STA 35C – Homework 4

Submission due: Tue, May 13 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_hw4.pdf`) to Canvas ("Homework 4" 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.

You may want to use R to solve Problem 1, Problem 2-(c),(d),(e), Problem 3-(b),(c), and Problem 4-(b).

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(123)
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 data (50% train, 50% test)and fit `glm(..., family=binomial)`. Measure classification error on the test set. Repeat for 100 random splits, plot a histogram of those 100 errors, then report mean, min, and max.

(c) **(5 points)** Compute the LOOCV classification error, and compare its value and stability (variance) to the results of single-split validation above in **(b)**.

(d) **(5 points)** Perform k-fold CV for $k \in \{2, 3, 4, 5, 10, 20\}$. Report the resulting classification errors, then plot error vs. k. Which k yields highest or lowest error? Any surprises?

(e*) **(*5 bonus points)** Add a quadratic term X^2 in data generating process:

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

Repeat **(b)**, **(c)**, and **(d)**. Compare and discuss your results.

Problem 2: The Bootstrap (25 points total + 5 bonus points).

(a) (5 points) Given a fixed sample $\{x_1, x_2, x_3, x_4, x_5\}$ having all distinct values, if we draw a bootstrap sample (with replacement) of the same size, what is the probability that we replicate the original sample *exactly* in the same order? What is the probability of replicating the original sample ignoring the order?

(b) (5 points) You flip a fair coin 10 times, observing 6 heads, 4 tails in a specific sequence:

$$\{H, T, H, H, T, H, T, T, H, H\}$$

For a size-10 bootstrap sample from these 10 tosses, find the probability of getting k heads. How does that compare to flipping a fair coin 10 times for real? Evaluate these probabilities for $k \in \{4, 5, 6, 7\}$ and note which outcome is most likely under each scenario. Are they same or different?

Recall Problem 5 from Homework 3:

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

Previously, we examined how many of the newly computed or bootstrapped slopes are contained within a single 95% confidence interval, to observe how the distribution of bootstrapped estimates differ from that of the true ones. Now, we assess 95% coverage more systematically, in a way that better aligns with the intended purpose of confidence intervals:

(c) (5 points) For each $k = 1, \ldots, 1000$:

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

Fit a linear model, get $\hat{\beta}_1^{(k)}$, form a standard 95% CI. What fraction of these intervals contain the true $\beta_1 = 2$? Is it near 0.95? If not, comment briefly.

(d) (10 points) For each of those 1000 datasets, draw 1000 bootstrap samples of size 100, re-fit to get $\{\hat{\beta}_{1j}^*\}$. Compute their standard deviation, then form a 95% CI around $\hat{\beta}_1^{(k)}$. What fraction of these cover the true slope? Is it ≈ 0.95? Discuss any discrepancy.

(e*) (*5 bonus points) Repeat **(c)**–**(d)** but now generate data under

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

Discuss any difference in coverage you observe.

Problem 3: Subset selection (25 points total + 5 bonus points).

(a) (9 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 using all).
- **(3 points)** Summarize the purpose and the procedure of the best subset selection in 3–5 sentences.
- **(3 points)** Compare best subset vs. forward stepwise: cost, search path, drawbacks, etc.

(b) (16 points total). Download the `Credit` dataset from the textbook's website and run best subset selection on it.

(i) (4 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) (4 points) For each k, find the subset with the lowest RSS (or highest R^2). Report these subsets.

(iii) (4 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) (4 points) Now, select the best model by **5-fold cross-validation** error instead of adjusted R^2. Does that yield a different final subset? Discuss any difference you see.

(c*) (5 bonus points)

- Implement forward stepwise selection on the same `Credit` dataset. Which predictors are chosen as $k = 1, 2, \ldots$?
- Compare the final selected model(s) to your best subset selection results above. Are they identical or different? Comment if the chosen subsets differ.

Problem 4: Regularization (20 points in total).

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

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

Problem 5: Multiple hypothesis testing (10 points in total).

Suppose that we test m hypotheses simultaneously, and control the Type-I error for each hypothesis at level α (e.g., 0.05). Assume that all m p-values are independent, and that all null hypotheses are true.

(a) 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.

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

(c) What is the mean and the variance of the 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.