COEN 140 Machine Learning and Data Mining

Lab Assignment #4: Gradient Descent

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.

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 Linear Regression to predict the crime rates with the vector-version gradient descent algorithm. Please do not use any machine learning library. Use all training samples to train the model. Note that the model weights should have a bias term w_0. Initialize the weights as zeros. Set the learning rate as $\eta = 10^{-5}$, and maxIter=10,000.

Define "converging" as the maximum absolute difference of any weight element between current iteration and the previous iteration is less than $\varepsilon = 10^{-5}$.

Report the following:

1.a At which iteration does your algorithm exit?

When the algorithm exits, what are the
1.b MSE values of both training data and test data?

1.c The first 10 elements of the returned weight vector?

1.d The predicted crime rates of the first 10 test samples?

1.e Compare the results of **1.b**, **1.c**, and **1.d** to the results of Problem 1 in Lab assignment 3, what do you observe?

1.f Plot the error function of the training samples $E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N_{train}}\{\mathbf{x}_n^T\mathbf{w} - t_n\}^2$ as a function of iteration number. Comment on the result.

Demo and explain to TA (10%):

For Problem 1, show in a certain iteration

1. How you update the gradient vector

2. How you update the weight vector

3. How you calculate the error function value

4. How you check the stopping criterion

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