CS 383 - Machine Learning

Assignment 3 - Linear Regression

Introduction

In this assignment you will explore gradient descent and perform linear regression on a dataset using cross-validation to analyze your results.

As with all homeworks, you cannot use any functions that are against the "spirit" of the assignment. For this assignment that would mean an linear regression functions. You *may* use statistical and linear algebra functions to do things like:

- mean
- std
- cov
- inverse
- matrix multiplication
- transpose
- etc...

And as always your code should work on any dataset that has the same general form as the provided one.

Grading

Although all assignments will be weighed equally in computing your homework grade, below is the grading rubric we will use for this assignment:

Part 1 (Theory)	20pts
Part 2 (Gradient Descent)	10pts
Part 3 (Closed-form LR)	40pts
Part 4 (S-Folds LR)	20pts
Report	10pts
TOTAL	100

Table 1: Grading Rubric

Datasets

Fish Length Dataset (x06Simple.csv) This dataset consists of 44 rows of data each of the form:

1. Index
2. Age (days)
3. Temperature of Water (degrees Celsius)
4. Length of Fish

The first row of the data contains header information.

Data obtained from: http://people.sc.fsu.edu/ jburkardt/datasets/regression/regression.html

1 Theory

1. (10pts) Consider the following data:

$$\begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \\ -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix}$$

   Compute the coefficients for the linear regression using least squares estimate (LSE), where the second value (column) is the dependent variable (the value to be predicted) and the first column is the sole feature. Show your work and remember to add a bias feature. Compute this model using **all** of the data (don't worry about separating into training and testing sets).

2. For the function $g(x, y) = (x+y-2)^2$, where x and y are a single valued variables (not vectors):

   (a) What are the partial gradients, $\frac{\partial g}{\partial x}$ and $\frac{\partial g}{\partial y}$? Show work to support your answer. (4pts)

   (b) Create a 3D plot of x vs y, vs $g(x, y)$ use a software package of your choosing. (4pts)

   (c) Based on your plot, what are the values of x and y that minimize $g(x, y)$? (2pts)

2 Gradient Descent

In this section we want to visualize the gradient descent process on the function $g(x, y) = (x+y-2)^2$. You should have already derived (pun?) the gradient of this function in the theory section. To bootstrap the process, initialize $x = 0, y = 0$ and terminate the process when the change in the $g(x, y)$ from one iteration to another is less that 2^{-32}. In addition, we'll use a learning rate of $\eta = 0.1$.

In your report you will need

1. Plot iteration vs $g(x, y)$
2. Plot iteration vs x
3. Plot iteration vs y

3 Closed Form Linear Regression

Download the dataset *x06Simple.csv* from Blackboard. This dataset has header information in its first row and then all subsequent rows are in the format:

$$Index, Age, Temp, LengthofFish$$

Using the $Temp$ and $LengthofFish$ as your features, we want to predict the Age of a fish.

Your code should work on any CSV data set that has the first column be header information, the first column be some integer index, the second column as the target value, then D columns of real-valued features.

Write a script that:

1. Reads in the data, ignoring the first row (header) and first column (index).
2. Randomizes the data
3. Selects the first 2/3 (round up) of the data for training and the remaining for testing
4. Computes the closed-form solution of linear regression
5. Applies the solution to the testing samples
6. Computes the *root mean squared error* (RMSE): $\sqrt{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}$. where $\hat{Y}_i$ is the predicted value for observation X_i.

Implementation Details

1. Seed the random number generate with zero prior to randomizing the data
2. Don't forget to add in the bias feature!

In your report you will need:

1. The final model in the form $y = \theta_0 + \theta_1 x_1 + ...$
2. The root mean squared error.

4 S-Folds Cross-Validation

Cross-Validation is a technique used to get reliable evaluation results when we don't have that much data (and it is therefore difficult to train and/or test a model reliably).

In this section you will do S-Folds Cross-Validation for a few different values of S. For each run you will divide your data up into S parts (folds) and test S different models using S-Folds Cross-Validation and evaluate via root mean squared error. In addition, to observe the affect of system variance, we will repeat these experiments several times (shuffling the data each time prior to creating the folds). We will again be doing our experiment on the provided fish dataset to predict the age of a fish.

Write a script that:

1. Reads in the data, ignoring the first row (header) and first column (index).
2. 20 times does the following:
   (a) Randomizes the data
   (b) Creates S folds.
   (c) For $i = 1$ to S
      i. Select fold i as your testing data and the remaining $(S-1)$ folds as your training data
      ii. Train a closed-form linear regression model
      iii. Compute the squared error for each sample in the current testing fold
   (d) You should now have N squared errors. Compute the RMSE for these.
3. You should now have 20 RMSE values. Compute the mean and standard deviation of these. The former should give us a better "overall" mean, whereas the latter should give us feel for the variance of the models that were created.

Implementation Details

1. Don't forget to add in the bias feature!
2. Set your seed value at the very beginning of your script (if you set it within the 20 tests, each test will have the same randomly shuffled data!).

In your report you will need:

1. The average and standard deviation of the root mean squared error for $S = 3$ over the 20 different seed values..
2. The average and standard deviation of the root mean squared error for $S = 5$ over the 20 different seed values.
3. The average and standard deviation of the root mean squared error for $S = 20$ over 20 different seed values.
4. The average and standard deviation of the root mean squared error for $S = N$ (where N is the number of samples) over 20 different seed values. This is basically *leave-one-out* cross-validation.

Submission

For your submission, upload to Blackboard a single zip file with no spaces in the file or directory names and contains:

1. PDF Writeup
2. Source Code
3. readme.txt file

The readme.txt file should contain information on how to run your code to reproduce results for each part of the assignment.

The PDF document should contain the following:

1. Part 1: Your solutions to the theory questions.
2. Part 2: Your three plots.
3. Part 3:
   (a) Final Model
   (b) RMSE
4. Part 4:
   (a) Average and standard deviations of RMSEs for different cross-validations.