Yeabin Moon, Ph.D.
BUS 212A-2 Spring 2023
February 27, 2023

Homework 4

Submission Instructions

1) You have to use Jupyter Notebook
2) Click the Save button at the top of the Jupyter Notebook.
3) Select Cell → All Output → Clear. This will clear all the outputs from all cells (but will keep the content of all cells).
4) Select Cell → Run All. This will run all the cells in order, and will take several minutes.
5) Once you've rerun everything, select File → Download as → PDF via LaTeX (If you have trouble using "PDF via LaTex", you can also save the webpage as pdf. Make sure all your solutions especially the coding parts are displayed in the pdf, it's okay if the provided codes get cut off because lines are not wrapped in code cells).
6) Look at the PDF file and make sure all your solutions are there, displayed correctly. The PDF is the only thing your graders will see!
7) Submit your PDF on Latte.

Question 1. We now review k-fold cross-validation.

1) Explain how k-fold cross-validation is implemented.
2) What are the advantages and disadvantages of k-fold cross-validation relative to:
   - The validation set approach?
   - LOOCV

Question 2. What is the curse of dimensionality?

Question 3. What are the main motivations for reducing a dataset's dimensionality? What are the main drawbacks?

Question 4. How can you evaluate the performance of a dimensionality reduction algorithm on your dataset?

Question 5. We will now perform cross-validation on a simulated data set.

1) Generate a simulated data set as follows:

```
import numpy as np
np.random.seed(1)
x = np.random.normal(0,1,100)
y = x - 2 * x ** 2 + np.random.normal(0,1,100)
```

   In this data set, what is n and what is p? Write out the model used to generate the data in equation form.
2) Create a scatterplot of X against Y. Comment on what you find.
3) Set a random seed, and then compute the LOOCV errors that result from fitting the following four models using least squares:
   (a) $Y = \beta_0 + \beta_1 X + \epsilon$
   (b) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$
   (c) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$
   (d) $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$
4) Repeat (3) using another random seed, and report your results. Are your results the same as what you got in (3)? Why?

5) Which of the models in (3) had the smallest LOOCV error? Is this what you expected? Explain your answer.
6) Comment on the statistical significance of the coefficient estimates that results from fitting each of the models in (3) using least squares. Do these results agree with the conclusions drawn based on the cross-validation results?

Question 6. In this question, you need to predict the number of applications received using the other variables in the College data set

1) Split the data set into a training set and a test set.
2) Fit a linear model using least squares on the training set, and report the test error obtained.
3) Fit a ridge regression model on the training set, with α chosen by cross-validation. Report the test error obtained
4) Fit a lasso model on the training set, with α chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates
5) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these three approaches?