STA 35C – Homework 5

Submission due: Tue, May 12 at 11:59 PM PT

Instructor: Dogyoon Song

Instructions: Upload a **single PDF file** to Gradescope via Canvas ("Homework 5" under "Assignments"). Name the file using the prefix of your UC Davis email ID and the homework number (e.g., `dgsong_hw5.pdf`). Include "STA 35C," your name, and the last four digits of your student ID on the front page. For coding problems, include your R code, relevant output, and any requested plots/tables in the PDF. No late submissions will be accepted; any submission received after the deadline will receive 0 points. For full information about submission requirements and the late submission policy, see the syllabus.

You may want to use R to solve Problem 1, Problem 2(b)–(c*), and Problem 3(c). The regular total score is **100 points**, with **up to 10 bonus points** available.

Problem 1: Cross-validation for classification (20 points total + 5 bonus points).

We will explore cross-validation for a *classification* (logistic regression) setting on a simulated dataset. Use the code below to generate `X` (numeric) and `Y` (binary):

```
set.seed(2026)
n <- 100
X <- rnorm(n)
log_odds <- -2 - 3*X
probs <- 1 / (1 + exp(-log_odds))
Y <- rbinom(n, 1, probs)
df <- data.frame(X, Y=as.factor(Y))
```

(a) **(5 points)** Draw a scatterplot of X vs. Y (color 0/1 classes).

(b) **(5 points)** Randomly split the data into 50% training and 50% test (validation) sets, and fit `glm(..., family=binomial)` on the training set. Use threshold 0.5 to classify test observations, and compute the test classification error. Repeat this for 100 random splits, plot a histogram of the 100 errors, and report their mean, minimum, and maximum.

(c) **(5 points)** Compute the LOOCV classification error using the same threshold 0.5. Compare its value to the distribution of errors from the single-split validation procedure in part **(b)**.

(d) **(5 points)** Perform k-fold CV for $k \in \{2, 3, 4, 5, 10, 20\}$, using threshold 0.5 to classify held-out observations. Report the resulting CV classification errors, then plot error versus k. Which k gives the highest and lowest estimated error? Briefly comment on any pattern you observe.

(e*) **(*5 bonus points)** Create a new dataset from the quadratic data-generating process

$$\log\big(\Pr(Y=1)/\Pr(Y=0)\big) = -2 - 3X + X^2.$$

Repeat parts **(b)**, **(c)**, and **(d)** for two candidate logistic models:

$$\text{Model 1: } Y \sim X, \qquad \text{Model 2: } Y \sim X + X^2.$$

Compare the estimated classification errors and briefly discuss what changes.

Problem 2: Subset selection (30 points total + 5 bonus points).

(a) (10 points total). *In your own words,* briefly answer or explain the following:

- **(3 points)** Why might we want to search through subsets of predictors rather than use them all?
- **(3 points)** Summarize the purpose and procedure of best subset selection in at most 5 sentences.
- **(4 points)** Compare best subset vs. forward stepwise: cost, search path, drawbacks, etc.

(b) (20 points total). Download the `Credit` dataset from the textbook's website. Use `Balance` as the response and the remaining variables as candidate predictors, and run best subset selection.

(i) (5 points) Generate a scatter plot of RSS versus subset size k *for all subsets*, stratified by k. (*Hint:* you can compute `RSS` for each model, store them in a data frame, and plot.)

(ii) (5 points) For each k, find the subset with the lowest RSS (or highest R^2). Report these subsets.

(iii) (5 points) Among these k-predictor "best" subsets of size $k = 0, \ldots, p$, use *adjusted* R^2 to pick the "overall best" model. Which subset is chosen?

(iv) (5 points) Now, select the best model using **5-fold cross-validation** error instead of adjusted R^2. Use the same folds when comparing different subset sizes. Does that yield a different final subset? Briefly discuss any difference you see.

(c*) (5 bonus points) Implement forward stepwise selection on the same `Credit` dataset.

- Which predictors are chosen along the forward stepwise path for $k = 1, 2, \ldots$?
- Compare the selected model(s) to your best subset selection results above. Are they identical or different? Briefly comment if the chosen subsets differ.

Problem 3: Regularization (30 points in total).

(a) (10 points) *In your own words,* briefly explain the following in at most 3 sentences each.

- **(4 points)** Regularization.
- **(3 points)** Ridge (ℓ_2-penalized) regression.
- **(3 points)** Lasso (ℓ_1-penalized) regression.

(b) (10 points) [JWHT21, Chapter 6, Exercise 4].

(c) (10 points) [JWHT21, Chapter 6, Exercise 9, parts (a)-(d) and (g) only]. You may want to use R to solve this problem. You can download the `College` data set from the textbook's website.

Problem 4: Multiple hypothesis testing (20 points in total).

(a) (5 points) In your own words, explain what it means to test a null hypothesis at level α and what a Type I error is, in at most 3 sentences.

Suppose that we test m hypotheses simultaneously. For this problem, assume all m null hypotheses are true, the tests are independent, and each true null is rejected with probability α (e.g., $\alpha = 0.05$).

(b) (5 points) Let the random variable A_j equal to 1 if the j-th null hypothesis is rejected, and 0 otherwise. What is the distribution of A_j? Write down its PMF.

(c) (5 points) What is the distribution of $\sum_{j=1}^{m} A_j$? Write down its PMF.

(d) (5 points) What are the mean and variance of the total number of Type I errors that we will make?

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.