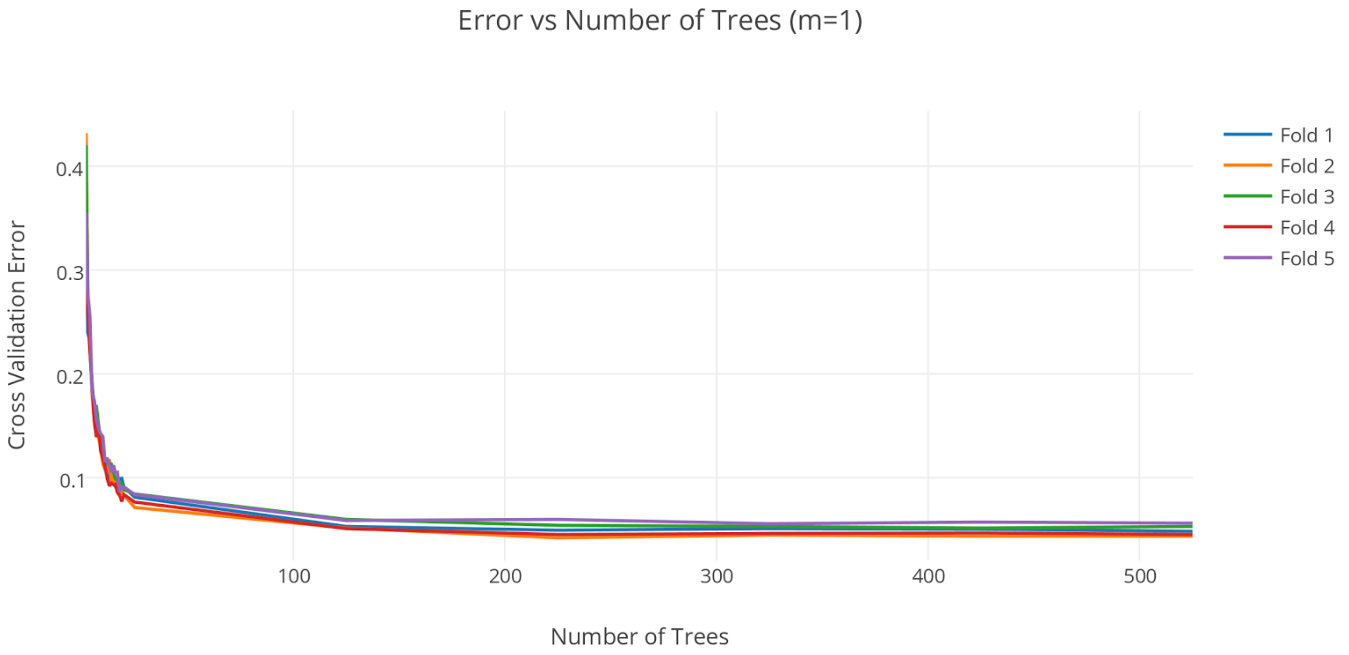
CS771A: Machine Learning: Tools, Techniques, Applications

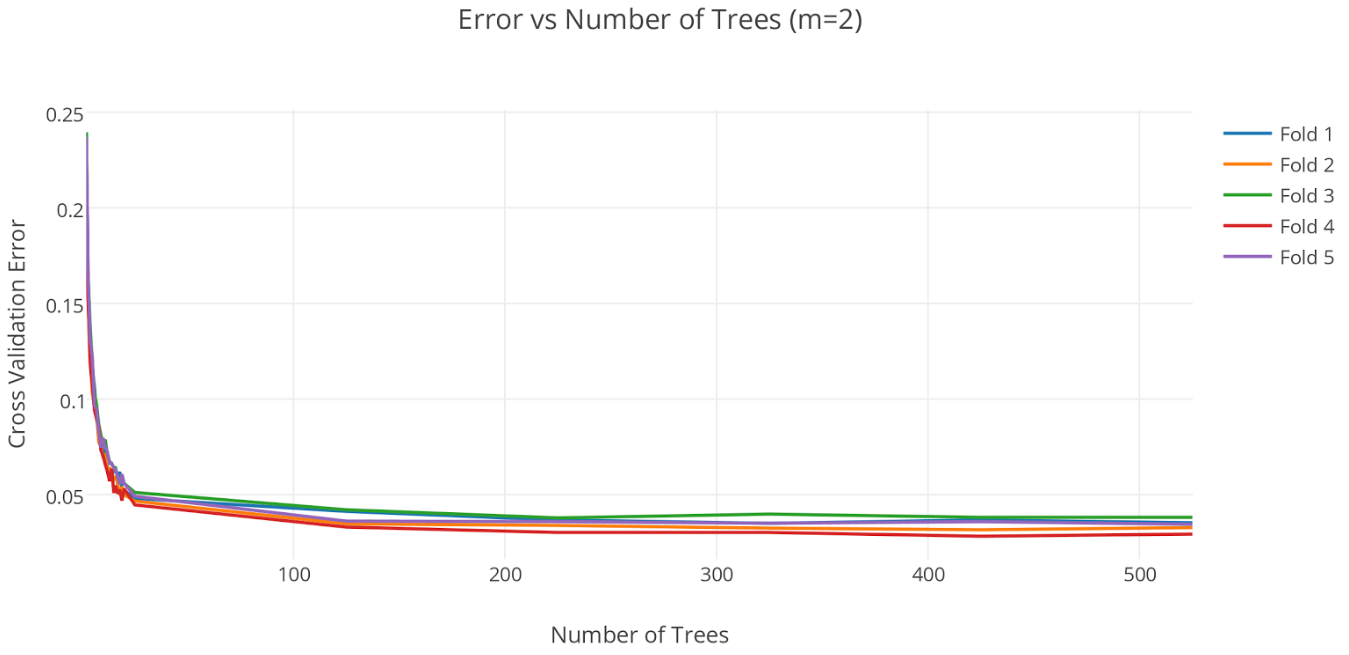
Saurav Kumar, 12641
31st January 2015

Assignment 2

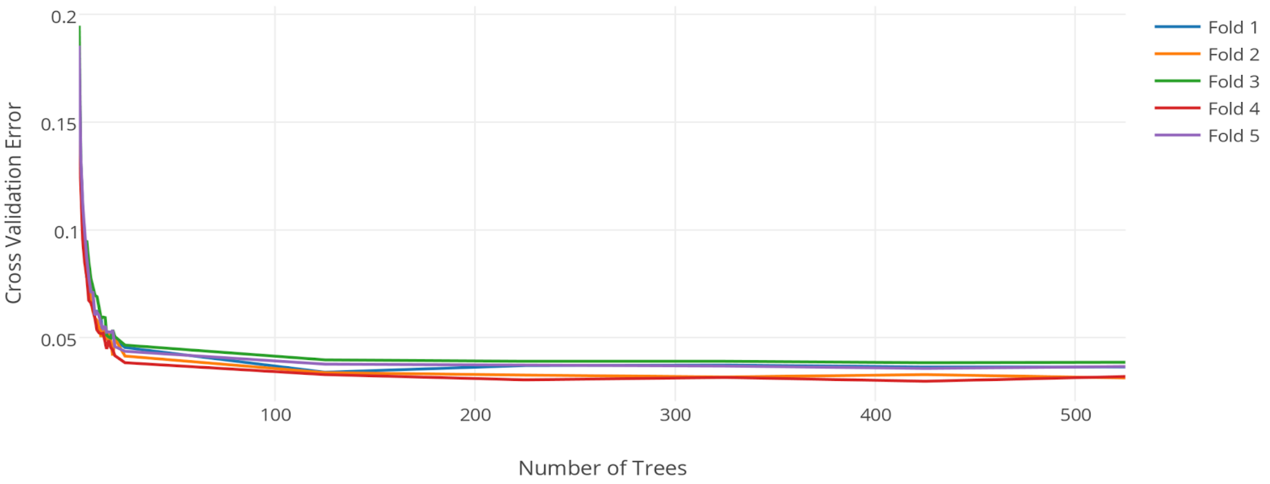
Part 1:

Number of Trees where Error levels off is close to 41 (using m = 5, and difference threshold = 0.005 from the error at 500 trees).
Following plots were recorded for Cross Validation Error (5 Fold) versus Number of Trees, for m = 1, 2, 4, 5 and 8. For each of the plots, error levels off around 40-50 trees.

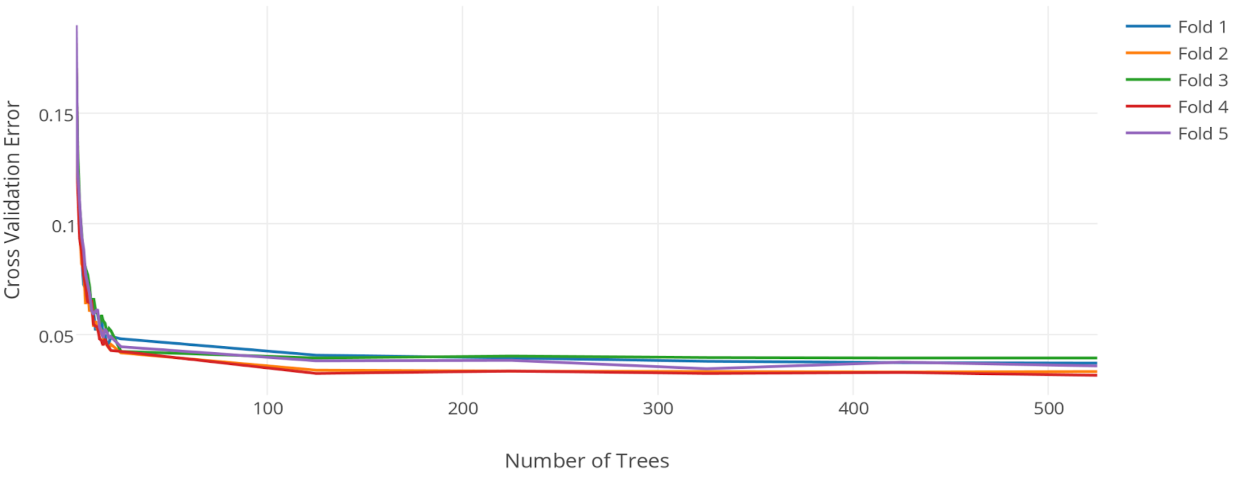




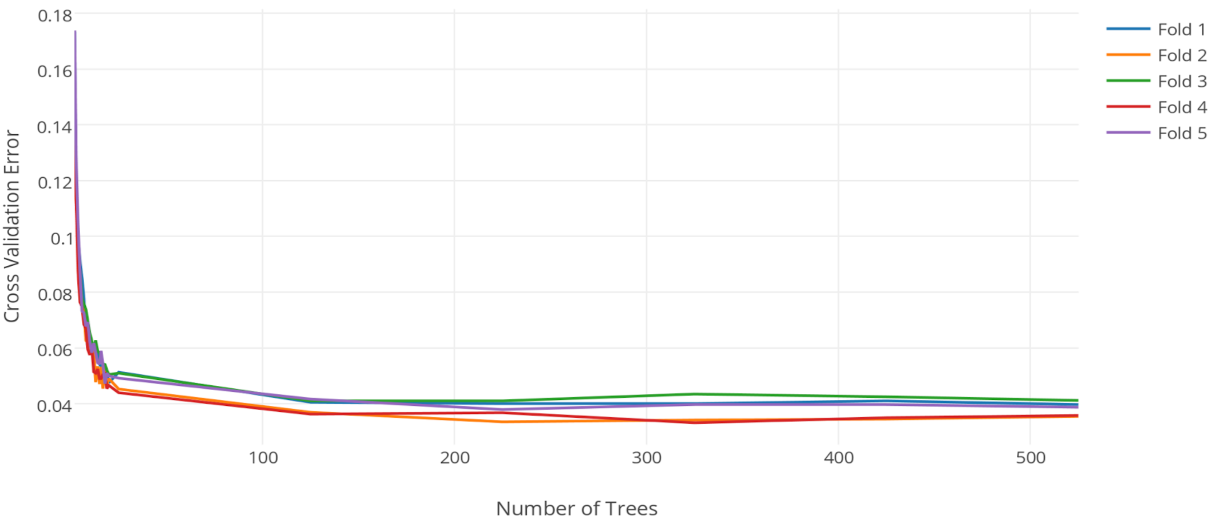
Error vs Number of Trees (m=4)



Error vs Number of Trees (m=5)



Error vs Number of Trees (m=8)



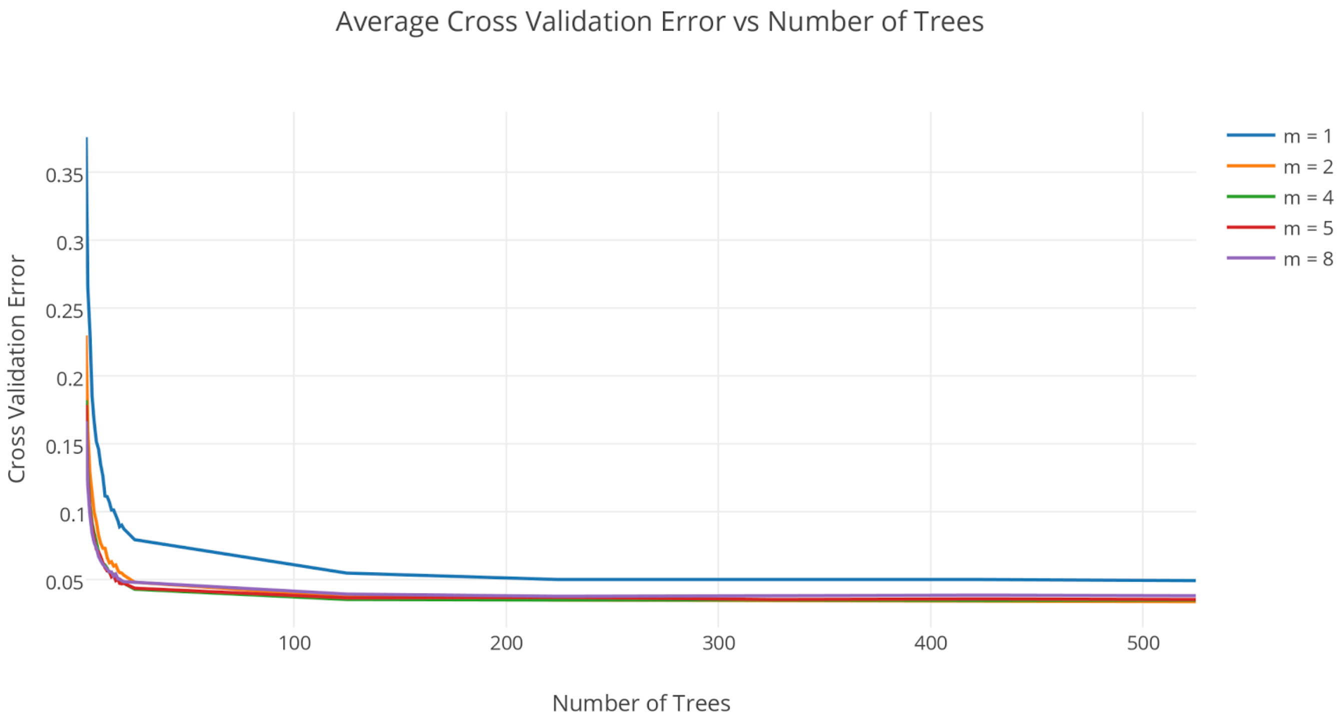
Part 2:

Out Of Bag (OOB) error for the forest with 41 Trees = **0.04675**
Out Of Bag (OOB) error for the forest with 500 Trees = **0.03095**

Part 3:

For number of trees = 52, which is 1.25 times the number of trees where error levelled off, following data was observed:

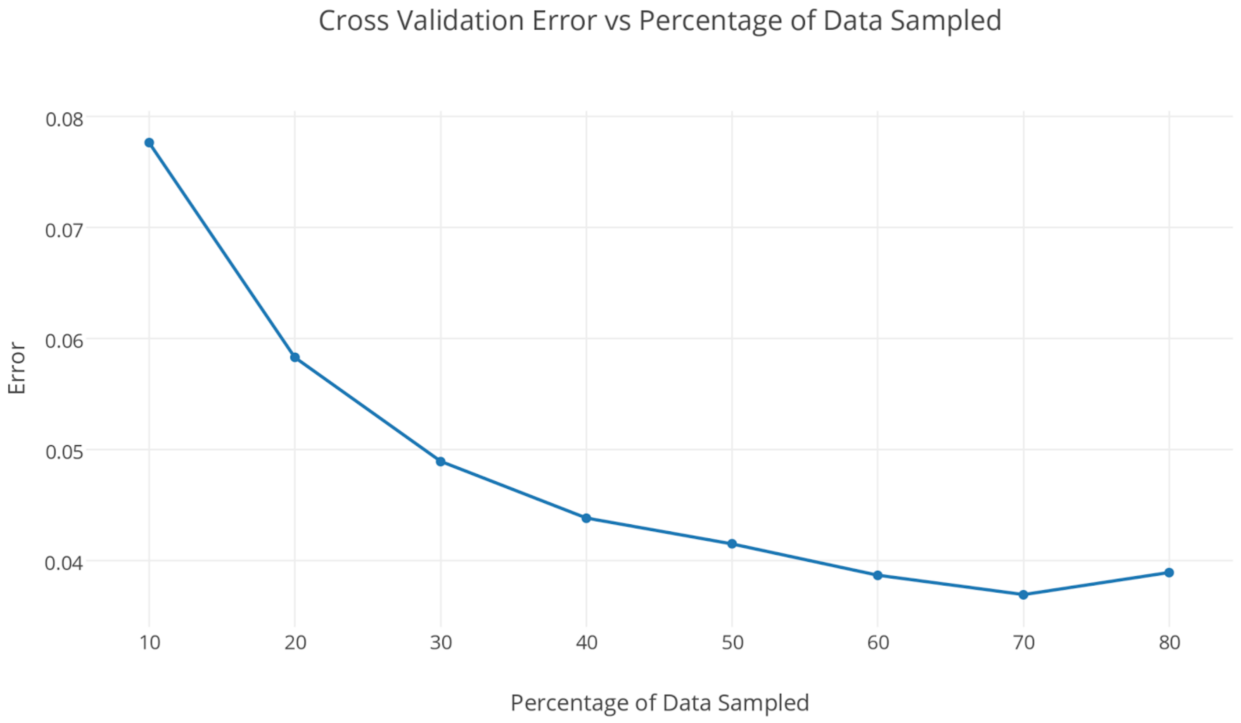
Fold	**Error for m = 1**	**Error for m = 2**	**Error for m = 4**	**Error for m = 8**
1	0.061000	0.042250	0.042250	0.042750
2	0.058250	0.036250	0.034250	0.042500
3	0.066500	0.044000	0.043500	0.043000
4	0.060750	0.034250	0.036000	0.037000
5	0.064250	0.041000	0.041750	0.042250
Mean	**0.062150**	**0.039550**	**0.039550**	**0.041500**



Part 4:

Effect of size of randomly sampled data while constructing the tree: It is clear from the plot and the table that error decreases as the size of learning set increases.

Sample Size	Error
10%	0.07765
20%	0.0583
30%	0.0489
40%	0.0438
50%	0.0415
60%	0.03865
70%	0.0369
80%	0.0389



Justification for bagging:

Since Random Forest is an unstable learning technique, ensembling by bagging will improve accuracy of the model. As evident from the above plot, error decreases with increase in percentage of data sampled to construct each tree in the forest, and reaches minimum around 60-70%. Bagging, on average, uses 63% of the data to construct each tree, which leads to less correlated trees in the forest, and hence improving the accuracy. From the plot above, we observe that error is minimum near 60-70%, justifying the bagging technique.