Homework 3
CMPSC 165B - Machine Learning

Kenny Hoang Nguyen
X299187

February 2020

In my code, the way I find the best split is to try different limits for each features and see which split that minimizes the total impurity $P_l i(N_l) + P_r i(N_r)$ where $i(N) = 1 - \sum_j P(w_j)^2$ is the gini impurity and j is the index of the classes. The test basically send each row of data that is tested on the feature defined for each node and try to place the test data in a leaf node and choose the class that has highest probability in that leaf node.
Doing the 5-fold validation I see that by depth=4 the tree converges to an average accuracy of about 91% for the test data, and 100% for the same data the tree was trained with.
The 5-fold validation splits the training data given to us in 5 buckets where each bucket has balanced data for each class. Then in each buckets the rows are shuffled and 80% of the rows (data) is used for training and the rest for testing. This is done 100 times for each bucket with depth from 0 to 30 levels. The accuracy is then given as an average accuracy for all these runs. For each of the two data sets we had, the runs resulted in the two graphs given in figure 1 and 2. The blue line are the average accuracy for the training data, meanwhile the red line is the accuracy for the testing data. For both data sets, iriz and bezdekiris we see that the testing data have an average accuracy of about 91% which is somewhat good.

We see that in the graph for accuracy we get an accuracy peak at depth 3 for both data sets, but then the accuracy become lower again. This seems to be a case of overfitted data, we have trained the tree so that it classifies the training data perfectly, but it misclassifies the testing data more. Using an ensemble (random forest) instead of a single tree, we can make the tree shorter and use boosting or bagging to make the final decision so that we do not get that drop in the accuracy of the classification. As we have gotten here when we increases the max depth of the tree. This will then on average result in a higher accuracy than we got here.

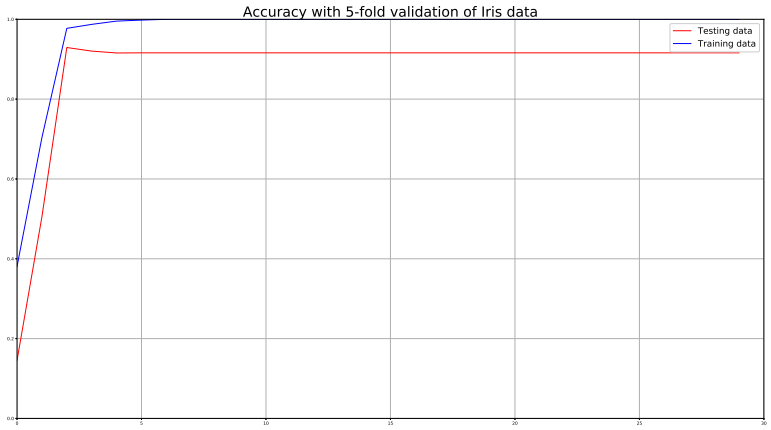

Figure 1: Accuracy for iris data

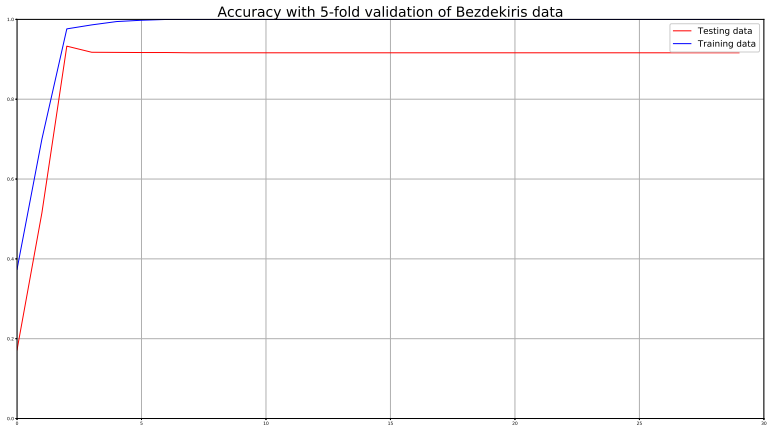

Figure 2: Accuracy for bezdekIris data