EP2420 Project 1 Advanced Project

Report 1 Task 1, 2.1

Name: Wenqi Rong

Date: 2020.10.31

Project Data: KV periodic

Task 1:

1. **Data set info**

The data size of feature and target files are shown in Figure 1, since both X and Y have a column for timestamp, the number of samples is 28962, the number of features is 1751, the number of types of targets is 2.

```
The shape of X.csv is:
(28962, 1752)
The shape of Y.csv is:
(28962, 3)
```

Figure 1: Data size of the feature and target files

2. **Randomly selected features' explanations**

The selected features and their explanations are shown in Table 1.

Feature	Explanation
usr	percentage of CPU utilization that occurred while executing at the user level (application)
sys	percentage of CPU utilization that occurred while executing at the system level (kernel)
rxkB	total number of kilobytes received per second
iowait	percentage of time that the CPU or CPUs were idle during which the system had an outstanding disk I/O request
idle	percentage of time that the CPU or CPUs were idle and the system did not have an outstanding disk I/O request
retrans	number of RPC requests per second, those which needed to be retransmitted
txkB	total number of kilobytes transmitted per second

Table 1: Feature table

3. **Statistics of selected features**

The statistics of selected features are shown in Figure 2.

	4_cpu15_.usr	1_cpu6_.usr	0_cpu3_.sys	3_eth0_rxkB.s	3_cpu0_.iowait	2_cpu17_.sys	1_cpu10_.idle	3_retrans.s_2	2_cpu0_.idle	3_eth1_txkB.s
mean	0.15	0.54	0.35	0.12	0.59	0.44	98.55	0.66	97.85	492.29
std	0.61	0.93	0.64	0.08	1.49	0.95	1.73	6.30	2.59	182.65
max	16.00	27.45	4.00	0.76	48.98	55.00	100.00	306.00	101.01	1,442.63
min	0.00	0.00	0.00	0.00	0.00	0.00	26.00	0.00	31.00	18.21
25%	0.00	0.00	0.00	0.06	0.00	0.00	97.98	0.00	97.03	345.78
90%	0.99	1.02	1.01	0.24	2.97	1.02	100.00	0.00	100.00	731.76

Figure 2: Statistics of selected features

Task 2.1:

1. **Model accuracy**

The model accuracy of linear, random forest, neural network and naïve regression are shown in Table 2.

	Training time	Target	Accuracy
Linear	16 seconds	ReadsAvg	2896859.71
		WritesAvg	2716549.91
Random forest	16 minutes	ReadsAvg	0.022
		WritesAvg	0.024
Neural network	8 minutes	ReadsAvg	0.044
		WritesAvg	0.046
Naïve		ReadsAvg	0.044
		WritesAvg	0.046

Table 2: Model accuracy

It is clearly seen that the random forest regression (with 100 trees and mse criterion) gets the best accuracy in both targets. The reason of good performance might be the characteristics of random forest that it is suitable for multi-feature decision. However, the raining time of random forest is also the longest.

By plotting the predict results of the neural network (3 layers, 500 iterations, logistic activation, adam optimizer and with early stopping), it is interesting to see the predict results are quite similar to the prediction of naïve model. Therefore, it is quite reasonable that their accuracies are similar.

As for the linear model, the accuracy is not satisfying. However, by observing the results, we can find that only a couple of predicted results are extremely different from the real targets, and then affect the average accuracy.

2. **Prediction results**

With the random forest regressor chosen, 600 samples are predicted considering the plot readability. The results comparisons are shown in Figure 3 and 4.

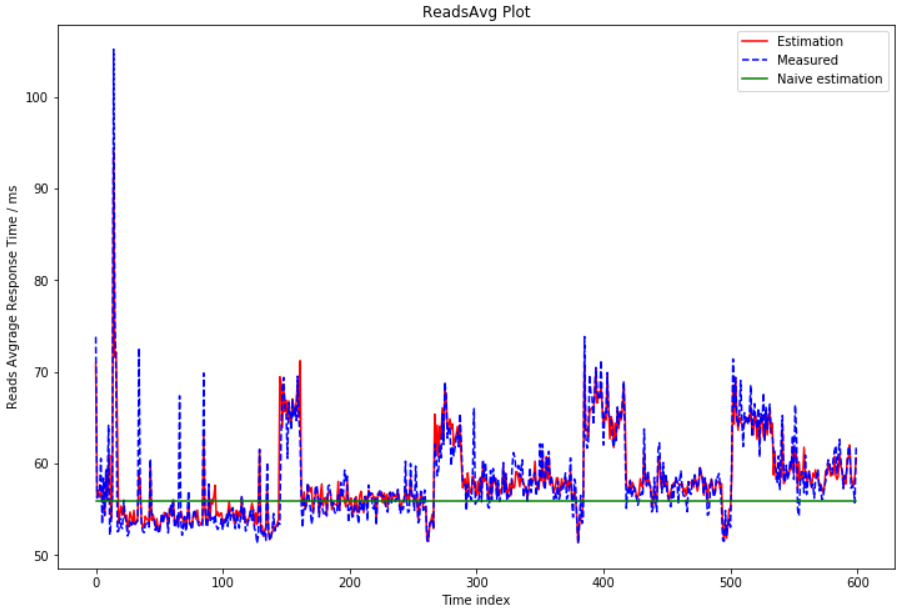

Figure 3: Predictions of ReadsAvg

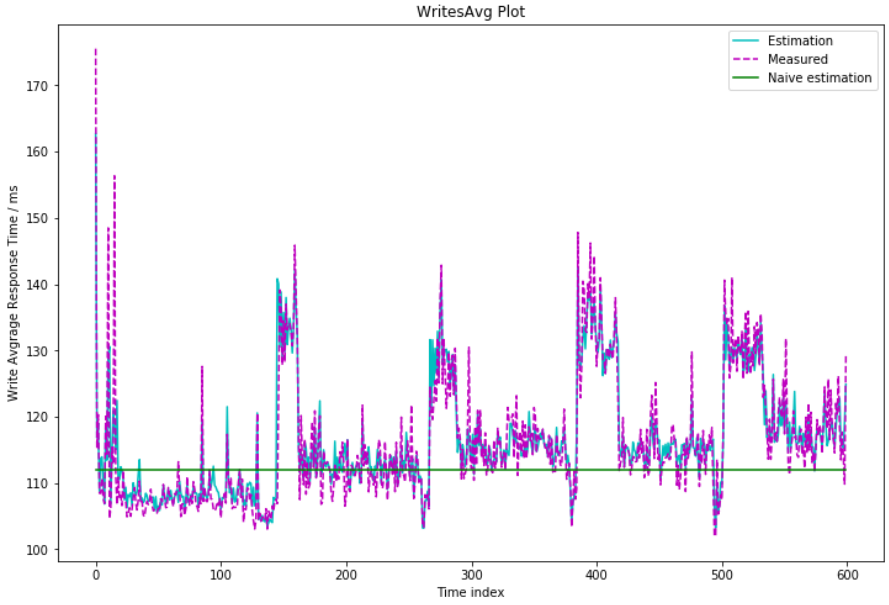

Figure 4: Predictions of WritesAvg

The prediction results of random forest seem coherent to the real targets, while the average is just a straight line since it is the average value of training data.

3. **Density plot and histogram**

By using the density plot and histogram, we can see the distribution of real targets. According to Figure 5 and 6, the network performance seems to remain at a certain level in most cases and sometimes get bursts. As for estimation errors shown in Figure 7, we can know that the random forest can predict more accurately the ReadsAvg than the WritesAvg. The reason of this fact might be the difference among targets of ReadsAvg is smaller than that of WritesAvg as shown in Figure 5 and 6.

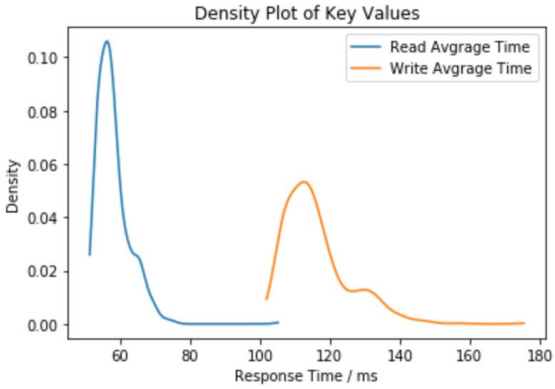

Figure 5: Density plot

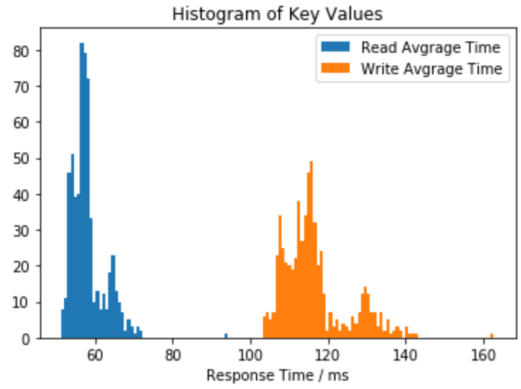

Figure 6: Histogram

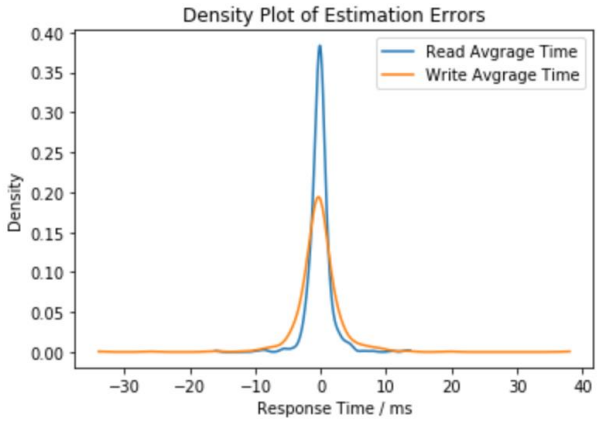

Figure 7: Density plot of estimation errors