Decision Trees and Random Forests

By

Charles Silkin

Contents

PROJECT INSTRUCTIONS

Using the training and test sets, perform the following tasks:

a) On the `madelon` dataset, train decision trees of maximum depth 1, 2, ... up to 12, for a total of 12 decision trees. If your package does not allow the max depth as a parameter, train trees with 2^1, 2^2, … , 2^{12} nodes, again a total of 12 trees. Use the trained trees to predict the class labels on the training and test sets, and obtain the training and test misclassification errors. Plot on the same graph the training and test misclassification errors vs tree depth (or log2 of nodes) as two separate curves. Report in a table the minimum test error and the tree depth (number of nodes or splits) for which the minimum was attained.

b) Repeat point a) on the `satimage` dataset.

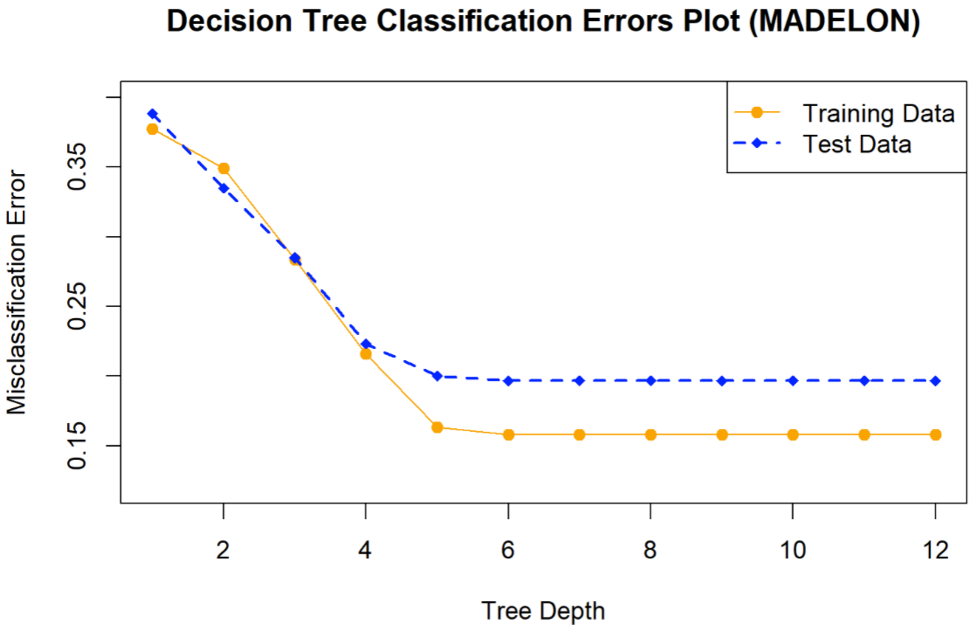
c) On the madelon dataset, for each of $k \in \{3, 10, 30, 100, 300\}$ train a random forest with k trees where the split attribute at each node is chosen from a random subset of $\sim \sqrt{500}$ features. Use the trained trees to predict the class labels on the training and test sets, and obtain the training and test misclassification errors. Plot on the same graph the training and test errors vs number of trees k as two separate curves. Report the training and test misclassification errors in a table.

d) Repeat point c) on the `madelon` dataset where the split attribute at each node is chosen from a random subset of $\sim LN(500)$ features.

e) Repeat point c) on the `madelon` dataset where the split attribute at each node is chosen from all 500 features.

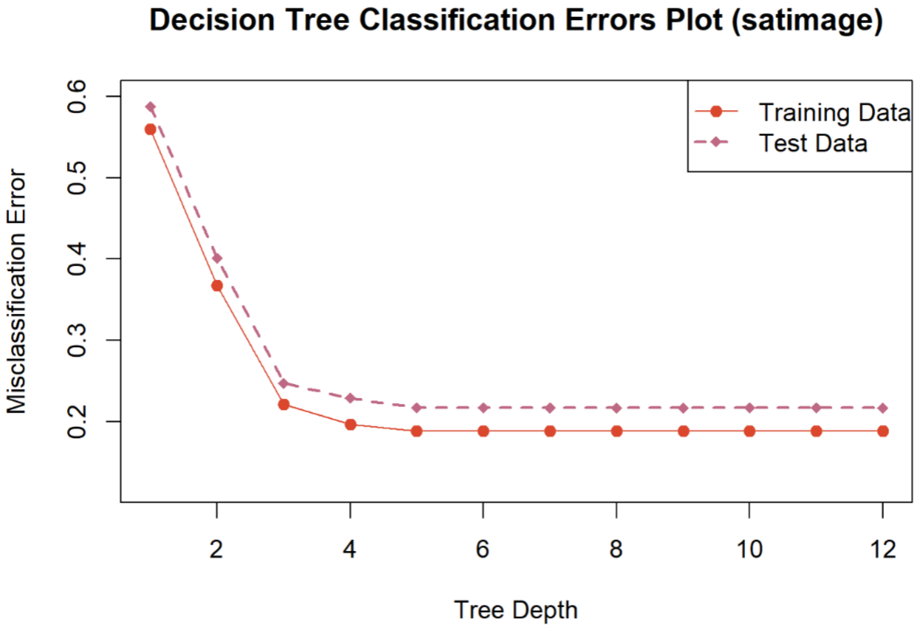
PART A

Minimum Test Error	Tree Depth
0.1966667	6



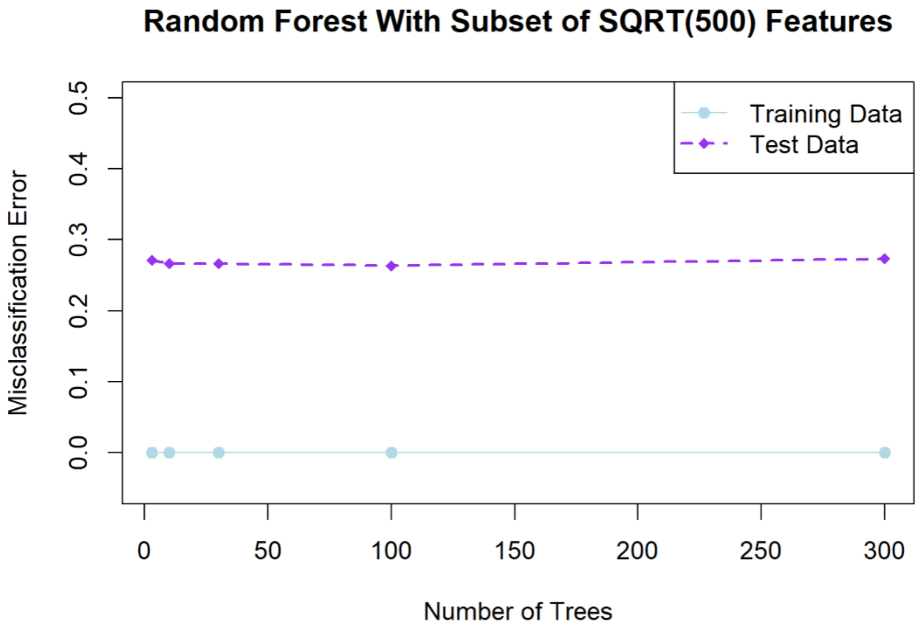
PART B

Minimum Test Error	Tree Depth
0.2170	5



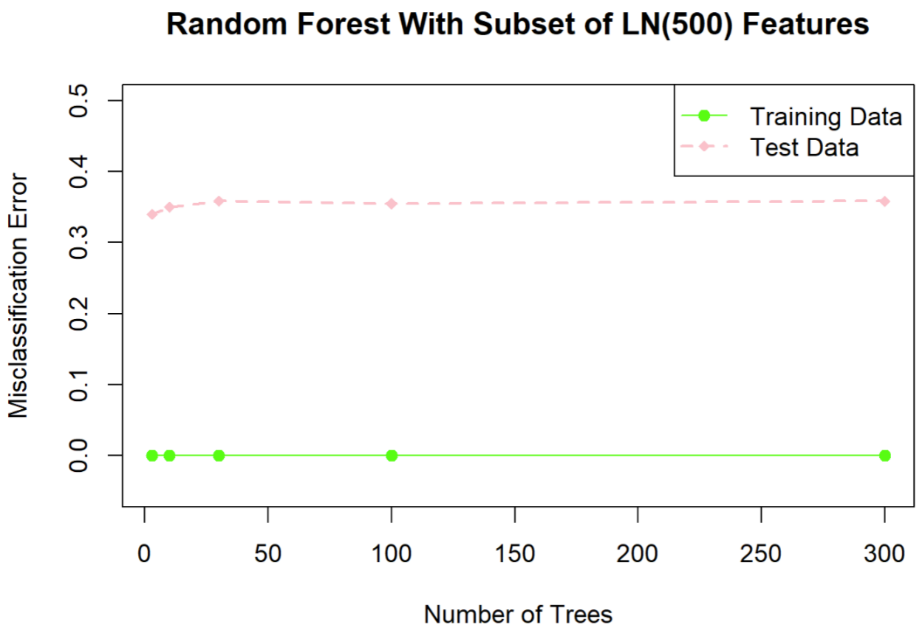
PART C

Number of Trees	Training Misclassification Error	Test Misclassification Error
3	0	0.2716667
10	0	0.2666667
30	0	0.2666667
100	0	0.2633333
300	0	0.2733333



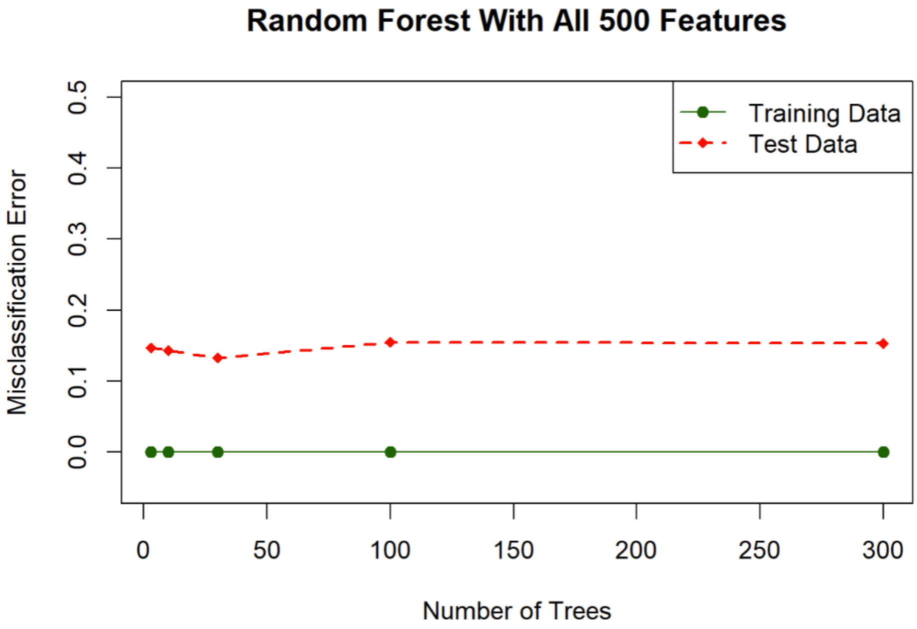
PART D

Number of Trees	Training Misclassification Error	Test Misclassification Error
3	0	0.34
10	0	0.35
30	0	0.3583333
100	0	0.355
300	0	0.3583333



PART E

Number of Trees	Training Misclassification Error	Test Misclassification Error
3	0	0.1466667
10	0	0.1433333
30	0	0.1333333
100	0	0.155
300	0	0.1533333



REFERENCES

1. https://www.guru99.com/r-random-forest-tutorial.html
2. https://www.r-bloggers.com/2021/04/decision-trees-in-r/
3. https://www.youtube.com/watch?v=HmEPCEXn-ZM
4. https://www.youtube.com/watch?v=HeTT73WxKIc
5. https://www.digitalocean.com/community/tutorials/plot-function-in-r