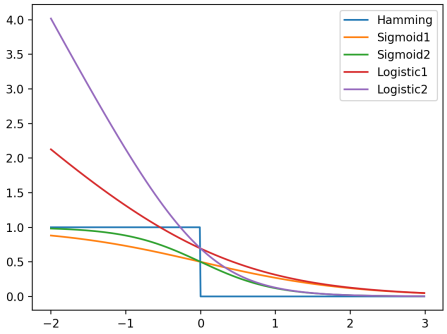
PS7 Machine Learning: Multiclass Classification, Bagging

Cai Glencross and Katie Li

March 23, 2018

1 Soybean Multi-class Classification



	Hamming	Sigmoid	Logistic
one-versus-all	41	48	42
one-versus-one	54	46	46
random (R1)	38	38	33
random (R2)	34	35	34

1. The random codes have less error likely because they encapsulate more complex relationships. They represent comparisons not necessarily from only one-to-one relationships but potentially across many classes.

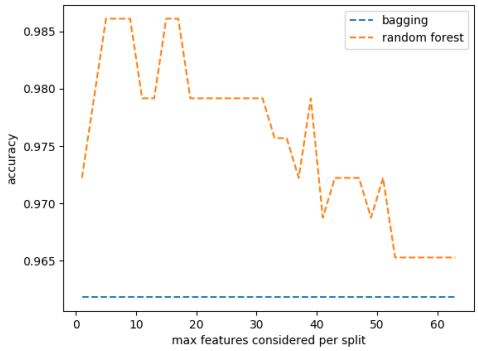
   As Hamming distance punishes all misclassifications the same, we can see that it performs best in one vs. all, since it is more likely that there will be some misclassifications, and we want to allow some outliers. For one vs one however it is less likely that any misclassifications will occur and therefore it is a disadvantage that Hamming Distance does not punish severe misclassifications, so sigmoid and logistic losses work better.

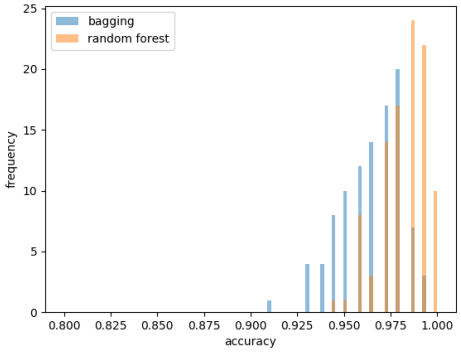
2. The random output code is the most suitable for this problem because it allows for more complex comparisons between multiple classes and not

simply one vs. one or one vs. all. For example, we could have a row with three +1's and all other negative ones, and fit a classifier that closely relates all of these together. The other output code types may be more prone to overfitting.

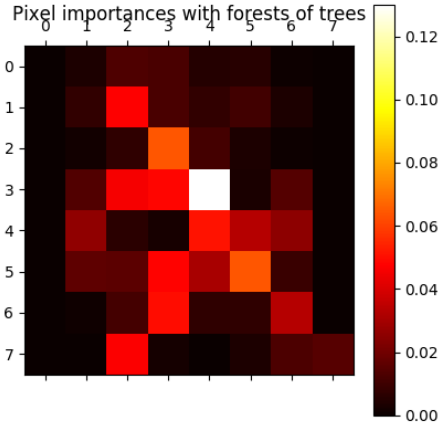
The random output code, unlike OVO, does not throw away any data. In addition, it does not contribute to data skew, which occurs in OVA, that does not classify well with imbalanced datasets.

2 Bagging Digits





(a.) We can see that bagging does not perform nearly as well as even the worst random forest. This is because the random forest uses the bagging approach but adds another layer of randomness in only considering a subset of features to split on for each branch. This helps because the greedy strategy of choosing the feature with the largest information gain to split on will not always produce the best decision tree, so only examining a subset gives us a greater chance to find the best decision tree. There is also more potential to overfit to training data using the greedy approach of picking the highest information gain because the feature in the training set with the highest information gain will not necessarily generalize to the test set. This is also the insight behind a lower number of features considered per split producing better results. The smaller the number of features (up to a point) produces more randomness and therefore less overfitting and less greedy strategies.



(b.) It was not surprising that the borders of the image were not particularly important. Initially, it was surprising that the middle pixel was most important but upon further reflection it appears that the middle pixel divides the dataset more or less in half. However , even upon further reflection it is difficult to gain intuition as to why the other important pixels are distributed the way they are.