CSCI-SHU 360 Machine Learning
Solution to homework 3

Josiah Li yl11912@nyu.edu

May 5, 2025

1 Programming Problem: Random Forests

1.1

Table 1: Train and Test RMSE for Different Models

Model	**Train RMSE**	**Test RMSE**
Random Forest	3.19	3.86
Linear Regression	4.82	5.20
Ridge Regression (α=0.5)	4.63	4.92

We can see that the random forest model performs the best compared to the previous models.

1.2

Table 2: Train and Test Accuracy for credit risk and breast canser prediction.

Model	**Train Accuracy**	**Test Accuracy**
Credit Risk	84.29%	74.67%
Breast Cancer	98.24%	96.50%

2 Programming Problem: Gradient Boosting Decision Trees

2.1

For the implementation that we first iterate m features, and perform a linear search on the current sample set to find the best threshold, we need to do the sorting based on the features. Costing $\mathcal{O}(mn \log n)$. After that, what really matters is the linear search we do on the tree. As there are d layers, and the total sample number on each layer will be n. Thus, on each layer, the cost of a linear search will be $\mathcal{O}(mn)$ in total. Thus, the total time complexity will be $\mathcal{O}(mn \log n + mnd)$.

2.2

According to Question 1, we know that the most expensive operation in GBDT training is scanning all candidate thresholds on every feature to find the best split. To alleviate this cost, we discretise each continuous feature into a small, fixed number of buckets (histogram bins). The idea is to transform an $O(n)$ scan over raw values into an $O(K)$ scan over bucket statistics. Thus, after we done this, the total time complexity will be reduced to $\mathcal{O}(mKd)$.

2.3

We can see that finding the splits for the features can be done in a parallel way. Here we modify the function "find_best_decision_rule".

2.4

Table 3: Train and Test RMSE for Different Models

Model	**Train RMSE**	**Test RMSE**
Random Forest	3.19	3.86
GBDT	1.59	3.59
Linear Regression	4.82	5.20
Ridge Regression (α=0.5)	4.63	4.92

2.5

Table 4: Train and Test Accuracy for credit risk and breast canser prediction by using GBDT.

Model	**Train Accuracy**	**Test Accuracy**
Credit Risk	89.14%	76.33%
Breast Cancer	99.75%	95.91%

2.6

By comparing the previous tables, we can see that most of the time, we can see that GBDT has at least the same or better performance on the datasets. This is probably because GBDT is correcting the mistakes of the tree to the direction that minimize the mistakes, while RF corrects the mistakes by randomly sampling from the dataset, which is trying to correct the mistakes by a random pattern.