CS 613 - Machine Learning

Assignment 1 - Linear Regression

Introduction

In this assignment you will perform linear regression on a dataset and 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

Part 1 (Theory)	10pts
Part 2 (Closed-form LR)	30pts
Part 3 (S-folds LR)	30pts
Part 4 (Local LR)	30pts
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. Consider the following supervised *training* dataset:

$$X = \begin{bmatrix} -2 \\ -5 \\ -3 \\ 0 \\ -8 \\ -2 \\ 1 \\ 5 \\ -1 \\ 6 \end{bmatrix}, Y = \begin{bmatrix} 1 \\ -4 \\ 1 \\ 3 \\ 11 \\ 5 \\ 0 \\ -1 \\ -3 \\ 1 \end{bmatrix}$$

   (a) Compute the coefficients (weights and bias) for linear regression using the direct method. Show your work by setting up the equations, and remember to add a bias feature. Since we're doing a direct solution, there is no need to zscore it (6pts).

   (b) Using your learned model in the previous part, what are your predictions, $\hat{Y}$, for the training data (2pts)?

   (c) What is the RMSE and MAPE for this training set based on the model you learned in the previous part (2pts)?

2 Closed Form (Direct) Linear Regression

Download the dataset *x06Simple.csv* from Blackboard. This dataset has header information in its first row and then all subsequent rows are observations. We will attempt to predict the age of the fish using the water temperature and length of the 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 being 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. Shuffles the rows of the data
3. Selects the first 2/3 (round up) of the data for training and the remaining for validation.
4. Computes the linear regression model using the direct solution.
5. Applies the learned model to the validation samples.
6. Computes the *root mean squared error* (RMSE) and mean absolute percent error (MAPE) for the training and validation sets.

Implementation Details

1. For reproducability, seed the random number generate with zero prior to shuffling the data.
2. Don't forget to add in a bias feature!

In your report you will need:

1. The final model in the form $y = w_0 + w_1 x_1 + ...$
2. The values for RMSE and MAPE for the training and validation sets.

3 S-Folds Cross-Validation

Cross-Validation is a technique used to use more data for training a system while maintaining a reliable validation score.

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 build 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.

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) Seeds the random number generator to the current run (out of 20).
   (b) Shuffles the rows of the data
   (c) Creates S folds.
   (d) For $i = 1$ to S
      i. Select fold i as your validation data and the remaining $(S - 1)$ folds as your training data.
      ii. Train a linear regression model using the direct solution.
      iii. Compute the squared error for each sample in the current validation fold
   (e) 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 a bias feature!

In your report you will need:

1. The average and standard deviation of the root mean squared validation error for $S = 4$ over the 20 different runs.
2. The average and standard deviation of the root mean squared validation error for $S = 11$ over the 20 different runs.
3. The average and standard deviation of the root mean squared validation error for $S = 22$ over 20 different runs.
4. The average and standard deviation of the root mean squared validation error for $S = N$ (where N is the number of samples) over 20 different runs. This is basically *leave-one-out* cross-validation.

4 Locally-Weighted Linear Regression

Next we'll do locally-weighted closed-form linear regression.

Write a script to:

1. Read in the data, ignoring the first row (header) and first column (index).
2. Shuffles the rows of the data
3. Select the first 2/3 (round up) of the data for training and the remaining for validation.
4. Then for each *validation sample*
   (a) Compute the necessary distances of the validation sample to the training data in order to establish your weight matrix.
   (b) Use the weight matrix to compute a local model via the direct method.
   (c) Evaluate the validation sample using the local model.
   (d) Compute the squared error of the validation sample.
5. Computes the RMSE and MAPE over the validation data.

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!
3. Use the L1 distance when computing the distances $d(a, b)$.
4. Let $k = 1$ in the similarity function $\beta(a, b) = e^{-d(a,b)^2/k^2}$.
5. Use *all* training instances when computing the local model.

In your report you will need:

1. The RMSE and MAPE for the validation data.

Submission

For your submission, upload to Blackboard a single zip file containing:

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:
   (a) Your solutions to the theory question
2. Part 2:
   (a) Final Model
   (b) RMSEs and MAPEs
3. Part 3:
   (a) Mean and Standard Deviations of validation RMSEs for different values of S.
4. Part 4:
   (a) RMSE and MAPE for validation set.