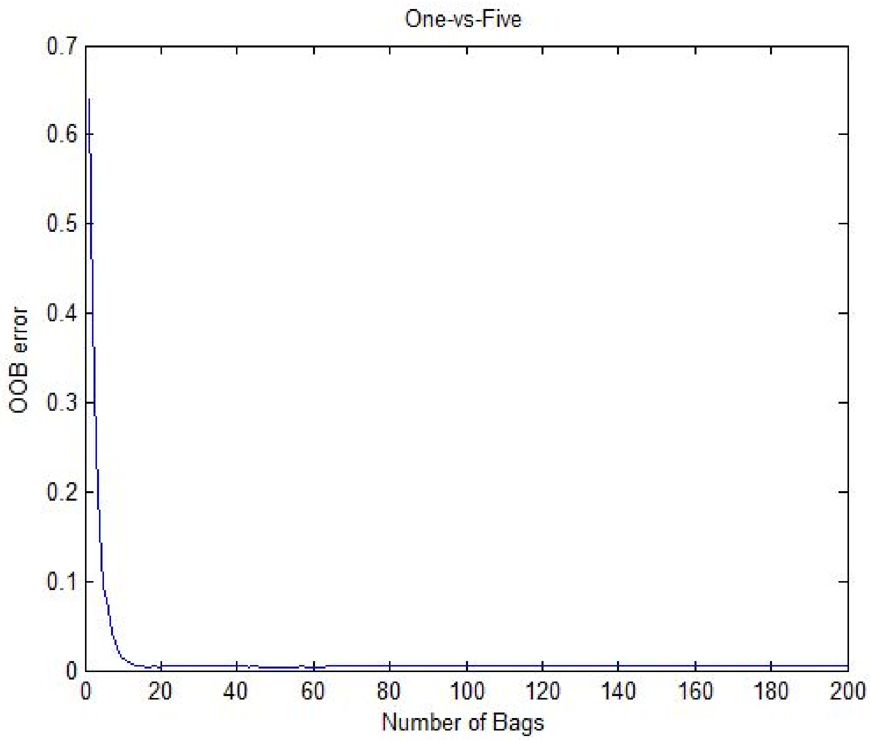
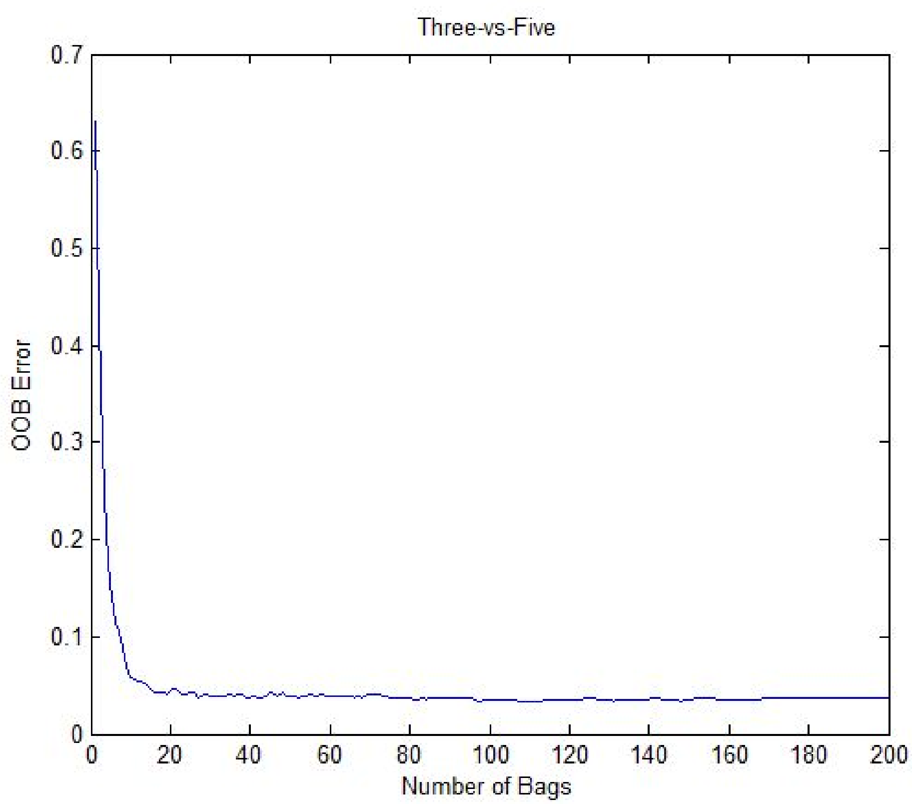
Tyler Schlichenmeyer- HW5



Cross-val error: 1.02%   OOB error: 0.45%



Cross-val error: 6.26%   OOB error: 3.71 %

results summary:

```
Working on the one-vs-five problem...

The test error for one tree is 0.0165
The test error for 200 bagged decision trees is 0.0118

Now working on the three-vs-five problem...

The test error for one tree is 0.1196
The test error for 200 bagged decision trees is 0.0920
```

d) As we create more bags, our out-of-bag error, which we are using to estimate Eout, rapidly decreases and rapidly approaches its asymptotic value after about 20-40 bags. Our bagging method predicted a better out of sample error than cross-validating a single decision tree (part a) and indeed our bagged ensemble performed better in practice on the test set (part b), though comparing the values from b to a we can see that our estimates from part a were noticeably optimistic estimates for the out of sample error. It is also interesting to note that 1vs5 was a much easier task than 3vs5, which makes sense intuitively since 1s should be relatively easy to detect because of their simple shape.

2) Looking at our data set and the attributes, it’s immediately clear we have a lot of superfluous information that will be implemented in our trees. So while bagging can help with inherently noisy data, I think that pruning will be more effective in eliminating less effective separators (for example, the classification of of a first grader is astronomically unlikely to be different than a fourth grader and this attribute distinction should not be included in our model). Therefore I decided to go with a pruning model rather than a bagging model to prevent overfitting to these parameters. To figure out the pruning level, I calculated which level minimizes the cross-validation error, and then pruned the tree at that level. The result was a 4% decrease in test error from a single decision tree.

error on unpruned tree: 23.96%
error on pruned tree: 20.04%