CS 383 - Machine Learning

Assignment 3 - Closed Form Linear Regression
Winter 2017

Introduction

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

You may **not** use any function from the Matlab ML library in your code. 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)	10pts
Part 2 (Closed-form LR)	20pts
Part 3 (S-Folds LR)	10pts
Report	5pts
TOTAL	45

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 data:

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

   (a) Compute the coefficients for the linear regression using global least squares estimate (LSE) where the second value is the dependent variable (the value to be predicted). Show your work and remember to add a bias feature and to standardize the features (10pts).

2 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:

$$ROWId, x_{i,1}, x_{i,2}, y_i$$

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

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. Standardizes the data (except for the last column of course) using the training data
5. Computes the closed-form solution of linear regression
6. Applies the solution to the testing samples
7. 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 obseration 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 offset 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.

RMSE:	853.38

Table 2: Closed Form Regression Evaluation

3 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 divide your data up into 5 parts and train/test 5 different models using the 5-Folds Cross-Validation. We can then look at the root mean squared error using all the errors.

Write a script that:

1. Reads in the data, ignoring the first row (header) and first column (index).
2. Randomizes the data
3. Creates S folds (for our purposes $S = 5$, but make your code generalizable, that is it should work for any legal value of S). NOTE: While making theses folds yourself should be relatively easy, Matlab does have a function called *cvpartition* that does exactly this for us.
4. For $i = 1$ to S
   (a) Select fold i as your testing data and the remaining $(S - 1)$ folds as your training data
   (b) Standardizes the data (except for the last column of course) based on the training data
   (c) Train a closed-form linear regression model
   (d) Compute the squared error for each sample in the current testing fold
5. Compute the RMSE using all the errors.

Implementation Details

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

In your report you will need:

1. The root mean squared error.

RMSE	627.67

Table 3: Evaluation Using 5-Fold Cross Validation

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) RMSE
3. Part 3:
   (a) RMSE