2019 Fall EE5183 FinTech - Homework 1

Machine Learning Basics: Regression

Due: October 2019

INSTRUCTIONS

1. In this homework, datasets from Student Performance Data Set from UCI Machine Learning Repository(https://archive.ics.uci.edu/ml/datasets/student+performance) are utilized to build regression models. Those two datasets were combined and shuffled into a single dataset. The last column, cat, represents the classes the students belong to.

2. **Please use train.csv to train/test your models and report regression results generated from the hidden test set,*test_no_G3.csv.*** The following columns should be included as predictors: *school, sex, age, famsize, studytime, failures, activities, higher, internet, romantic, famrel, freetime, goout, Dalc, Walc, health, absences,* and you need to transform *binary* columns to one-hot encoding vectors. The target is *G3*.

3. Name your source code that contains your *main* function as *hw1_StudentID.py* and your report as *hw1_StudentID.pdf*. You should provide your predictions for hidden test set following the format of example submissions (*StudentID_1.txt*). Please use *tabs* to separate IDs and predictions.

4. **You should write your own codes independently. Plagiarism is strictly prohibited.**

PROBLEMS

1. (80%)Linear Regression

   (a) (20%) Split *train.csv* into training set (80%) and test set (20%). Both the training and test set should be normalized by subtracting the (column-wise) means of training set from them and then divided by the (column-wise) standard deviations of the training set. **Please elaborate on how you obtain your training and test sets in your report.** Notice that you should use identical training and test sets for (b) - (e).

   (b) (10%) Implement a linear regression model *without* the bias term to predict *G3*. Use pseudo-inverse to obtain the weights. Record the root mean squared error (RMSE) of the test set.

   (c) (10%) Regularization is often adopted to avoid over-fitting. Regularization for linear regression model by adding an additional term in your function $\mathbf{J(w)}$:

   $$\mathbf{J(w) = MSE_{train} + \frac{\lambda}{2} w^T w}$$

   Implement a *regularized* linear regression model without the bias term where $\lambda = 1.0$. **Please describe how to find the optimal weights with maximum likelihood criterion in your report**. Record the RMSE of the test set.

   (d) (10%) Repeat (c) but *include* the bias term in your model.

   (e) (10%) Follow *Example: Bayesian Linear Regression* in the textbook (Chapter 5) and implement a Bayesian linear regression model with the bias term. Let $\mu_{\mathbf{0}} = \mathbf{0}$ and $\mathbf{\Lambda_0} = \frac{\mathbf{1}}{\alpha}\mathbf{I}$ in (5.78) where $\alpha = 1.0$. Use the mean of the posterior as weights for your model. Record the RMSE of the test set.

   (f) (20%) Plot the ground truth (real *G3*) versus all predicted values generated by models (b) - (e) as examplied in Figure 1. **Please compare the RMSEs and predicted *G3* values in your report. Also, please explain mathematically why predicted *G3* values are closer to the ground truth for (d) and (e).**

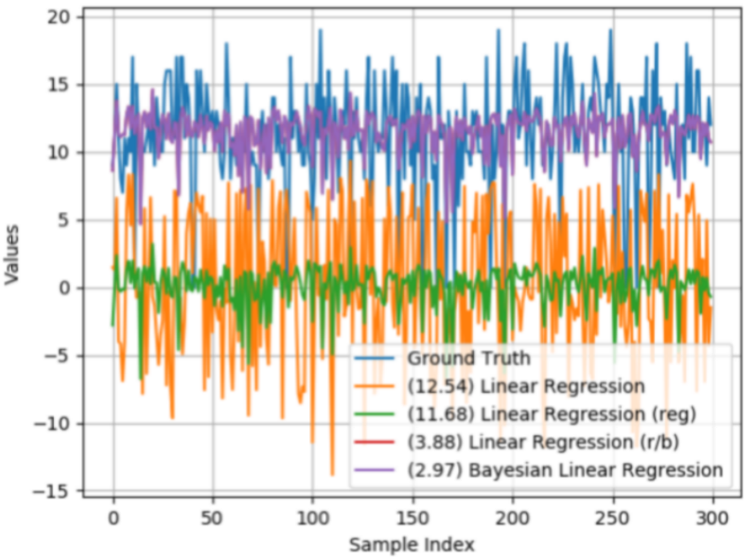

Figure 1: Regression result comparison.

2. (20%) Hidden Test Set

   (a) Apply the model from 1. (d) to *test_no_G3.csv* and save your results as *StudentID_1.txt*. You are allowed to tune α.