 6.2, \qquad \hat{Y}_B = 10 - 0.6 \times (7) = 5.8.$$

(ii) (5 points) Model A: $R^2 = 0.50$, Model B: $R^2 = 0.64$. B is better at explaining Y's variance. R^2 is fraction of Y's variation (variance) explained by the model. Equivalently, $R^2 = 1 - \frac{RSS}{TSS}$.

(iii) (5 points) A model with both X_1, X_2 yields $R^2 = 0.70$. This *might* be better, capturing both factors, evidenced by the increase in R^2. However, this is *not necessarily* better if overfitting, and adjusted R^2 might not have increased by much, as R^2 might have increased just by chance with additional predictors.

(b) Coefficients & Inference (15 points). We fit (1) $Y \sim X_1$ and (2) $Y \sim X_1 + X_2$ with:

$$\text{(1) Simple: } \hat{\beta}_1 = -5,\ SE = 1.67, \quad \text{(2) Multiple: } \hat{\beta}_1 = -1.6,\ SE = 1.6.$$

(i) (5 points)

- **Simple model:** $t = \frac{-5}{1.67} \approx -3.0 \Rightarrow p \approx 0.003 < 0.05$ (significant).
- **Multiple model:** $t = \frac{-1.6}{1.6} = -1.0 \Rightarrow p \approx 0.317 > 0.05$ (not significant).

(ii) (5 points)

- **Simple:** $\beta_1 = -5$ means participants with 2+ years' experience solve the puzzle about 5 minutes faster than $X_1 = 0$ on average.
- **Multiple:** $\beta_1 = -1.6$ indicates participants with 2+ years' experience solve the puzzle about only 1.6 minutes faster than $X_1 = 0$ on average, once short-term memory score (X_2) is controlled.

(iii) (5 points) If X_1 correlates with X_2, omitting X_2 can inflate X_1's effect by including its indirect influence via X_2. For example, people with higher memory scores may be likelier to enjoy puzzles and thus more experience, or puzzle experience might improve memory. Controlling for X_2 isolates X_1's direct effect, thereby mitigating confounding.

(iv*) (*2 bonus points) In this model, $\beta_1 = -7$ is the intercept difference at $X_2 = 0$. Thus, at X_2 fixed at 0, participants with 2+ years' experience solve puzzles 7 minutes faster on average than those without.

(c) Adding More Predictors (10 points).

(i) (4 points) Y–X_2 relation looks nonlinear; we can consider adding a higher-order term in X_2, e.g., X_2^2.
* **Note:** X_1 is a dummy variable, and Y is numeric (not categorical); seeing two clusters is normal, and classification methods (e.g., logistic) don't apply here.

(ii) (3 points) As X_3 seems uncorrelated with Y, it might not help to explain Y. Possibly skip X_3.

(iii) (3 points) Although X_4 seems strongly associated with Y, it also strongly correlates with X_2. Including both can cause collinearity, and we should choose only one or carefully interpret; perhaps skip X_4.

Problem 4: Solution (35 points + 4 bonus)

(a) 2D Logistic Regression (20 points).

(i) (5 points) Compute $\hat{p}(x_{\text{test}})$
With $\hat{\beta}_0 = -2$, $\hat{\beta}_1 = -1$, $\hat{\beta}_2 = 2$, for $x_{\text{test}} = (1, 1)$:

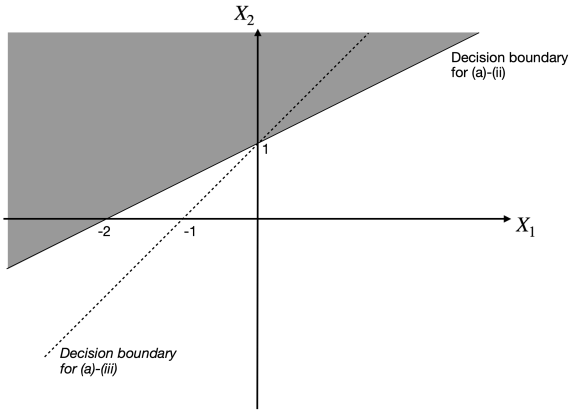
$$\log\left(\frac{p}{1-p}\right) = -2 + (-1)\cdot 1 + 2\cdot 1 = -1. \quad \Longrightarrow \quad p = \frac{e^{-1}}{1+e^{-1}} \approx 0.269 < 0.5 \quad \Longrightarrow \quad \hat{y}_{\text{test}} = 0.$$

(ii) (5 points) Decision Boundary
The given decision rule predicts $\hat{y} = 1$ if and only if

$$\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 \geq \log\left(\frac{p^*}{1-p^*}\right) = 0, \qquad \text{which is equivalent to} \qquad x_2 \geq -\frac{\hat{\beta}_1}{\hat{\beta}_2} x_1 - \frac{\hat{\beta}_0}{\hat{\beta}_2} = \frac{1}{2} x_1 + 1.$$

In the last equality, we plugged in $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (-2, -1, 2)$.

(iii*) (*2 bonus) Changing to $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2) = (-1, -1, 1)$, we now predict $\hat{y} = 1$ if and only if $x_2 \geq x_1 + 1$.



(iv) (5 points) Confusion Matrix

	Pred = 1	Pred = 0
$Y = 1$	35	5
$Y = 0$	15	45

TPR $= \frac{35}{35+5} = 0.875$, FPR $= \frac{15}{15+45} = 0.25$.

(v) (5 points) Lowering p^* With $p^* = 0.1$ vs. 0.5, more borderline cases → "1." FP ↑, and FN ↓.

(b) 1D LDA (15 points). Each species has a normal distribution with the *same* variance but different means, and we apply LDA to classify by weight.

(i) (4 points) Sample Means & Pooled s

Species A: $\{1.5, 2.5\} \implies \bar{x}_A = 2.0.$ **Species B:** $\{2.0, 3.0, 4.0\} \implies \bar{x}_B = 3.0.$

Pooled covariance:

$$s^2 = \frac{\left[(1.5-2)^2 + (2.5-2)^2\right] + \left[(2-3)^2 + (3-3)^2 + (4-3)^2\right]}{5-2} = \frac{0.5+2}{3} = \frac{5}{6} \approx 0.8333.$$

(ii) (4 points) Linear Discriminants

$\pi_A = \frac{2}{5},\ \pi_B = \frac{3}{5}$. Then $\log(\frac{2}{5}) \approx -0.916,\ \log(\frac{3}{5}) \approx -0.511$.

Thus, the linear discriminant functions in this problem reduce to:

$$\delta_A(x) = \frac{x\,\bar{x}_A}{s^2} - \frac{\bar{x}_A^2}{2\,s^2} + \log(\pi_A) = \frac{12}{5}(x-1) + \log\left(\frac{2}{5}\right),$$
$$\delta_B(x) = \frac{x\,\bar{x}_B}{s^2} - \frac{\bar{x}_B^2}{2\,s^2} + \log(\pi_B) = \frac{18}{5}\left(x - \frac{3}{2}\right) + \log\left(\frac{3}{5}\right).$$

(iii) (4 points) Predict at $x_{\text{new}} = 2.5$

From the linear discriminant functions above, we get

$$\delta_A(x) - \delta_B(x) = -\frac{6}{5}x + 3 + \log\left(\frac{2}{3}\right). \tag{1}$$

Inserting $x = 2.5$, we get

$$\delta_A(2.5) - \delta_B(2.5) = -3 + 3 + \log\left(\frac{2}{3}\right) \approx -0.405 < 0.$$

So classify **Species B**.

(iv) (3 points) Adding 4 More A's

With additional data points, A has $2 + 4 = 6$ crabs vs. B has 3, so $\pi_A = \frac{6}{9} = 0.6667$, $\pi_B = 0.3333$. $\log(\frac{0.6667}{0.3333}) = \log(2) \approx 0.693$, a positive shift. Following the same steps as above in (ii)–(iii), we get

$$\delta_A(2.5) - \delta_B(2.5) = -3 + 3 + \log\left(\frac{6}{3}\right) = \log 2 \approx 0.693 > 0.$$

Now we predict **Species A** instead of Species B.

(v*) (*2 bonus points)

Requiring $\Pr(A \mid x) \geq p^* > 0.5$ translates to $\delta_A(x) - \delta_B(x) \geq \log(\frac{p^*}{1-p^*})$. Thus, with $p^* = 0.9$,

$$\text{predict } \hat{Y} = A \quad \text{if and only if} \quad \delta_A(x) - \delta_B(x) > 2\log(3) \approx 2.197.$$

Following (1) in (iii),

$$\delta_A(2) - \delta_B(2) = -\frac{6}{5} \times 2 + 3 + \log\left(\frac{2}{3}\right) \approx 0.195 < 2.197.$$

Thus, we would classify the crab with x'_{new} as **Species B** under $p^* = 0.9$ to avoid missing B crabs; note that it would have been **Species A** with the original threshold $p^* = 0.5$ as $\delta_A(2) - \delta_B(2) \approx 0.195 > 0$.