STA 35C – Homework 3

Submission due: Tue, May 6 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a PDF file, named with your UC Davis email ID and homework number (e.g., `dgsong_hw3.pdf`) to Canvas ("Homework 3" under "Assignment"). Please make sure to include "STA 35C," your name, and the last four digits of your student ID on the front page. No late homework will be accepted for any reason; submissions after the deadline will receive 0 points. For more details on submission requirements and the late submission policy, see the syllabus.

Choose 2 problems from Problems 1–3 (the ones you wish to review), plus **Problems 4 and 5**. **In total, you will submit answers to 4 problems**: (1) **2 of your choice** from Problems 1–3; (2) **Problem 4**; and (3) **Problem 5**. If you (bravely) complete all 5 problems, your final score will be the sum of the best 2 from Problems 1–3 + a 5-point hard-worker bonus + Problem 4's score + Problem 5's score.

You may want to use R to solve Problem 2-(c), Problem 4-(b), and Problem 5.

Problem 1: Probability (25 points in total).

A company packages products in boxes of two items each. Let

$$X = \text{number of defective items in a box of size 4.}$$

Assume each item is defective with probability $\frac{1}{4}$, independently of others.

(a) (5 points) Compute $\mathbb{E}[X]$ and $\mathrm{Var}(X)$. (*Hint:* You may use $X = X_1 + X_2 + X_3 + X_4$ where $X_i \sim \text{Bernoulli}(\frac{1}{4})$ i.i.d.)

(b) (8 points) Suppose the inspection time Y (in minutes) for each box is a continuous random variable with PDF

$$f_Y(y) = \begin{cases} \frac{1}{4}e^{-\frac{1}{4}y}, & y \geq 0, \\ 0, & y < 0. \end{cases}$$

We define the total cost

$$W = 10 + X + 4\,Y.$$

Compute $\mathbb{E}[W]$ and $\mathrm{Var}(W)$, assuming $\mathrm{corr}(X, Y) = 0.2$. (*Hint:* $\mathrm{Cov}(X, Y) = \rho\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}$.)

(c) (12 points total) Suppose that boxes come from either Factory A or B, with probability 0.5 each.

- Factory A: Probability of defect per item is 0.25 (i.i.d.).
- Factory B: Probability of defect per item is 0.10 (i.i.d.).

(i) (6 points) What is the probability that a randomly chosen box is from Factory A *and* has exactly one defective item?

(ii) (6 points) Given you observe $X = 1$ defective item, find the posterior probability that the box came from Factory A.

Problem 2: Logistic regression (25 points in total).

Consider the logistic model for binary classification:

$$\log \frac{p(X)}{1-p(X)} \;=\; \beta_0 + \beta_1 X_1 + \beta_2 X_2, \quad p(X) = \Pr[Y = 1 \mid X].$$

Given $X = (x_1, x_2)$, we classify $\hat{Y} = 1$ if $p(x_1, x_2) \geq p^*$ and $\hat{Y} = 0$ otherwise. Suppose the estimated coefficients are:

$$\hat{\beta}_0 = -1.2, \quad \hat{\beta}_1 = 2.0, \quad \hat{\beta}_2 = -0.5.$$

(a) (5 points) You observe a new test point $x_{\text{test}} = (x_1, x_2) = (2, 3)$. Decide whether $\hat{y}_{\text{test}} = 1$ or $\hat{y}_{\text{test}} = 0$ at $p^* = 0.5$.

(b) (7 points) Derive the equation of the decision boundary for $p^* = 0.5$, and sketch this line in the (x_1, x_2) plane, indicating where $\hat{Y} = 1$ vs. $\hat{Y} = 0$.

(c) (13 points total) Use R to fit the model on a dataset of 100 points.

```
set.seed(1)  # for reproducibility
X1_1 <- rnorm(40, mean = 2, sd = 1)
X2_1 <- rnorm(40, mean = 2, sd = 1)

# Generate (X1, X2) for label=0
X1_0 <- rnorm(60, mean = 0, sd = 1.2)
X2_0 <- rnorm(60, mean = 0, sd = 1.2)

# Combine into a single dataset
X1 <- c(X1_1, X1_0)
X2 <- c(X2_1, X2_0)
Label <- c(rep(1, 40), rep(0, 60))

# Create a data frame
df <- data.frame(X1, X2, Label)
```

(i) (5 points) Create a confusion matrix, and compute the true positive rate (TPR = sensitivity), and the false positive rate (FPR = 1 - specificity) at $p^* = 0.5$.

(ii) (8 points) For $p^* \in \{0, 0.05, 0.1, \ldots, 0.95, 1\}$, compute TPR and FPR, then plot the ROC curve (FPR, TPR).

Problem 3: Linear discriminant analysis (25 points total + 5 bonus points).

You catch crabs of two species, A and B, recording these weights (pounds):

$$\text{Species A: } x \in \{1.0,\ 2.0,\ 3.0\}, \quad \text{Species B: } x \in \{3.0,\ 4.0,\ 5.0,\ 6.0\}.$$

Assume:

- Species A: Gaussian with mean μ_A, variance σ^2.
- Species B: Gaussian with mean μ_B, variance σ^2.

(a) (5 points) Compute the sample means $\bar{x}_A$, $\bar{x}_B$ and the pooled sample variance s^2.

(b) (5 points) Write the linear discriminant functions $\delta_A(x)$ and $\delta_B(x)$, using $\bar{x}_A$, $\bar{x}_B$, s^2 obtained above.

(c) (5 points) If $x_{\text{new}} = 3.2$, which species do you predict? Show your reasoning.

(d) (5 points) If you add three more data points for Species A (with about the same mean and variance), would your prediction at $x_{\text{new}} = 3.2$ change? Explain briefly.

(e) (5 points) Suppose you do *not* want to miss any crabs of Species B. You decide to predict A only if $\Pr(Y = A \mid X) \geq p^*$ with $p^* > 0.5$ (e.g. $p^* = 0.9$). How does this modify the LDA decision rule in terms of $\delta_A(x)$ and $\delta_B(x)$? State your new decision boundary and apply it to $x_{\text{new}} = 3.2$.

(f*) (*5 bonus points) Suppose Species A and B might be Laplace-distributed instead of Gaussian. The PDF of Laplace with mean μ and variance $2b^2$ is

$$f(x; \mu, b) = \frac{1}{2b} \exp\Big(-\frac{|\,x - \mu|}{b}\Big).$$

(i) Draw the graph of $f(x; 0, 1)$ and compare it with the graph of Gaussian density.

(ii) Derive the discriminant function under Laplace density and compare it with the linear discriminant function under Gaussian. (Hint: consider $\log(\pi_k f_k(x))$.)

Problem 4: Cross-validation (20 points in total + 5 bonus points).

(a*) (*5 bonus points) In your own words:

(i) Explain how LOOCV is implemented.

(ii) Briefly discuss advantages/disadvantages of LOOCV vs. the validation set approach and k-fold CV.

(b) (20 points) [JWHT21, Chapter 5, Exercise 8]. You may want to use R to solve this problem.

Problem 5: The Bootstrap (20 points in total).

Using R, create a dataset with:

```
set.seed(1)
x <- rnorm(100)
y <- -1 + 2*x + rnorm(100)
```

(a) Fit a simple linear regression of Y on X to obtain $\hat{\beta}_1$. Using the standard formula for a linear regression coefficient's standard error, write down a 95% confidence interval for β_1.

(b) Now imagine drawing entirely new datasets from the true model. For each $k = 1, \ldots, 1000$:

```
set.seed(k)
x <- rnorm(100)
y <- -1 + 2*x + rnorm(100)
```

Fit a linear model for each k and collect $\hat{\beta}_1^{(k)}$.

Draw a histogram of these 1000 estimates and report the fraction of $\hat{\beta}_1^{(k)}$ values contained in the *single* 95% CI from part **(a)**. Discuss whether this fraction is close to 0.95.

(c) We often have just one dataset (the original one from `set.seed(1)`). Use the *bootstrap* to draw 1000 bootstrap samples of size 100 with replacement. Re-fit the linear model on each bootstrap sample, yielding $\hat{\beta}_1^*$. Draw a histogram of $\hat{\beta}_1^*$, compare its shape/spread to part **(b)**, and compute the fraction of $\hat{\beta}_1^*$ values inside the 95% CI from part **(a)**. Discuss how close it is to 0.95, and why it might differ.

(d) Repeat steps **(a)**–**(c)**, but now assuming the data are generated from a quadratic model:

```
set.seed(1)
x <- rnorm(100)
y <- -1 + 2*x - x^2 + rnorm(100)
```

References

[JWHT21] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning: with Applications in R*, volume 112 of *Springer Texts in Statistics*. Springer, New York, NY, 2nd edition, 2021.