CSE 291 Homework2

Wen liang A53214852
Chao Yu A99049546

Part I. Random Forests

Covertype

1. Train 2 types of Matrix, max depth is five and full depth.

2. Investigate the number of trees.

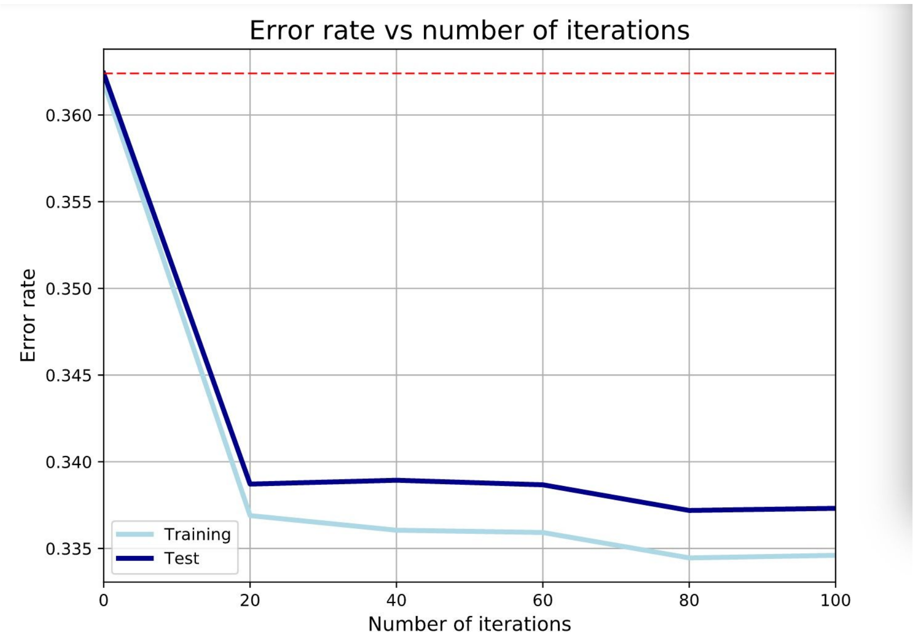


Figure 1. The plot of error rate and number of trees (5 depth)

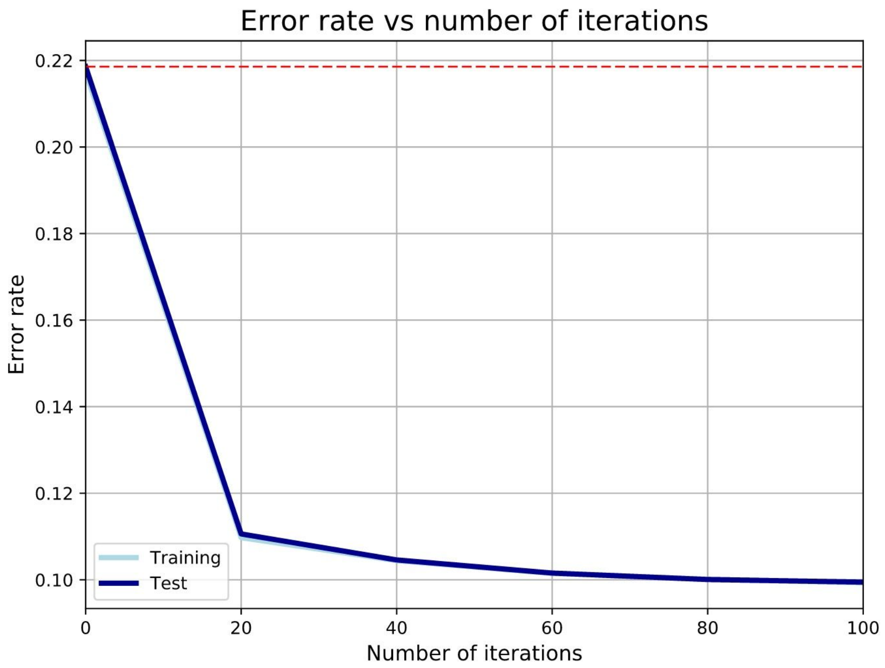


Figure 2. The plot of error rate and number of trees (full tree)

Comments: The error rate decreases along with the increase of number of trees. However, after 80 trees, the error reach a plateau and stop to decrease because the process is saturated. And this could be a reasonable number of trees in the forests.

3. Visualize the features importance

We compute the information gain for each feature, sort by the gain and plot their importance

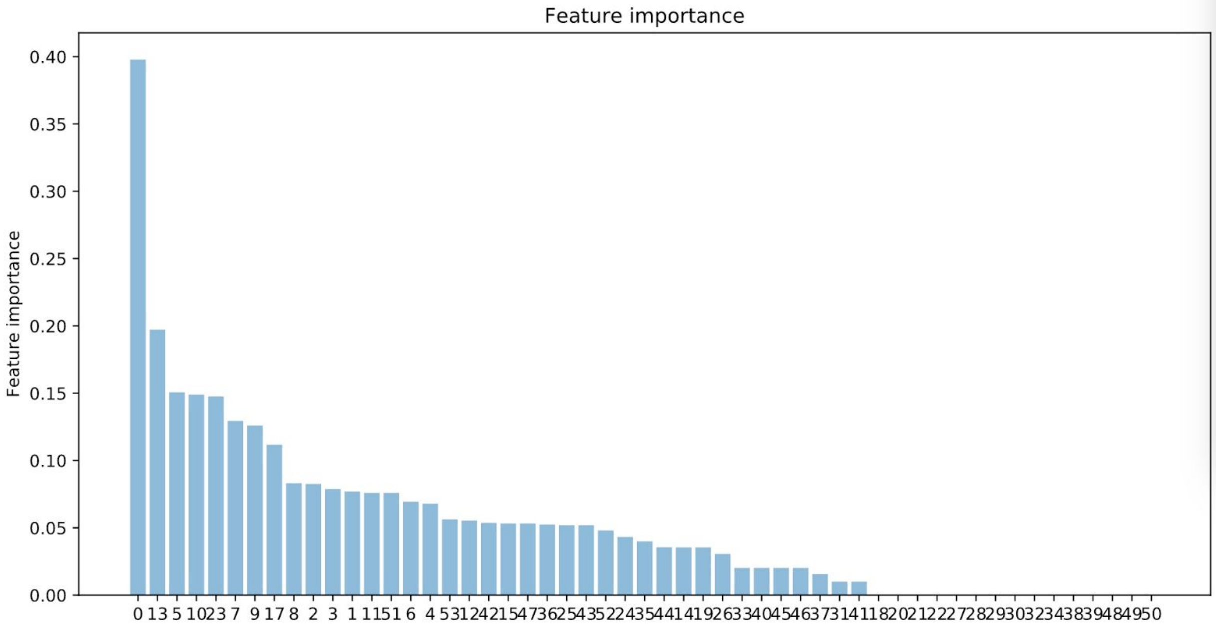
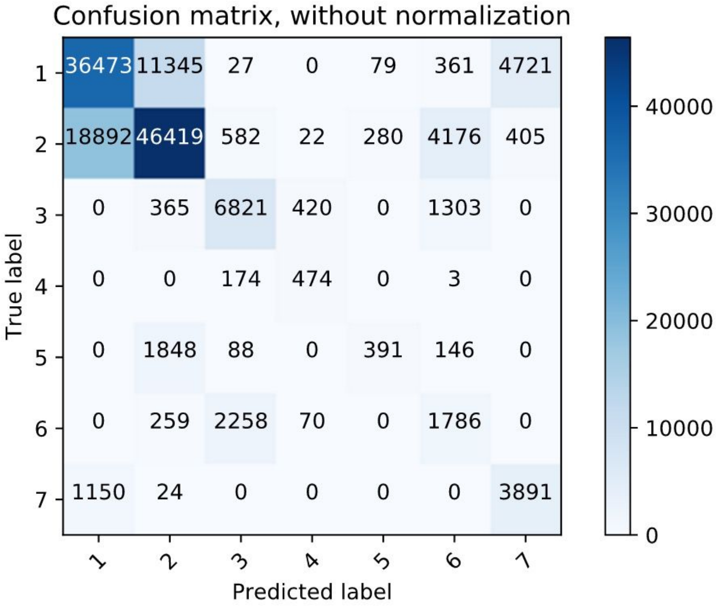


Figure 3. Feature importance

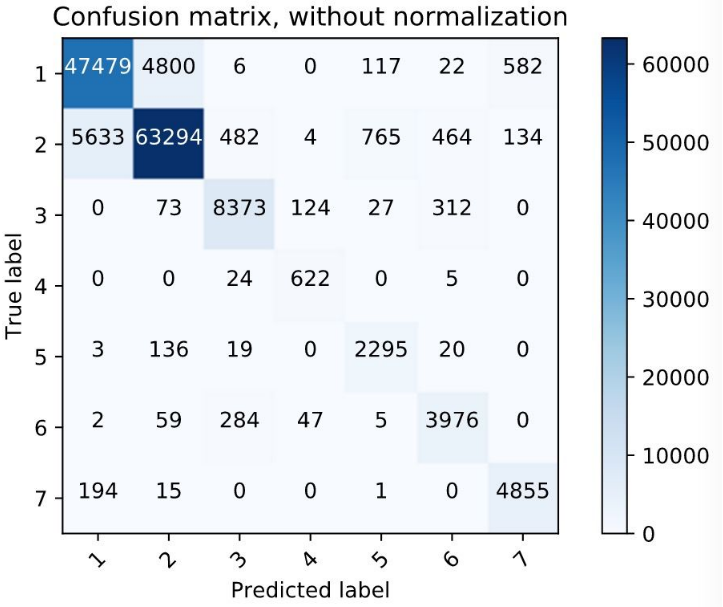
The feature that is most important is feature 0 and it is dominant in this figur. We can also see some feature have zero information gain for the classification and this means they may do not have relation with the label and cannot help to classify it.

4. Confusion matrix for 2 random forests



True label \ Predicted label	1	2	3	4	5	6	7
1	36473	11345	27	0	79	361	4721
2	18892	46419	582	22	280	4176	405
3	0	365	6821	420	0	1303	0
4	0	0	174	474	0	3	0
5	0	1848	88	0	391	146	0
6	0	259	2258	70	0	1786	0
7	1150	24	0	0	0	0	3891

Figure 4. Confusion matrix for 5-depth random forest classifier



True label \ Predicted label	1	2	3	4	5	6	7
1	47479	4800	6	0	117	22	582
2	5633	63294	482	4	765	464	134
3	0	73	8373	124	27	312	0
4	0	0	24	622	0	5	0
5	3	136	19	0	2295	20	0
6	2	59	284	47	5	3976	0
7	194	15	0	0	1	0	4855

Figure 5. Confusion matrix for full extent random forest

5. The error rate plot: Figure 1 and Figure 2

Here, the performance of full extent tree random forest is better than 5-depth. This is because the tree we build is 2 splits so depth = 5 is not enough to get a good result. Also, this dataset is highly skewed and these short trees are prone to be misleaded by dominant classes 1 and 2 which makes it harder to classify.

MNIST

6. Train 2 types of Matrix, max depth is five and full depth.

7. Investigate the number of trees.

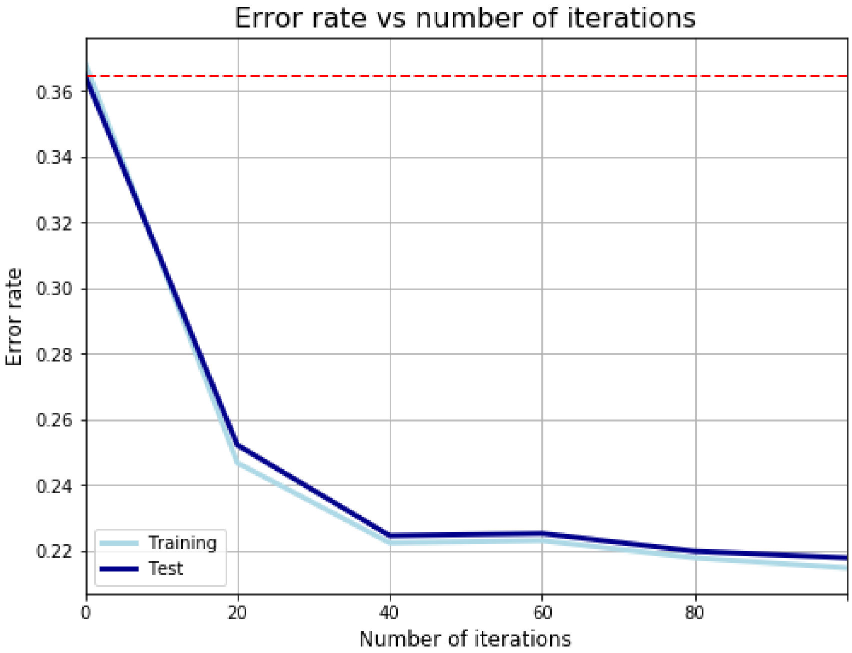


Figure 6. The plot of error rate and number of trees (5 depth)

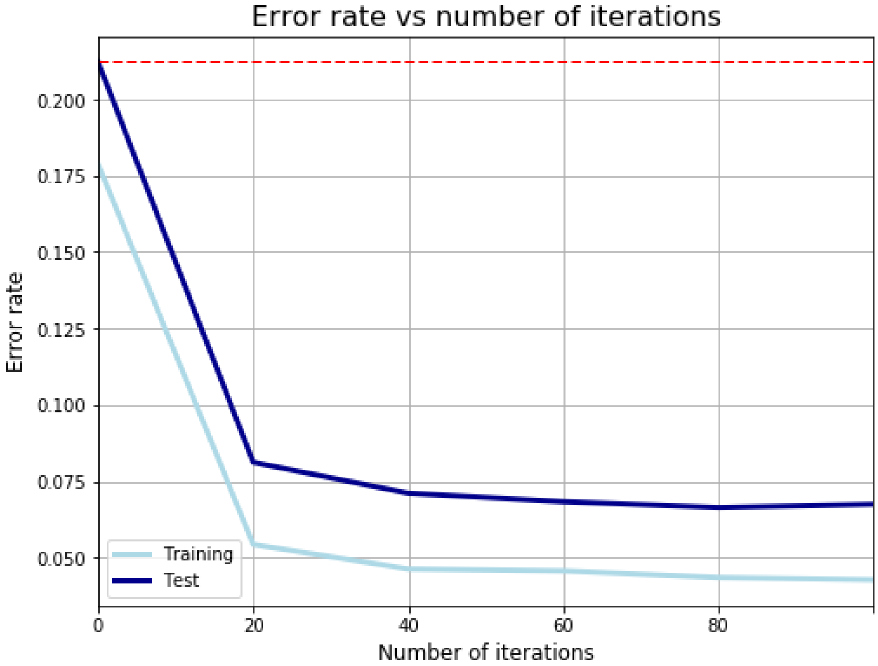


Figure 7. The plot of error rate and number of trees (full tree)

Similarly, the error rate decreases with increase of the number of trees. And 80 could be a reasonable number of trees.

8. Visualize the features importance

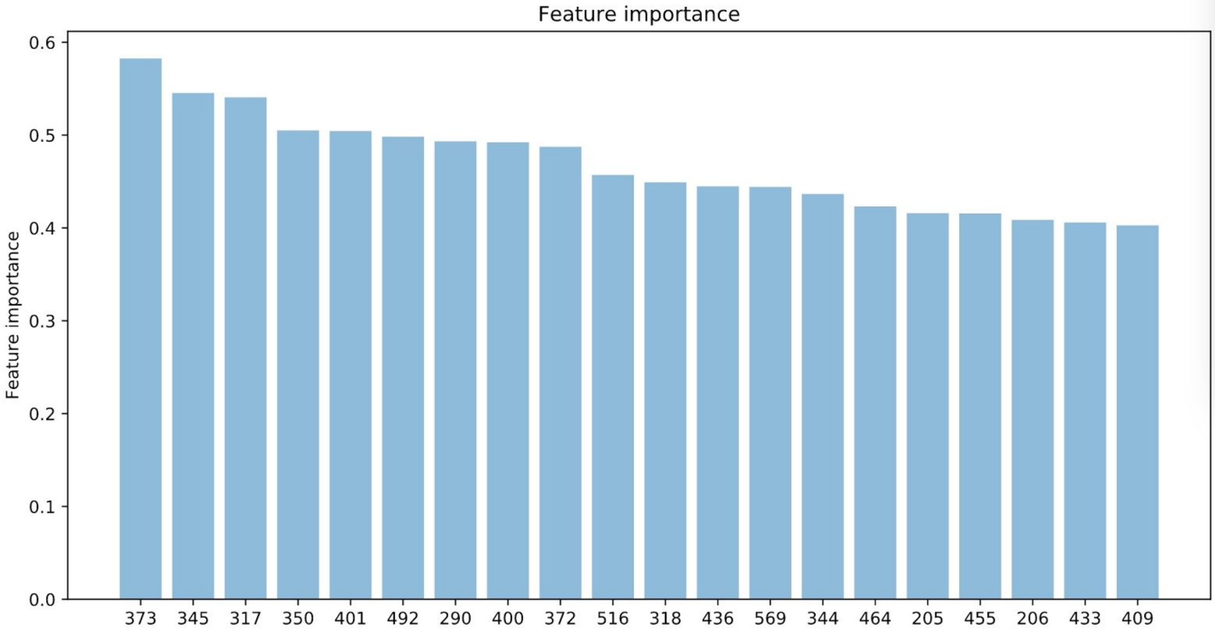
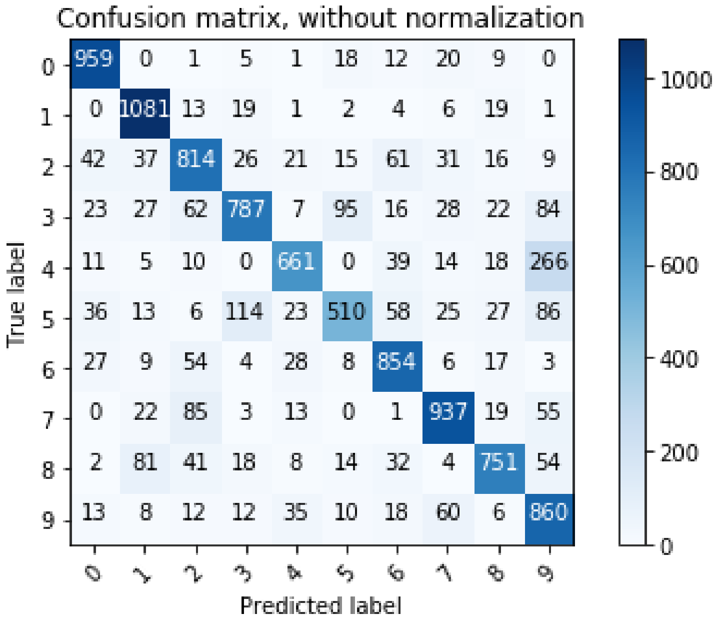


Figure 8. Feature importance

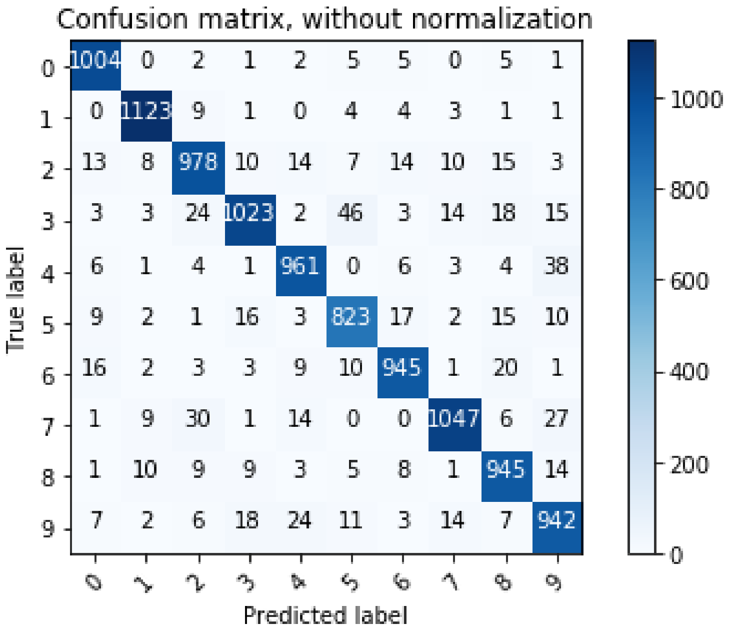
For feature ranking, because the data set is too large and it takes very long time for my laptop to run, so I picked those dimensions with very high variance to do the feature ranking and choose the top 20 most important features to plot. We can see that many dimensions in the MNIST dataset have rather high information gain.

9. Confusion matrix for 2 random forests



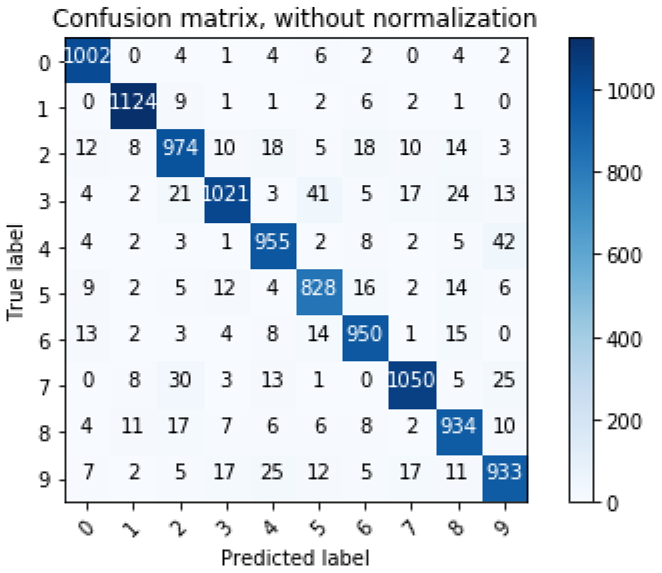
	0	1	2	3	4	5	6	7	8	9
0	959	0	1	5	1	18	12	20	9	0
1	0	1081	13	19	1	2	4	6	19	1
2	42	37	814	26	21	15	61	31	16	9
3	23	27	62	787	7	95	16	28	22	84
4	11	5	10	0	661	0	39	14	18	266
5	36	13	6	114	23	510	58	25	27	86
6	27	9	54	4	28	8	854	6	17	3
7	0	22	85	3	13	0	1	937	19	55
8	2	81	41	18	8	14	32	4	751	54
9	13	8	12	12	35	10	18	60	6	860

Figure 9. Confusion matrix for 5-depth random forest classifier



	0	1	2	3	4	5	6	7	8	9
0	1004	0	2	1	2	5	5	0	5	1
1	0	1123	9	1	0	4	4	3	1	1
2	13	8	978	10	14	7	14	10	15	3
3	3	3	24	1023	2	46	3	14	18	15
4	6	1	4	1	961	0	6	3	4	38
5	9	2	1	16	3	823	17	2	15	10
6	16	2	3	3	9	10	945	1	20	1
7	1	9	30	1	14	0	0	1047	6	27
8	1	10	9	9	3	5	8	1	945	14
9	7	2	6	18	24	11	3	14	7	942

Figure 10. Confusion matrix for full extent random forest



	0	1	2	3	4	5	6	7	8	9
0	1002	0	4	1	4	6	2	0	4	2
1	0	1124	9	1	1	2	6	2	1	0
2	12	8	974	10	18	5	18	10	14	3
3	4	2	21	1021	3	41	5	17	24	13
4	4	2	3	1	955	2	8	2	5	42
5	9	2	5	12	4	828	16	2	14	6
6	13	2	3	4	8	14	950	1	15	0
7	0	8	30	3	13	1	0	1050	5	25
8	4	11	17	7	6	6	8	2	934	10
9	7	2	5	17	25	12	5	17	11	933

Figure 11. Confusion matrix for 15 depth random forest

We realized that depth = 5 is not proper for our 2 split method so we set depth = 15 and got the confusion matrix. We can see that it has very similar result with full extent trees random forests and the “full tree” may have some overfit problems.

10. The error rate plot: Figure 6, 7

Similarly, the full extent tree random forest accuracy is better than 5-depth.

Part II. Boosting

Covertype

1. Implement adaboost with 2 types of weak learners

2. Investigate the reasonable number of iterations

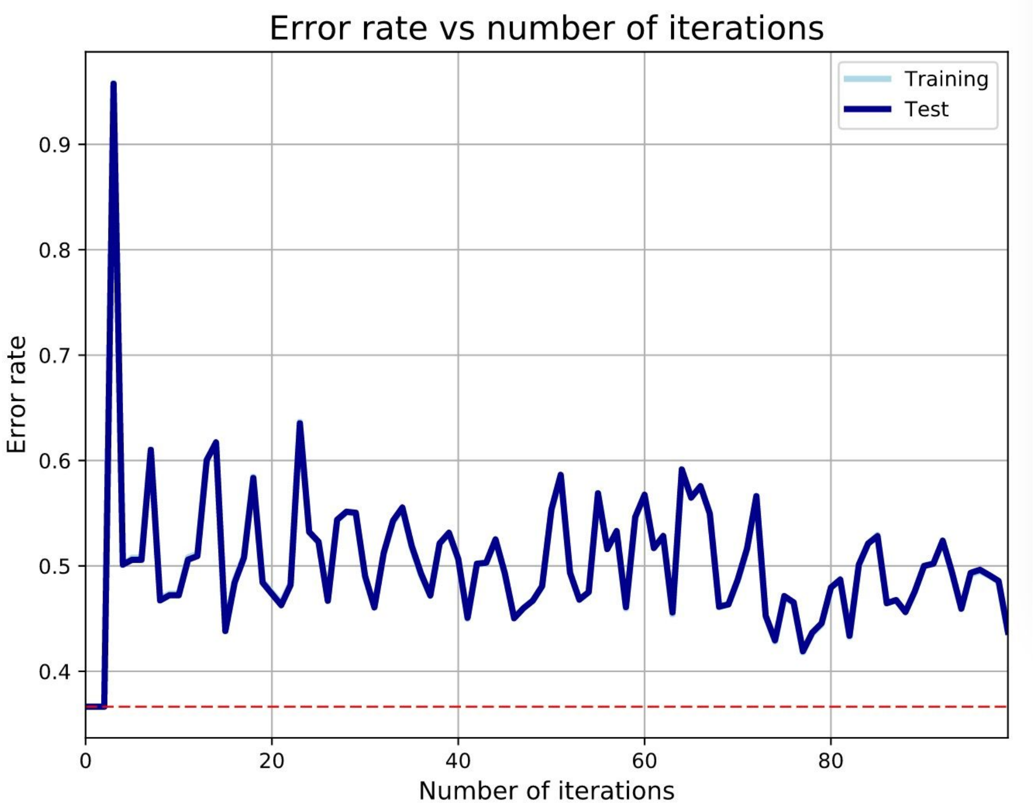


Figure 12. Plot of error rate and number of iterations (1-level-trees)

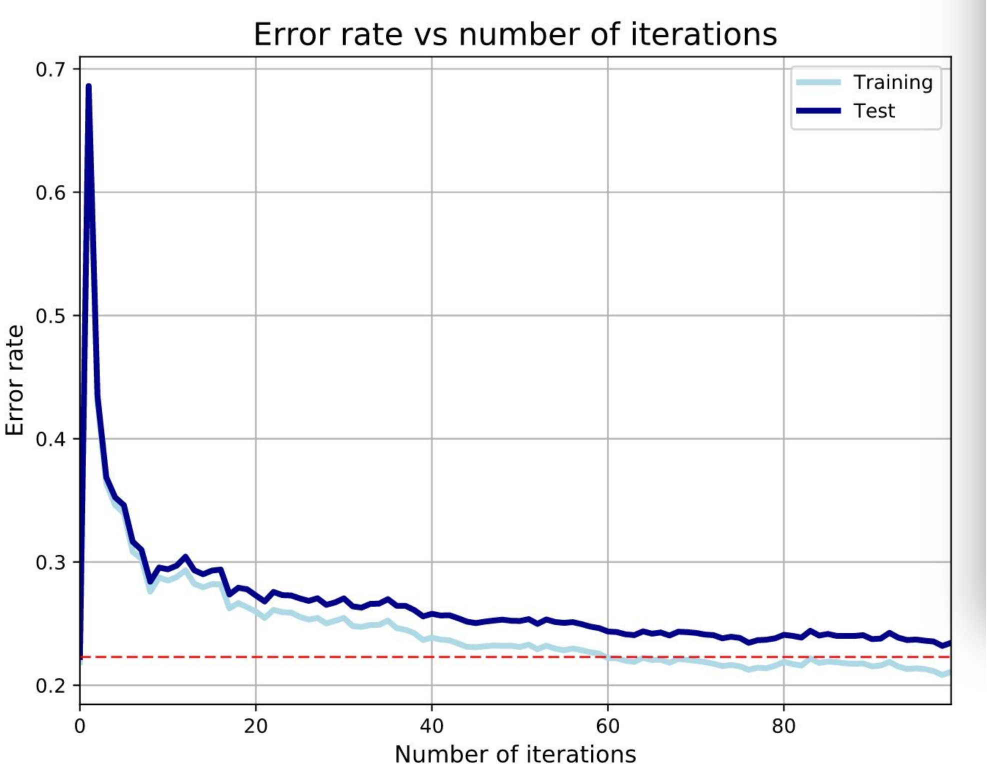


Figure 13. Plot of error rate and number of iterations (10-depth)

We got a very strange error rate plot due to the highly unbalanced dataset. We print out the result of first iteration and found that the prediction results only have the dominant classes. It got a very good error rate but not so reasonable result. According to the boosting method, it improves weight to those misclassified data, the predictor became more prone to choose those minority and increase the error rate extremely. Now, the error rate of one depth trees adaboost classifier just vibrate at a rather high value because the limitation of the depth makes them hard to explain the situation. However, the deeper trees can solve this problem and decrease the result gradually. Here, If you want higher accuracy, just one iteration is good. If you want high accuracy and reasonable results, choose 80 or more iterations is better.

3. Confusion matrix

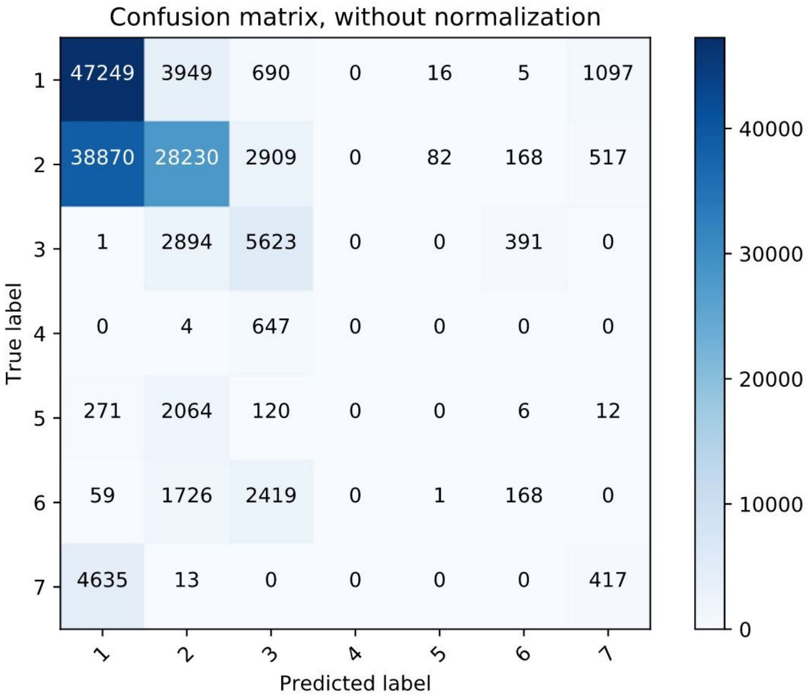
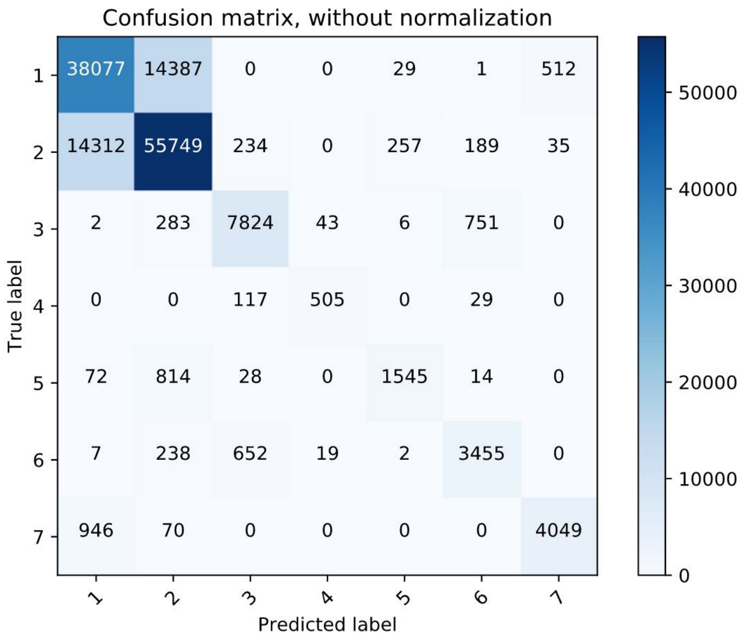


Figure 14. Confusion matrix for 1-depth adaboost



True label \ Predicted label	1	2	3	4	5	6	7
1	38077	14387	0	0	29	1	512
2	14312	55749	234	0	257	189	35
3	2	283	7824	43	6	751	0
4	0	0	117	505	0	29	0
5	72	814	28	0	1545	14	0
6	7	238	652	19	2	3455	0
7	946	70	0	0	0	0	4049

Figure 15. Confusion matrix for 10-depth adaboost

In figure 13, we can see that limitation of depth make it hard to get correct predictions for minority classes such as 4 and 5. However, the deeper trees adaboost partially solve this problem with more complex classifiers.

4. The error rate plot: Figure 12 and Figure 13

Due to the limitation of depth, the first classifier is worse than the second type. However, if the situation of data is not so that complex, we can have smaller accuracy difference between 2 classifiers. For the following MNIST example, the one-depth is also affected by the limitation and has very limited ability to classify dataset with many classes.

MNIST

5. Implement adaboost with 2 types of weak learners

6. Investigate the reasonable number of iterations

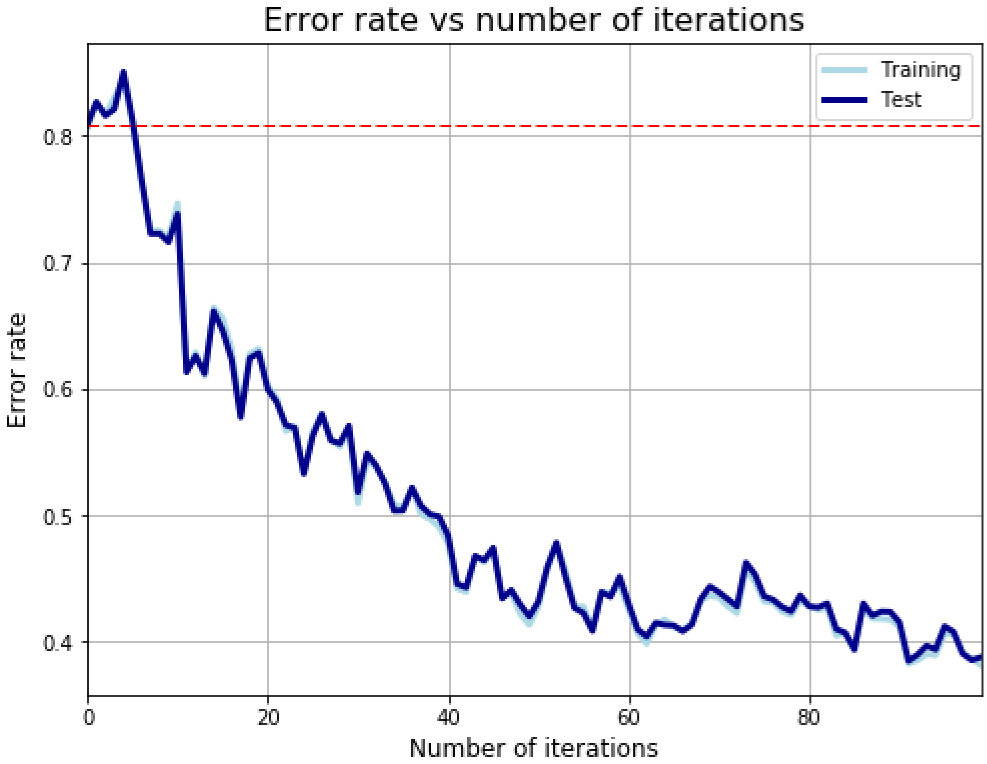


Figure 16. Plot of error rate and number of iterations (1-level-trees)

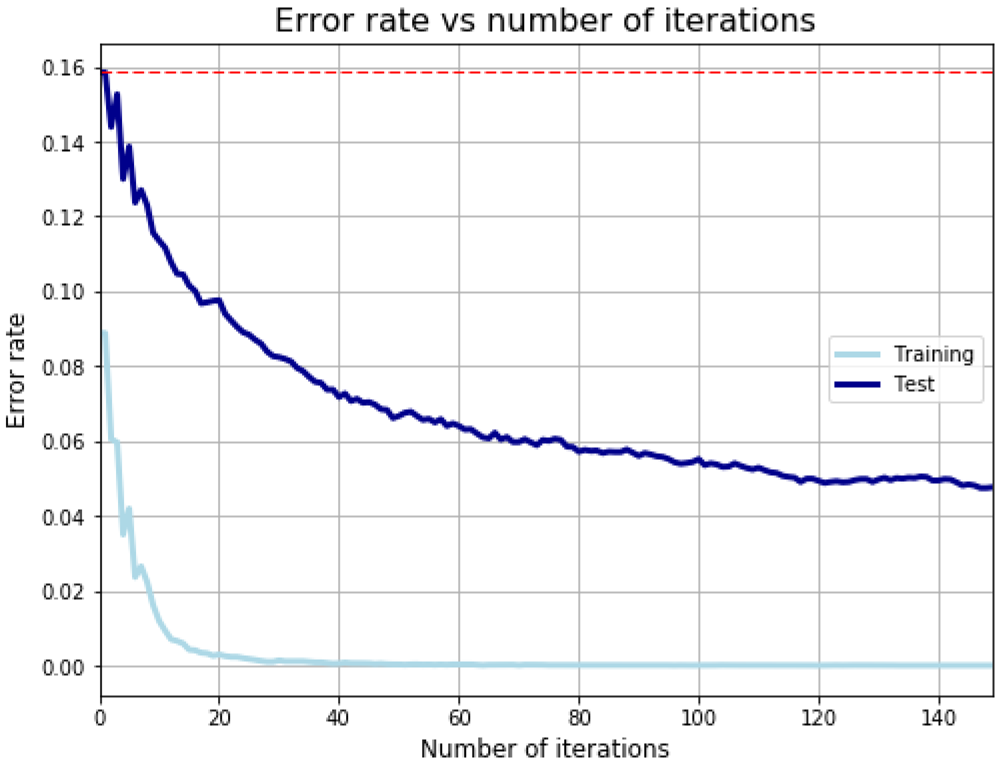


Figure 17. Plot of error rate and number of iterations (10-depth)

We chose 60 as a reasonable number of iterations for the first type classifier. However, the second type has higher ability to classify and saturate at about 120 iterations.

7. Confusion matrix

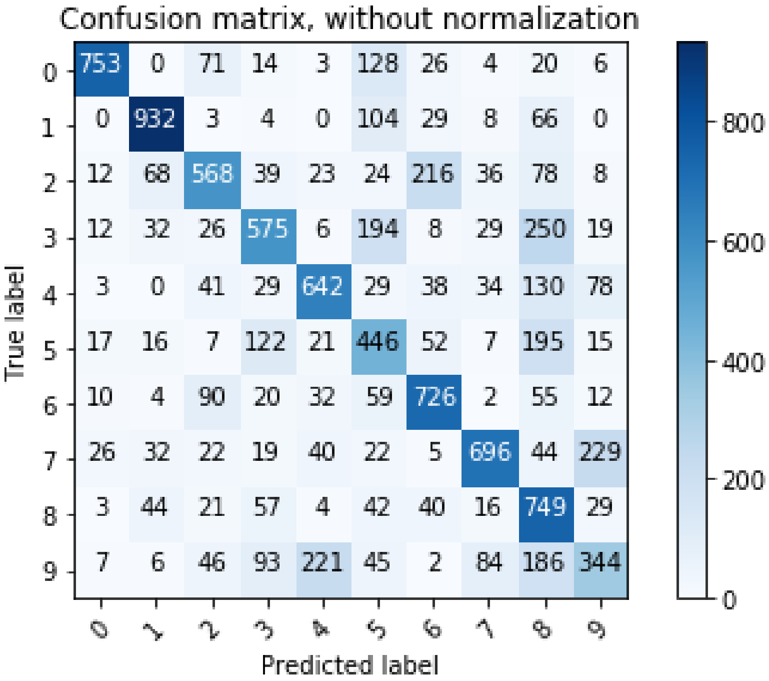


Figure 18. Confusion matrix (1-level-trees)

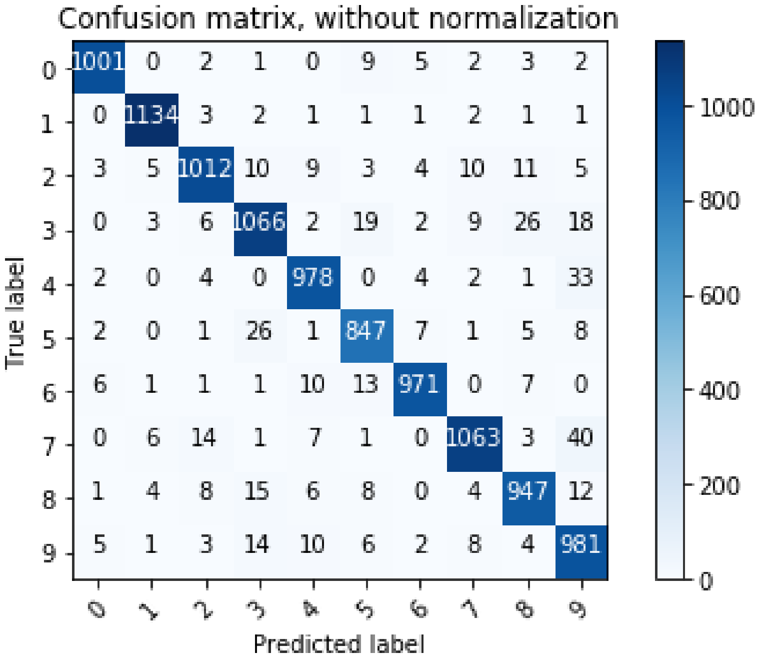


Figure 19. Confusion matrix (10-levels)

8. The error rate plot: Figure 16 and Figure 17

The 10-depth trees adaboost classifier have more than 94% accuracy which is much higher than 1-depth adaboost classifiers.