Homework 5:
Ridge Regression and K-Fold Cross Validation

Submit your assignments on Gradescope.

Please name your coding assignment as 'HW5.py'.

Use the provided Python template file, and complete the functions ONLY. (DO NOT edit function definitions, code outside the function, or use other libraries).

In this homework, you have been provided with the prepare_data() and MSE() functions to help you with your implementation. The functions that you will be graded on have pre-initialized variables as a guide for the data types allowed in the return statements.

This is a coding assignment.

Linear Regression

In part one, you will implement the linear regression model you saw last week in homework 5. The functions for this linear regression model will be the same as in the previous homework assignment.

In this part, you will first implement the Ridge Regression model, then use K-Fold cross-validation to train 10 different models.

1. *Split Function*

Implement the split function for this assignment. The function takes in the data and a percentage. You will be splitting the dataset into training and test data based on the training data percentage.

2. *Ridge Regression* (L2-Regularization)

As seen in the lecture, L2-Regularization with linear regression is called Ridge Regression. You will be implementing a ridge regression model in this problem. The default parameters for this model are *learning_rate = .00001*, *iterations = 1000*, and *penalty = 1*. Use the default parameters when implementing your model.

Use page 63 in Su_24_W3_B_OverUnderFit_Validation

3. *K-Fold Cross Validation*

In this problem, you will be implementing K-Fold Cross Validation. There will be 10 folds. In each fold, the training dataset will be split into 10 parts. One part will be used for validation, then the rest will be used for training the model. A diagram has been provided below to illustrate this process. Refer back to the lecture slides for more information on K-Fold CV.

Final evaluation:
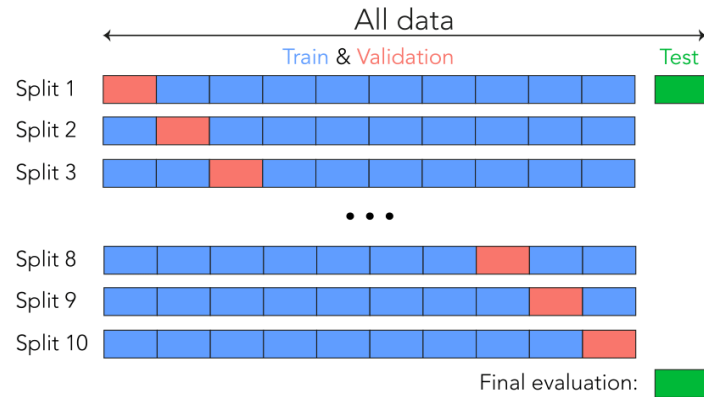
After completing the K-Fold validation, you will select the best model from the 10 you have trained and return its mean square error.

This is a report assignment.

There will be **separate graphs** with their respective lines of regression. You will submit one graph for your linear regression model, one for Ridge regression, and one for your *best-performing* ridge regression model with k-fold validation. The code for the plots has already been filled in for you. This problem will be graded based on the correctness of your models.

What was the lowest resulting MSE value? Compare the results for your linear regression model and ridge regression model. Which model was the best for this data?

Discuss the use of Regularization and k-fold validation; and how they affect the training.