Introduction to Machine Learning
Lab Assignment 1: Regression

Summary: In class, we have discussed different regression models. For this assignment, you will demonstrate you understand how to apply, evaluate, and analyze these models.

1. **Synthetic Data: Comparative Analysis & Learning Curves [40 points]**

   (a) Generate a dataset by sampling 1,200 values from a cubic function with noise added. Set values to range from 0 to 1. (Code)

   (b) Create a 70/30 train/test split of the dataset. (Code)

   (c) Plot learning curves for the following two regression models: linear and polynomial (to the 4th degree). Vary the amount of training data by increments of at most 84 (i.e., at least 10 training sizes from the 840 training examples), evaluate using mean squared error, and remember to show curves for both the training and test datasets. (Code and Write-up)

   (d) Write a discussion analyzing and comparing the two models. Explain which model performs better and why. Also, address which models you think are underfitting versus overfitting and explain why. Final, discuss the impact of increasing the amount of training data on both models. Your discussion should consist of two to three paragraphs. (Write-up)

2. **Real Data: Comparative Analysis & Feature Analysis [40 points]**

   (a) Load a real dataset not covered in class that is designed for the regression problem; e.g., from sklearn.datasets, Kaggle, your own data, etc. (Code)

   (b) Create a 70/30 train/test split of the dataset. (Code)

   (c) Train and evaluate the predictive performance for each of the following regression models: linear, ridge, lasso, and polynomial (to the 4th degree). Evaluate using mean squared error. Report all values in a single table. (Code and Write-up)

   (d) For the top-performing regression model, report the learned feature coefficients. (Code and Write-up)

   (e) Write a discussion analyzing and comparing the four models. Explain which model(s) perform the best/worst and why you think this occurs. Also discuss which features are most predictive for the top-performing model and why this may be. Your discussion should consist of two to three paragraphs. (Write-up)

3. **Real Data: Tuning Hyperparameters for Regularized Models [20 points]**

   (a) Use the same training and test datasets from step 2c.

   (b) Plot performance curves for the following regression models when varying α: ridge and lasso. Vary α to have at least 10 values and evaluate using mean squared error. (Code and Write-up)

   (c) Write a discussion analyzing and comparing the two types of regression models. Explain the impact of the α parameter. Also, discuss which model performs the better and why you think this occurs. The discussion should consist of two to three paragraphs. (Write-up)

How to Submit Lab Assignment 1: Please submit a pdf that provides hyperlinks to your code and answers to the questions, as deemed appropriate for the task. The pdf file should be named using your first and last name; i.e., firstname_lastname.pdf. The material you submit must be your own.