**Random Forrest**

A close up of a newspaper

Description automatically generated

A screenshot of a cell phone

Description automatically generated

(a). OOB Error as a function of the number of bags (from 1 to 200) for one-vs-three cases

A close up of a map

Description automatically generated

OOB Error as a function of the number of bags (from 1 to 200) for three-vs-five cases

A close up of a map

Description automatically generated

1(b).

>> OneThreeFive

Working on the one-vs-three problem...

The cross-validation error of decision trees is **0.0090**

The OOB error of 200 bagged decision trees is **0.0036**

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

The cross-validation error of decision trees is **0.0692**

The OOB error of 200 bagged decision trees is **0.0346**

1(c).

>> TestError

Working on the one-vs-three problem...

The test error of a single decision tree is **0.0163**

The test error of an ensemble of 200 decision trees is **0.0116**

Working on the three-vs-five problem...

The test error of a single decision tree is **0.1196**

The test error of an ensemble of 200 decision trees is **0.0859**

1(d).

From part a, we can observe that when the number of bags increases, the out of bag error decreases. When the number of bags reach a threshold, like 50 in our example, the out of bag error basically remains unchanged, or fluctuates around a certain value. We can conclude that when we have more bags, the performance will be better initially, and when we have enough number of bags, the performance will converge finally.

From part b, we can observe that for both cases (one-vs-three and three-vs-five), the out of bag errors of the random forest are less than the cross-validation errors of a single decision tree. From part c, we can observe that when we apply the trained model to the training data set, the test errors of an ensemble of 200 decision trees are also less than those of a single decision tree. Therefore, in both cases, we can conclude that the performance of a random forest with 200 bootstrapping datasets each associated with a max-depth tree is better than that of a single max-depth decision tree. With 200 decision trees and taking the aggregation of them, we can significantly reduce the high variance of a single decision tree, while maintain the low bias. It is reasonable that a random forest performs better.

Moreover, if we compare part b and part c, we can observe that for a single decision tree, the cross-validation error is less than the test error; for the random forest, the out of bag error is also less than the test error. This is also reasonable, because validation error/oob error are more optimistic than the actual test error.