CS 771A: Assignment #2

Due on Sunday, February 1, 2015

Harish Karnick

Srijan R. Shetty (11727)

Contents

List of Figures

Problem 1

A binary search to find out the optimum value of k yielded the following result. (All errors are 5 fold cross validated misclassification error).

m	1 random split
min error ($k = 500$)	0.91%
threshold	10% of min error
optimum k	105 trees
error at optimum k	1.077%

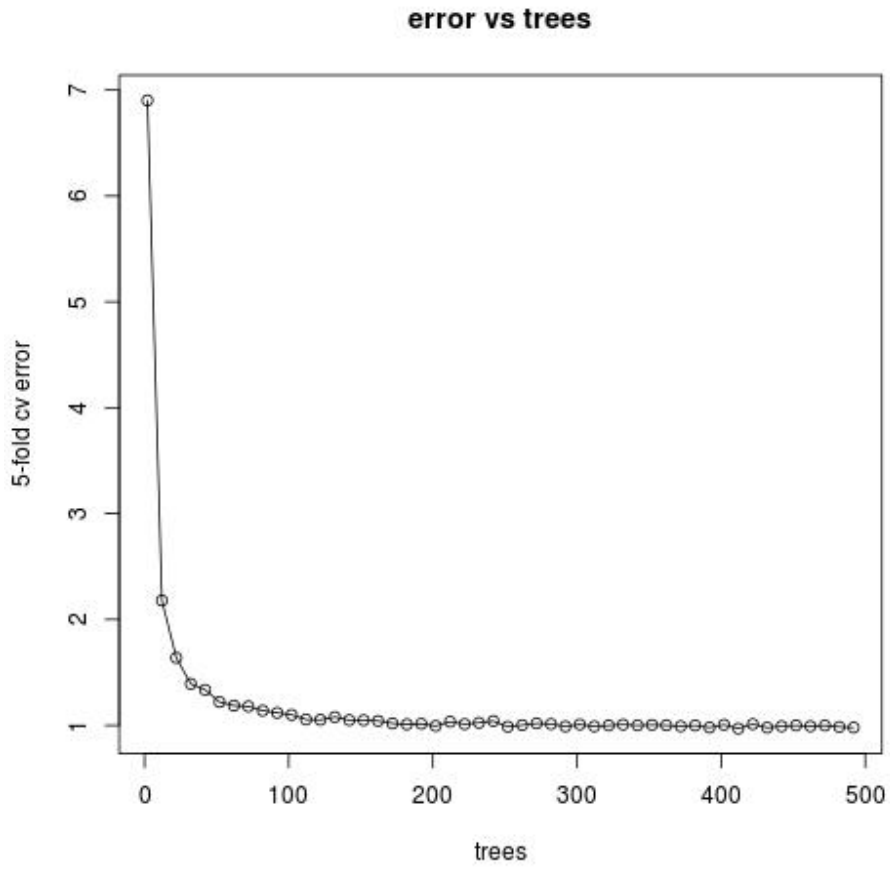

Figure 1: A plot of 5-fold misclassification error versus the number of trees follows the expected decreasing trend.

Problem 2

The **Out of Bag (OOB) error for** 1.25 **times the optimum value as found in Problem** 1 **is** 6.17%

Problem 3

m	5-fold misclassification error
1	1.047%
3	0.714%
5	0.745%
7	0.772%
9	0.796%

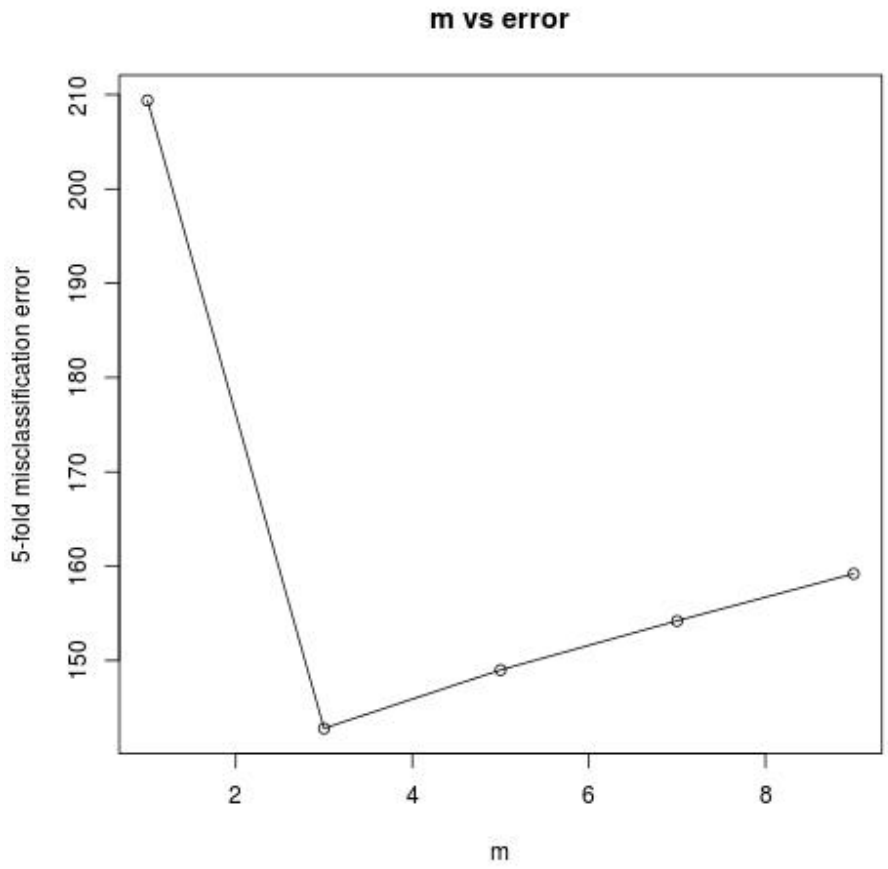

Figure 2: Plot of m vs 5-fold-misclassification-error shows an optimum at approximately log of the number of attributes in data.

Problem 4

In light of the observations tabled below, bagging is justified as a method of randomization to select samples because the cross-validation error levels off after 70% and hence the relative gain in making sure that each tree in the forest sees 70% or more is very marginal considering that each tree sees atleast 63% distinct vectors from the learning set. Additionally, the theoretic bounds work only we sample $|L|$ vectors from the learning set L, so bagging is a justified method for randomization.

Problem 4 continued on next page...

Percentage Trees	5-fold misclassification error	
10	289	1.81%
20	251	1.44%
30	230	1.283%
40	225	1.154%
50	200	1.10%
60	188	1.027%
70	248	0.987%
80	157	0.976%

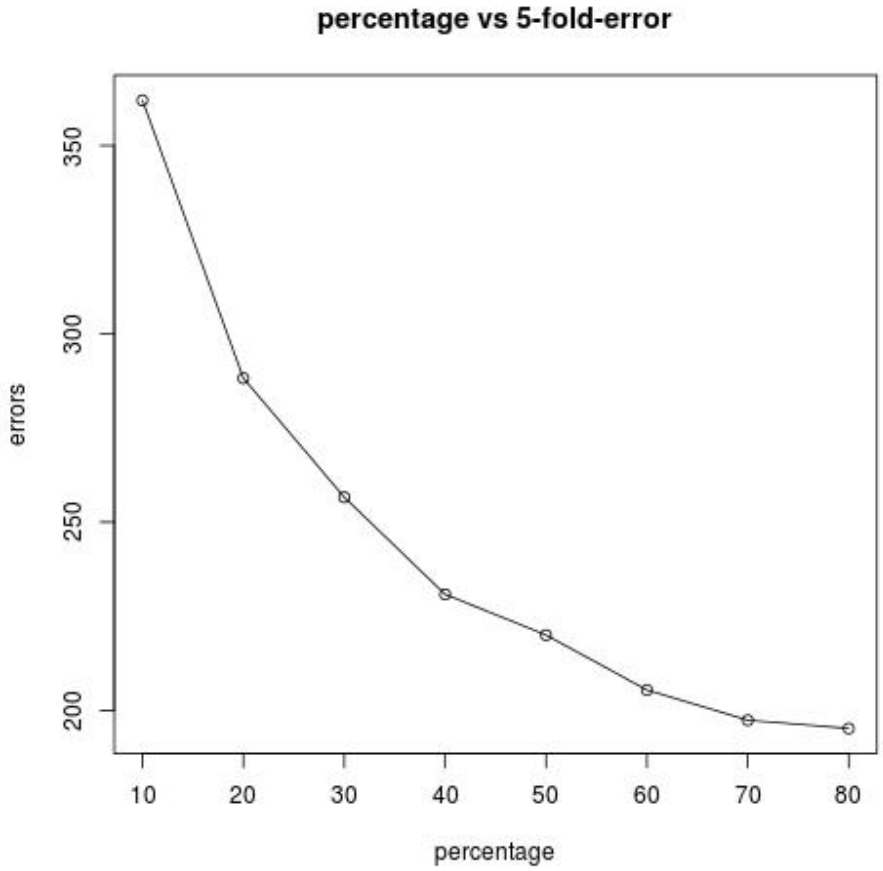

Figure 3: Plot of percentage vs 5-fold-error shows a decreasing trend