AI5213/EC4213: Machine Learning and Deep Learning **Student ID:**

Homework 2 (Resampling, Extending Linear Framework) **Name:**

Due: Sunday, Oct 27, 11:59PM

This homework comprises a mix of conceptual problems and coding exercises. Some problems are straightforward, while others requires deeper thought. Please start early!

> Guideline for those new to data analysis using Python:
> We recommend reviewing the Lab section within each chapter (e.g., p. 267 of Ch. 6 or https://islp.readthedocs.io/en/latest/labs/Ch06-varselect-lab.html) prior to tackling the programming tasks. Additionally, find datasets and Jupyter notebooks at https://github.com/intro-stat-learning/ISLP_labs/.

Deliverables: Submit a PDF of your homework, with an appendix listing all your code, to the Gradescope assignment entitled "HW2 Write-Up". You may typeset your homework in LaTeX or Word or submit neatly handwritten and scanned solutions. Make sure each solution is on a new page, and graphs are included in the correct sections. We need each solution to be self-contained on pages of its own.

Guideline:

1. On the first page of your write-up, please sign your signature next to the following statement. (Mac Preview, PDF Expert, and Foxit PDF Reader.) We want to make extra clear the consequences of cheating.
2. On the first page of your write-up, please list students who helped you or whom you helped on the homework. (Note that sending each other code is not allowed.)
3. Please write your answers in English. Korean is not allowed. Non-Korean staff members may not be able to grade responses in Korean.
4. Please don't forget to select all the pages that are related to each question during the Gradescope submission! (Submissions that do not clearly reference the exact pages containing the solution may not be graded.)

For staff use only

Honor Code	Q1	Q2	Q3	Q4	Total
/ 5	/ 15	/ 16	/ 24	/ 40	/ 100

Honor Code [5 pts]

Declare and sign the following statement:

"I certify that all solutions in this document are entirely my own and that I have not looked at anyone else's solution. I have given credit to all external sources I consulted."

Signature:

We welcome group discussions, but the work you submit should be entirely your own. If you use any information or pictures not from our lectures or readings, make sure to say where they came from. Please note that breaking academic rules can lead to severe penalties.

(a) Did you receive any help whatsoever from anyone in solving this assignment? If your answer is 'yes', give full details (e.g. "Junho explained to me what is asked in Q2-a")

(b) Did you give any help whatsoever to anyone in solving this assignment? If your answer is 'yes', give full details (e.g. "I pointed Josh to Ch. 2.3 since he didn't know how to proceed with Q2")

(c) Did you find or come across code that implements any part of this assignment? If your answer is 'yes', give full details (book & page, URL & location within the page, etc.).

Q1. Regression with regularization [15 pts; 3 pts each]

Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 \text{ subject to } \sum_{j=1}^{p}|\beta_j| \leq s$$

for a particular value of s. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase s from 0, the training RSS will:

i. Increase initially, and then eventually start decreasing in an inverted U shape.

ii. Decrease initially, and then eventually start increasing in a U shape.

iii. Steadily increase.

iv. Steadily decrease.

v. Remain constant.

(b) Repeat (a) for test RSS.

(c) Repeat (a) for (squared) bias.

(d) Repeat (a) for variance.

(e) Repeat (a) for the irreducible error.

Q2. Splines [16 pts; 4 pts each]

Suppose that a curve $\hat{g}$ is computed to smoothly fit a set of n points using the following formula:

$$\hat{g} = \arg\min_{g} \left(\sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int \left[g^{(m)}(x) \right]^2 dx \right),$$

where $g^{(m)}$ represents the mth derivative of g (and $g^{(0)} = g$). Provide example sketches of $\hat{g}$ in each of the following scenarios.

(a) $\lambda = \infty$, $m = 0$.

(b) $\lambda = \infty$, $m = 3$.

(c) $\lambda = 0, m = 0.$

(d) $\lambda = 0, m = 3.$

Q3. Bootstrap [24 pts] ⌨

We will now consider the Boston housing data set, from the `ISLP` library.

(a) Based on this data set, provide an estimate for the population mean of `medv`. Call this estimate $\hat{\mu}$. [3 pts]

(b) Provide an estimate of the standard error of $\hat{\mu}$. Interpret this result. [3 pts]

Hint: We can compute the standard error of the sample mean by dividing the sample standard deviation by the square root of the number of observations.

(c) Now estimate the standard error of $\hat{\mu}$ using the bootstrap. How does this compare to your answer from (b)? [3 pts]

(d) Based on your bootstrap estimate from (c), provide a 95% confidence interval for the mean of medv. Compare it to the results obtained by using Boston['medv'].std() and the two standard error rule. [3 pts]

Hint: You can approximate a 95% confidence interval using the formula $[\hat{\mu} - 2SE(\hat{\mu}), \hat{\mu} + 2SE(\hat{\mu})]$.

Two standard error rule: For linear regression, the 95% confidence interval for β_1 approximately takes the form

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1). \tag{1}$$

(e) Based on this data set, provide an estimate $\hat{\mu}_{med}$, for the median value of medv in the population. [3 pts]

(f) We now would like to estimate the standard error of $\hat{\mu}_{med}$. Unfortunately, there is no simple formula for computing the standard error of the median. Instead, estimate the standard error of the median using the bootstrap. Comment on your findings. [3 pts]

(g) Based on this data set, provide an estimate for the tenth percentile of medv in Boston census tracts. Call this quantity $\hat{\mu}_{0.1}$. (You can use the np.percentile() function. [3 pts]

(h) Use the bootstrap to estimate the standard error of $\hat{\mu}_{0.1}$. Comment on your findings. [3 pts]

Q4. Cross Validation [40 pts]

Consider a classification problem with a large number of predictors. One possible strategy for analysis is:

(a) Screen the predictors to find a subset of “good” predictors that show strong univariate correlation with the class labels.

(b) Use just this subset of predictors to build a multivariate classifier.

(c) Use cross-validation to estimate the unknown tuning parameters (e.g. k in k-NN or a regularization parameter λ in Ridge), to estimate the prediction error of the final model.

Part A: Theoretical Analysis [16 pts] ☕

Is this a correct application of cross-validation? Choose either (a) or (b) to answer:

(a) If yes, explain why this is a valid application of cross-validation and describe how it helps prevent overfitting in this scenario.
   (1) Describe how the predictor screening step (step 1) might be justified
   (2) Explain how using cross-validation for parameter tuning (step 3) can help prevent overfitting
   (3) Discuss any potential benefits of this approach in terms of computational efficiency or interpretability
   (4) Propose a scenario where this method might produce reliable results, specifying: sample size and number of predictors, characteristics of the data that would make this approach more acceptable, any additional steps or precautions that should be taken.

(b) If no, describe a scenario that demonstrates why this approach is problematic. Include the following in your explanation:
   (1) Specify sample size, number of predictors, and their relationship to class labels
   (2) Outline the steps taken in the flawed analysis process
   (3) Explain the unexpected results obtained and why they occurred
   (4) Discuss what this reveals about the limitations of this cross-validation approach

Note: Regardless of whether you answer (a) or (b), make sure to justify your choice and provide a thorough explanation.

Your answer on Q4, Part A:

Part B: Case Study and Implementation [24 pts] ⌨

In this part, you will analyze a specific scenario through coding and interpret the results. Consider the following scenario inspired by Chap.5 lecture slides (page 17 of 44).

- $N = 50$ samples in two equal-size classes.
- $p = 5,000$ standard Gaussian predictors that are independent of the class labels.
- True test-error rate of any classifier should theoretically be 50%.

Implement a Python script to demonstrate both the incorrect and correct ways of performing cross-validation for this scenario. Your script should:

(a) Generate a dataset as described above. [3 pts]

(b) Implement the incorrect cross-validation method: [6 pts]

(1) Choose the 100 predictors having the highest correlation with the class labels.

(2) Build a 1-NN classifier using these 100 predictors.

(3) Estimate the prediction error using cross-validation.

(c) Implement the correct cross-validation method. [6 pts]
(Hint: The correct method should perform feature selection within each fold of the cross-validation)

(d) Run both methods multiple times (e.g., 50 repetitions) and compare their results. [3 pts]

(e) Visualize the results using histograms similar to Figure 1, showing the distribution of correlations for selected predictors in both methods. [6 pts]

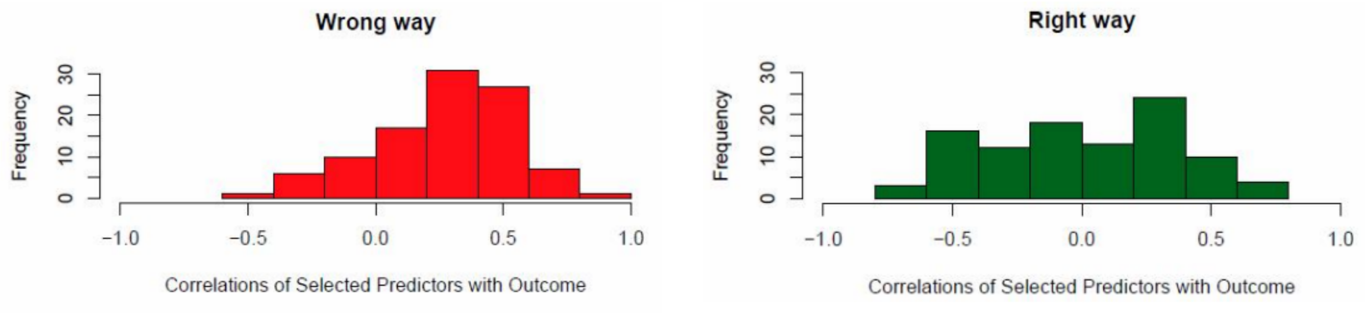

Figure 1: Expected histogram: Cross-validation the wrong and right way: Histograms show the correlation of class labels, in 10 randomly chosen samples, with the 100 predictors chosen using the incorrect (red) and correct (green) versions of cross-validation.