Instructions

- This homework assignment is worth 65 points.
- Please submit a **.ipynb** file to Blackboard.
- **Please strive for clarity and organization.**
- **Due Date: September 30, 2022 by 11:59 pm.**

Exercise 1

(5 points) You are given a dataset having more variables than observations. Assuming that there seems to be a linear relationship between the target variable and the input variables in the dataset, why ordinary least squares (OLS) is a bad option to estimate the model parameters? Which technique would be best to use? Why?

Exercise 2

(5 points) For Ridge regression, if the regularization parameter, λ, is equal to 0, what are the implications?

(a) Large coefficients in the linear model are not penalized.

(b) Overfitting problems are not accounted for.

(c) The objective function is the same as ordinary least squares objective function.

(d) All of the above.

(e) (a) and (b)

(f) (a) and (c)

(g) (b) and (c)

(h) None of the above.

Exercise 3

(5 points) For Lasso Regression, if the regularization parameter, λ, is very high, which options are **true**? Select all that apply.

(a) Can be used to select important features of a dataset.

(b) Shrinks the coefficients of less important features to exactly 0.

(c) The loss function is as same as the ordinary least square loss function.

(d) The objective function is as same as the Ridge Regression objective function.

(e) All of the above.

(f) (a) and (b)

(g) (a) and (c)

(h) (a) and (d)

(i) (b) and (c)

(j) (b) and (d)

(k) (c) and (d)

Exercise 4

An important theoretical result of statistics and Machine Learning is the fact that model's generalization error can be expressed as the sum of two very different errors:

- **Bias:** This part of the generalization error is due to wrong assumptions, such as assuming that the data is linear when it is actually quadratic. A high-bias model is most likely to under-fit the training data.
- **Variance:** This part is due to the model's excessive sensitivity to small variations in the training data. A model with many degrees of freedom (such as a high-degree polynomial model) is likely to have high variance and thus overfit the training data.

(5 points) Suppose you are using Ridge Regression and you notice that the training error and the validation error are almost equal and fairly high. Would you say that the model suffers from high bias or high variance? Should you increase the regularization parameter, λ, or reduce it?

Exercise 5

Consider the `CarPrice_Assignment.csv` data file. This data is public available on the Kaggle website, and has information on cars (characteristics related to car dimensions, engine and more). The goal is to use car information to predict the price of the car. **In Python**, answer the following:

(a) (5 points) Load the data file to you S3 bucket. Using the pandas library, read the csv data file and create a data-frame called `car_price`.

(b) (15 points) Using the `wheelbase`, `enginesize`, `compressionratio`, `horsepower`, `peakrpm`, `citympg`, and `highwaympg` as the predictor variables, and `price` is the target variable. Do the following:

(1) Split the data into train (80%) and test (20%)

(2) Using the train dataset:

(i) Estimate the optimal lambda using default values for lambda in scikit-learn and 5-folds. Make sure to normalize the data (`normalize = True`).

(ii) Perform LASSO as a variable selector (using the optimal lambda from previous step (i)). Make sure to normalize the data (`normalize = True`).

Repeat steps (1) and (2) 1000 times. Store the estimated model coefficients of each iteration in a data-frame. Remove the variables, whose estimated coefficients is 0 more than 500 times, from the training and testing datasets.

(c) (5 points) Split the data into train (80%) and test (20%). Then, normalize the inputs variables of the train and test datasets using the L2 normalization. That is, for each input variable subtract the mean of that variable, then divide by the L2-norm of that variable.

(d) (5 points) Using the train dataset, build a linear regression model. After that, use this model to predict on the test dataset. Report the MSE of this model.

(e) (10 points) Using the train dataset, build a Ridge regression model as follows:

(i) Using the train dataset, estimate the optimal lambda from the following set [0.001, 0.01, 0.1, 1, 10, 100] and using 5-folds.

(ii) Repeat (i) 100, store the optimal lambda of each iteration.

Using the most common lambda of the 100 optimal lambdas and the train dataset, build a Ridge regression model. After that, use this model to predict on the test dataset. Report the MSE of this model.

(f) (5 points) Using the results from parts (d) and (e), what model would you use to predict car prices? Explain.