Assignment 2

Mukul Sati [msati3@gatech.edu]

February 28, 2016

1 Random forests trained with bagging

I carried out two iterations of the following experiment: I trained multiple random forests, with varying number of decision trees (I used the implementation in sklearn). I let the decision trees grow unconstrained during the first iteration, and limited their depth to 5 levels during the second.

1.1 Wine dataset

For the wine dataset, I selected a random subset of size 75% of the original data for training. I did not ensure that the samples in the training set are equally distributed for each label. While this is sub-optimal as mentioned by the instructor on Piazza, I think the split of 33%, 40%, 27% amongst the classes is not substantial to adversely affect the results.

The plots for the errors for the two iterations are shown in Fig. 1.1.

The confusion matrix for the wine-data for a Random Forest with 100 depth-not-limited trees and one with 150 depth-limited (to 5 levels) trees respectively is shown in Table 1

1.2 MNIST dataset

For MNIST, I used PCA to reduce dimensionality, retaining enough principal component vectors that explain 80% of the variance in the data. I feel this gives

Label	1	2	3
1	10	1	0
2	0	21	0
3	0	0	12

Label	1	2	3
1	10	1	0
2	1	20	0
3	0	0	12

Table 1: The confusion matrix for a Random Forest with 100 depth-not-limited trees (left). The confusion for a Random Forest with 150 depth-limited to 5 trees (right). These results are for the wine-dataset.

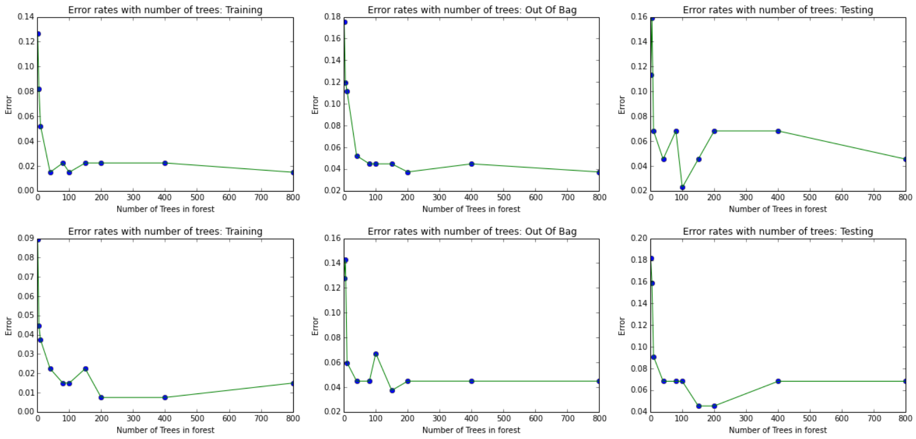

Figure 1: The errors on the (a) Training, (b) Out of bag and (c) Testing data when using unconstrained depth binary trees (top) and when using binary trees that are only allowed to grow to depth 5 (bottom) for the wine dataset

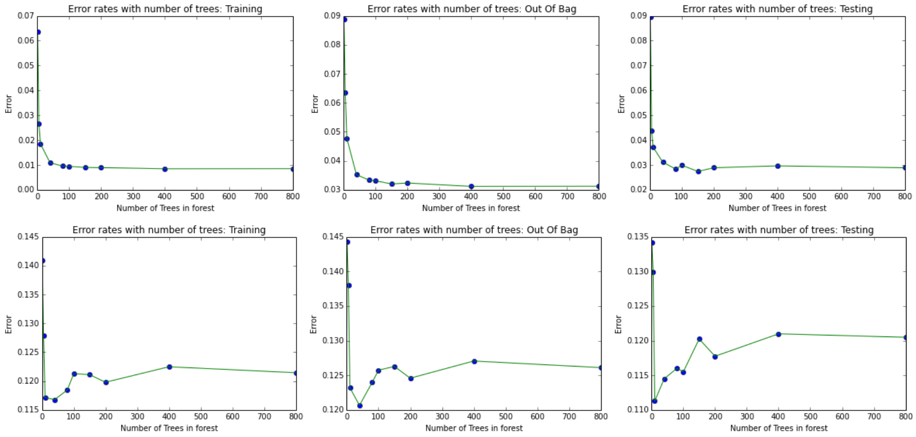

Figure 2: The errors on the (a) Training, (b) Out of bag and (c) Testing data when using unconstrained depth binary trees (top) and when using binary trees that are only allowed to grow to depth 5 (bottom) for the MNIST dataset

Digit	0	1	3	5
0	964	0	7	9
1	0	1126	5	4
3	4	4	972	30
5	18	1	28	845

Digit	0	1	3	5
0	852	0	26	102
1	0	1092	14	29
3	13	8	847	142
4	42	4	80	766

Table 2: The confusion matrix for a Random Forest with 150 depth-not-limited trees (left). The confusion for a Random Forest with 40 depth-limited to 5 trees (right). These results are for the MNIST dataset.

me a good balance between a performant algorithm and execution time. The plots for the errors for the two iterations are shown in Fig. 1.2.

The confusion matrix for the MNIST dataset for a Random Forest with 100 depth-not-limited trees and one with 150 depth-limited (to 5 levels) trees respectively is shown in Table 2

2 Boosting