FinTech Homework 1

Wu, Bo-Run (r08942073)

24th October 2020

1. Linear Regression

(a) For splitting training and testing sets, I use the numpy package to shuffle the indexes, by calling *random.permutation* function. Then, I choose the first 80% indexes to be training set, and else to be testing set. Also, I use the pandas package to transform the binary columns to one-hot encoding vectors by calling *get_ dummies()*, and normalize the data by apply normalize function column by column.

(b) RMSE: 12.14.

(c) RMSE: 11.95. The loss function that add the regularization term is

(d) $\frac{1}{2}||\boldsymbol{X^{(train)}w} - \boldsymbol{Y}||_2^2 + \frac{\lambda}{2}\boldsymbol{w^T w}$ and we let the first derivative of the loss function to be zero. $\boldsymbol{X^{(train)}}(\boldsymbol{X^{(train)}w} - \boldsymbol{Y})+$ $\lambda \boldsymbol{w} = 0$ and we can get the $\boldsymbol{w} = (\boldsymbol{X^{(train)T}X^{(train)}} + \boldsymbol{\lambda})^{-1}\boldsymbol{X^{(train)T}Y}$ as the closed form solution.

(e) RMSE: 3.68.

(f) RMSE: 3.68.

(g) The figure 1 is the predicted G3 values and RSMEs for (b) - (e). The reason why (d) and (e) are more closer to the Ground Truth is that the $\boldsymbol{w}$ from (d) and (e) are the same under the assumption of (e). As we can see in the figure 1, (b) with the regularization has the improvement that variance between the predicted values get smaller, and the (c) with both regularization and bias shift the (b) curves with constants letting (c) to be more closer to the Ground Truth.

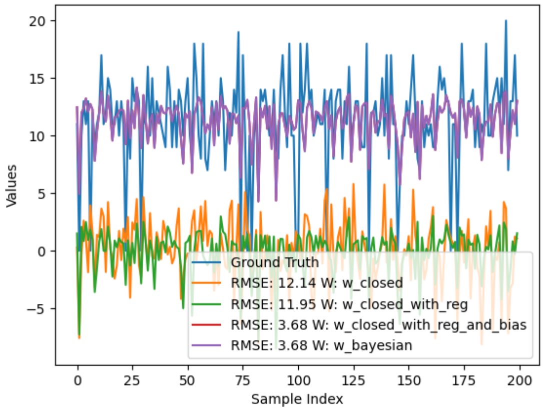

Figure 1: Student Line Chart

(h) The result is written in the r08942073_1.txt

2. Census Income Data Set

As we change the data set to census income data set, the problem become logistic regression instead of linear regression. The first problem that we encounter is the missing value in this data set. Fortunately, the missing value are all in the categorical columns, e.g. the columns without continuous values, so I specify another label for the columns that have missing value to represent the missing value. Next, I solve this problem by using linear regression to minimize the root-mean-square error between the predicted results and the ground truths. Here, I transform the ground truths into 0 and 1, representing two different level of income. The solution is not the traditional way of solving logistic regression, but we can see the differences from Problem 1. As we can see in the figure 2, the regularization didn't help the model to predict, and the bias did.

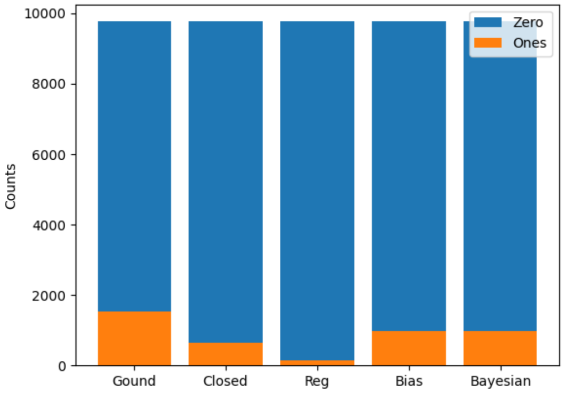

Figure 2: Census Bar Chart