Danny North

3/6/16

CS 478

Dr. Martinez

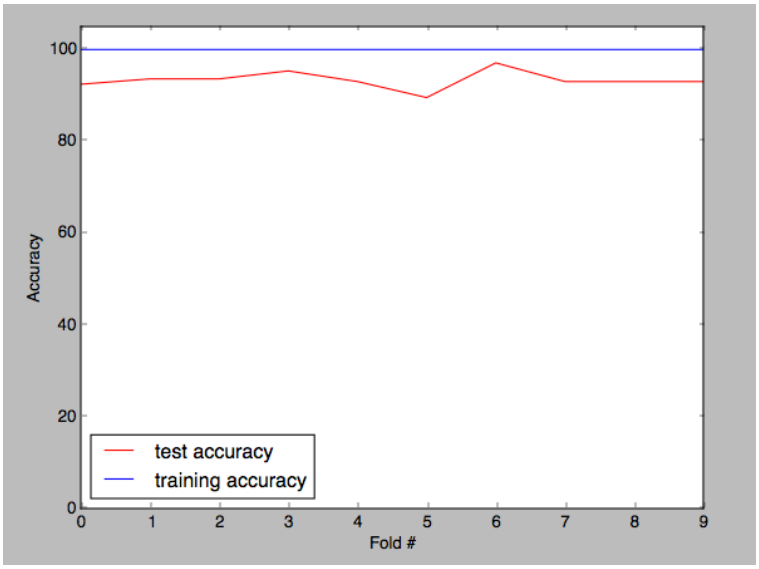
Decision Tree Lab

I. The Cars Problem

Using 10-fold cross validation, the result of the cars dataset can be seen below:

```
Training: [100.0, 100.0, 100.0, 100.0, 100.0, 100.0, 100.0, 100.0,
100.0, 100.0]
Training Average Accuracy: 100.000%

Test: [90.11627906976744, 93.6046511627907, 94.76744186046511,
94.76744186046511, 95.93023255813954, 93.02325581395348,
93.6046511627907, 94.76744186046511, 94.18604651162791,
92.44186046511628]
Test Average Accuracy: 93.721%
```



It's interesting to note in both tables that the training accuracy is 100% across the board. This is because we trained the data until we got pure leaf nodes, so we didn't stop until we were able to label the entire training set with 100% accuracy because we took it to its base form. The test accuracy stays in its correct range for both the cars and voting data.

For the cars dataset, the highest information gain and the tree's first split is on the "safety" attribute. It quickly labels all cars with LOW safety as "unacceptable". Both MED and HIGH safety decide to split on the "people" attribute. Both label any cars with 2 persons as "unacceptable" and split on the "buying cost". If the buying cost is

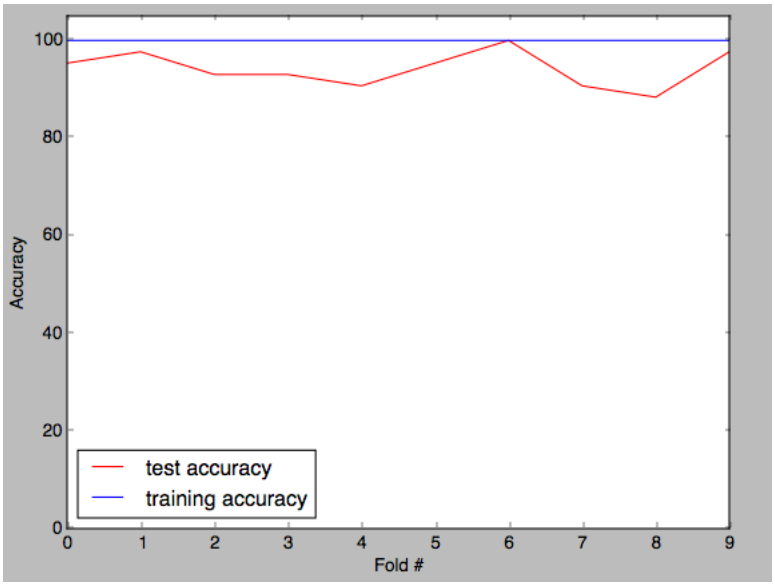
VHIGH then it splits on the “maintenance cost”. The decision tree seems to only label cars as “vgood” if the safety is medium or high, the buying and maintenance costs are medium or low, the luggage boot is medium or big, and there are 3 or more doors.

II. The Voting Problem

Using 10-fold cross validation, the result of the voting dataset can be seen below:

```
Training: [100.0, 100.0, 100.0, 100.0, 100.0, 100.0, 100.0, 100.0,
100.0, 100.0]
Training Average Accuracy: 100.000%

Test: [93.02325581395348, 90.69767441860465, 93.02325581395348,
95.34883720930233, 90.69767441860465, 90.69767441860465, 100.0,
88.37209302325581, 93.02325581395348, 95.34883720930233]
Test Average Accuracy: 93.023%
```



The way I handled unknowns in this problem was to include a new label of classification specifically for unknowns. I chose this approach because adding a new label made the tree more complete and lessens the risk of missing important value from the fact that it is unknown. This might include over fit into the tree because of unnecessary complexity, but for this problem it worked well.

The highest information gain for the voting dataset came from the “physician-fee-freeze” attribute. Those who voted NO were then split on the “adoption-of-the-budget-resolution” attribute. Those who voted YES or UNKNOWN for the “adoption-of-the-budget-resolution” attribute were put into a “democrat” label while those who voted NO were split further. Those who voted YES on the “physician-fee-freeze”

were then split on the “synfuels-corporation-cutback” attribute. Those who voted YES on immigration were labeled “republican” while those who voted NO or UNKNOWN continued splitting. Those who were UNKNOWN on the “physician-fee-freeze” attribute were split on the “mx-missile” attribute where those who voted NO were labeled “democrats” and those who were UNKNOWN were labeled “republicans”. Those who voted YES on “mx-missile” were split on the “anti-satellite-test-ban” attribute where those who voted NO were labeled “republican” and those who voted YES or UNKNOWN were labeled “democrat”.

III. Reduced Error Pruning

Below is a table regarding the reduced error pruning. From my experiments, the vote task was reduced significantly while the cars task did not produce better results upon pruning.

	# Nodes	Depth	% Accuracy
Unpruned-car-avg	376	7	94.1
Pruned-car-avg	376	7	94.1
Unpruned-vote-avg	54	8	94.0
Pruned-vote-avg	4	2	95.6

IV. Experiment

My experiment was to test how forcing maximum levels of tree depth changed accuracy. For this test I will use a table for both the vote and car data to see how forced depth changes accuracy. I will use 10-fold cross-validation and display the averages of each test. You can see from the experiment that the reduced error pruning seems to have worked as it was supposed to. The best accuracy for the voting set is depth 2, and the best for the car is depth 7.

Voting Dataset

Depth	% Accuracy
1	61.2
2	95.6
3	94.7
4	94.2
5	94.9
6	93.3
7	93.0
8	93.2
9	93.5

Cars Dataset

Depth	% Accuracy
1	70.0
2	70.1
3	76.6
4	81.1
5	85.9
6	92.6
7	93.7