COEN 140 Machine Learning and Data Mining

Lab Assignment #3: Linear Regression and Ridge Regression

Guideline: Submit a pdf report to Camino. Also submit all the source code needed to generate the results as a separate zip file to Camino.

Implement Linear Regression and Ridge Regression in Python. Please do not use any machine learning library.

Training data: crime-train.csv
Test data: crime-test.csv
A description of the variables: communities.names
Use the above datasets provided without modifications. Do not rename or alter the file's contents.

Load Datasets

Load the training and test data from crime-train.csv and crime-test.csv, using pandas.

```
import numpy as np
import pandas as pd
df_train = pd.read_csv("crime-train.csv")
df_train_np = pd.DataFrame(df_train).to_numpy()
df_test = pd.read_csv("crime-test.csv")
df_test_np = pd.DataFrame(df_test).to_numpy()
```

The data consist of local crime statistics for 1,994 US communities. The target t is the crime rate. The name of the target variable is *ViolentCrimesPerPop*, and it is held in the first column of df_train_np and df_test_np. There are 95 features. These features include possibly relevant variables such as the size of the police force or the percentage of children that graduate high school. The data have been split for you into a training and test set with 1,595 and 399 samples, respectively.

Problem 1
Implement the linear regression model. Use all training samples to train the model, and use the Least Squares method to find the solution. Note that the model weights should have a bias term w_0. Compute the MSE value on the training data and test data, respectively.

Report the following:
1.a MSE values of both training data and test data
1.b The first 10 elements of the optimal weight vector
1.c The predicted crime rate of the first 10 test samples

Problem 2
2.a Repeat the same experiments in Problem 1, using only the first 100 training samples to train the linear regression model. Report the MSE values of both training data and test data, and report the first 10 elements of the optimal weight vector.

2.b Implement Ridge Regression with $\lambda = 100$ to find the optimal model weights, again using only the first 100 training samples to train the model. Report the MSE values of both training data and test data, and report the first 10 elements of the optimal weight vector.

2.c Compare and comment on the results you obtained in **2.a** and **2.b**.

Demo and explain to TA (10%):

For both Problem 1 and 2, show

1. How you construct matrices **X**_train and **X**_test, with rows as the samples.
2. How you accommodate the bias term w_0 (i.e. what did you do such that your weights have a bias term?)
3. How you calculate the optimal weight vector **w**
4. How you obtain the predicted crime rates **y**, for test samples

Grading:
Demo: 10%
Report: 60%
Source Code: 30%