STA 35C – Homework 2

Submission due: Tue, April 22 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_hw2.pdf`) to Canvas ("Homework 2" 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.

Problem 1 (20 points in total).

(a) (4 points) Explain supervised and unsupervised learning in your own words, highlighting their purposes, similarities and differences.

(b) (4 points) Explain regression and classification in your own words, highlighting their similarities and differences.

(c) (4 points) Explain prediction and inference in your own words, with at least one example scenario/tasks for each.

(d) (4 points) Describe the difference between parametric and non-parametric methods, comparing their relative advantages and disadvantages.

(e) (4 points) Explain overfitting and the bias-variance tradeoff in your own words.

Problem 2 (30 points in total + 5 bonus points).

Scenario: You are analyzing an outcome measure Y (e.g., blood pressure or some clinical score) collected from n patients. Some patients received a *new treatment* ($D = 1$), while others received *standard* (control) treatment ($D = 0$). You also have an additional continuous predictor X (e.g., patient age or baseline risk). You suspect that Y may differ substantially between the two groups, possibly due to the different treatments.

Data: Your dataset consists of n tuples:

$$(y_1, d_1, x_1),\ \ldots,\ (y_n, d_n, x_n),$$

where y_i is the response, $d_i \in \{0, 1\}$ is treatment indicator, and x_i is the additional contextual predictor for patient $i \in [n]$.

For concreteness, suppose you collected 10 data points as given below:

$(-9.3704, 1, 39.107)$, $(-9.2399, 1, 38.245)$, $(-8.1464, 0, 24.318)$, $(-7.1728, 1, 36.825)$, $(-6.3464, 1, 29.314)$,
$(-5.8743, 0, 23.497)$, $(-5.6763, 1, 31.137)$, $(-1.8406, 0, 14.919)$, $(-0.8549, 0, 13.691)$, $(1.0222, 0, 10.631)$.

Objective: Suppose that

$$\begin{cases} \text{If } D = 1: & Y = a + b\,X + \varepsilon \\ \text{If } D = 0: & Y = c + d\,X + \varepsilon \end{cases}$$

where a, b, c, d are unknown constants and ε is random noise.

We want to test if the new treatment is more effective than the standard one. For instance, we can try to estimate the *average treatment effect*

$$\tau = \frac{1}{10}\sum_{i=1}^{10}\Big\{ \underbrace{(a + bx_i)}_{\text{response under treatment}} - \underbrace{(c + dx_i)}_{\text{response under control}} \Big\},$$

even though only one outcome (treated or control) is observed per patient. In this problem, we make several attempts to tackle this problem with methods learned so far. You may use R to solve this problem.

(a) (5 points) Initially, you compare the average outcome in each group:

$$\bar{y}_{\text{ctrl}} = \frac{1}{n_0}\sum_{i:\, d_i=0} y_i, \quad \bar{y}_{\text{treat}} = \frac{1}{n_1}\sum_{i:\, d_i=1} y_i,$$

where n_0 is the number of patients with $D = 0$, and n_1 with $D = 1$ (here, $n_0 = n_1 = 5$). Conclude whether $\bar{y}_{\text{ctrl}} > \bar{y}_{\text{treat}}$, $\bar{y}_{\text{ctrl}} < \bar{y}_{\text{treat}}$, or you cannot tell, and justify. (*Hint:* Use a two-sample t-test covered in STA 35B, cf. Dr. Frei's lecture materials from Winter 2024, or any valid test to see if you can reject $H_0 : \bar{y}_{\text{ctrl}} = \bar{y}_{\text{treat}}$.)

(b) (10 points) Now consider a simple linear model that only takes D into account, *ignoring* X:

$$Y \;=\; \beta_0 + \beta_1 D + \varepsilon.$$

(i) Interpret β_0 and β_1.

(ii) Is β_1 significantly different from 0, and what does that imply about the difference between treatment vs. control?

(iii) Compare your conclusion drawn from this dummy model to that from simply taking $\bar{y}_{\text{treat}} - \bar{y}_{\text{ctrl}}$ in **(a)**.

(c) (10 points) Next, you include the additional predictor X into your model:

$$Y = \beta_0 + \beta_1 D + \beta_2 X + \varepsilon.$$

(i) Explain how β_2 is interpreted in this model.

(ii) Does including X change how we interpret β_1 relative to part **(b)**? If yes, how?

(iii) Why might controlling for X (i.e., conditioning on a value of X) alter the apparent difference between groups? (*Hint:* It may be helpful to try a scatter plot to see how X distributes by group.)

(d) (5 points) Finally, you propose:

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

which is equivalent to

$$Y = \begin{cases} \beta_0 + \beta_1 \;+\; (\beta_2 + \beta_3)X + \varepsilon & \text{if } D = 1, \\ \beta_0 \;+\; \beta_2 X + \varepsilon & \text{if } D = 0. \end{cases}$$

(i) Explain how β_3 captures a possible difference in *slopes* across the two groups.

(ii) Would you conclude the slopes for the two groups are same or not?

(e*) (5 bonus points) Provide your estimates of a, b, c, d, and τ. What can you conclude about τ?

Problem 3 (30 points in total + 5 bonus points).

Scenario: A company administers a mandatory test to new hires in a training program. Each candidate either *Passes* ($Y = 1$) or *Fails* ($Y = 0$). The HR department believes that prior work experience (X, in months) influences the probability of passing.

(a) (8 points) Assume a logistic regression model

$$\Pr(Y = 1 \mid X = x) = \frac{1}{1 + e^{-(\alpha+\beta x)}}.$$

(i) Interpret α and β in this context. (*Hint:* Think about how log-odds is affected.)

(ii) How does a 1-month increase in X affect the *odds* of passing?

(b) (12 points) Suppose HR policy says if $\hat{p}(x) \geq 0.4$, the candidate is predicted to pass and placed in an "advanced" track; otherwise, they go to a "basic" track.

(i) If $\alpha = -3.0$ and $\beta = 0.07$, compute $\hat{p}(x)$ for $x = 20$ months.

(ii) Which track would that candidate be assigned to?

(iii) If the *true* pass probability is 0.7 for a certain candidate, but they are assigned to the "basic" track under the 0.4 cutoff, is this a false positive or a false negative? Briefly explain.

(iv) Discuss the pros and cons of using 0.4 rather than 0.5 as a threshold. In particular, consider which type of misclassification might be costlier for the company.

(c) (10 points) Implement the following in an R script and report your results.

(i) Simulate a dataset of size $n = 50$ by letting $X \sim \text{Uniform}(0, 30)$ and

$$\Pr(Y = 1 \mid X = x) = \frac{1}{1 + e^{-(-5+0.25\,x)}}.$$

Generate Y accordingly (*Hint:* Use `rbinom()` in R).

(ii) Fit a logistic model in R (via `glm(..., family=binomial)`).

(iii) Print the estimated coefficients. Compare them to ($\alpha = -5$, $\beta = 0.25$). Are they close?

(iv) Plot the fitted logistic curve and the data points.

(v) Produce predictions at thresholds 0.4 and 0.5, and create a small confusion matrix for each. Which threshold yields more false negatives vs. false positives?

(vi) Briefly discuss how a larger sample size might improve the parameter estimates.

(d*) (5 bonus points) Suppose HR also records whether a new hire has a professional certificate, say $C \in \{0, 1\}$.

(i) Propose how to include C in your logistic regression (along with X).

(ii) Discuss how you would interpret the coefficient of C and, potentially, any interaction $C \times X$.

(iii) Briefly outline how you might compare the performance of this extended model to your original logistic regression model.

Problem 4 (20 points in total + 5 bonus points).

Scenario: You survey smartphone users in three usage categories: *Light* ($C = 1$), *Moderate* ($C = 2$), or *Heavy* ($C = 3$). Denote each user's class by C. You measure two numeric features for each user:

$$X_1 = \text{daily screen time (hours/day)}, \quad X_2 = \text{weekly voice-call duration (minutes/week)}.$$

You plan to build a Linear Discriminant Analysis (LDA) classifier based on (X_1, X_2).

For concreteness, suppose you collect the following $n = 9$ data points:

User	X_1 (hrs/day)	X_2 (min/week)	C (class)
1	1.5	25	1 (Light)
2	2.0	50	1 (Light)
3	2.3	55	1 (Light)
4	3.0	60	2 (Moderate)
5	3.5	80	2 (Moderate)
6	4.1	75	2 (Moderate)
7	4.4	120	3 (Heavy)
8	5.2	100	3 (Heavy)
9	5.4	110	3 (Heavy)

(a) (10 points) You may use R in this problem.

(i) Using the data, estimate each class mean vector,

$$\mu_k = (\mu_{k,1}, \mu_{k,2}), \quad k = 1, 2, 3,$$

and the common covariance matrix Σ. (You may do matrix arithmetic in R for convenience.)

(ii) Write down the LDA discriminant functions for each class, assuming equal priors $\pi_k = 1/3$.

(iii) Explain how the decision boundaries are determined among the three classes, and sketch them in the (X_1, X_2) plane.

(b) (10 points) You may use R in this problem.

(i) Fit a *multinomial* logistic model (`nnet::multinom`) to the same data. Compare its decision boundaries with those from the LDA analysis above.

(ii) Suppose a new user has $(X_1 = 3.1,\ X_2 = 90)$. Which class would LDA assign them to, and which class would multinomial logistic assign? Are these predictions the same or different?

(c*) (5 bonus points) Suppose you learn that in the general population, 10% are Light, 50% are Moderate, and 40% are Heavy. That is, $\pi_1 = 0.1,\ \pi_2 = 0.5,\ \pi_3 = 0.4$.

(i) How do these priors change the LDA discriminant functions you wrote in (i)?

(ii) Illustrate how the boundaries might shift due to the fact that π_2 is larger than the others.Would you expect more or fewer classifications into class 2?

(iii) In practice, how might you decide whether to use equal or unequal priors?