David D'Attile
Prof. Kauchak
CS158
30 January 2022

Assignment 02: DT Experiment Writeup

1. *What is the accuracy of the random classifier on the Titanic data set from Assignment 01? To calculate this, generate a random 80/20 split and train the model on the 80% fraction and then evaluate the accuracy on the 20% fraction. Repeat this 100 times and average the result:*

   The random classifier achieved a classification accuracy of ~50.05% (7,157 correct classifications of 14,300 total attempts).

2. *What is the accuracy of your decision tree classifier on the Titanic data set with unlimited depth? As above, average the results over 100 random 80/20 splits:*

   The decision tree classifier with no depth limit achieved a classification accuracy of ~75.22% (10,756 correct classifications of 14,300 total attempts).

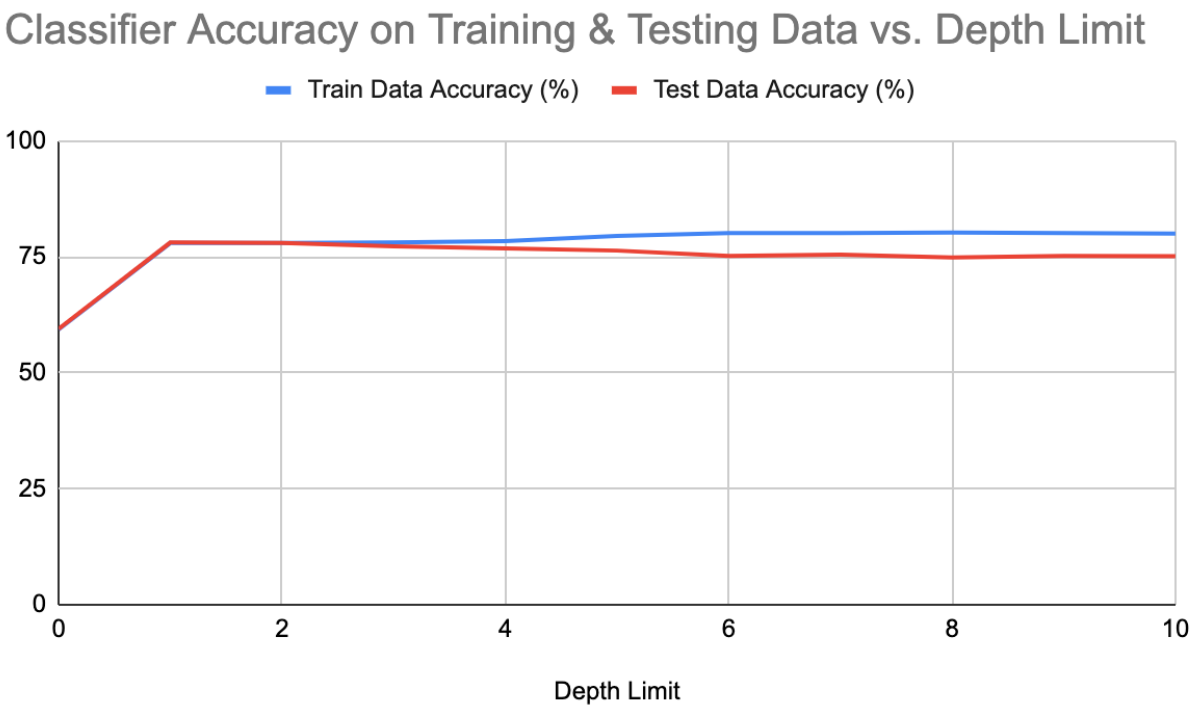
3. *What is the best depth limit to use for this data? To answer this, do the same calculations as above (average 100 experiments), but do it for increasing depth limits, specifically 0, 1, 2,..., 10. Show all of your results:*

   The decision tree classifier with a depth limit of 2 achieved a classification accuracy of ~78.75% (11,261 correct classifications of 14,300 total attempts). All results were as follows:

   - Depth limit of 0: ~59.16% (8,460/14,300)
   - Depth limit of 1: ~77.87% (11,136/14,300)
   - **Depth limit of 2: ~78.75% (11,261/14,300)**
   - Depth limit of 3: ~77.57% (11,092/14,300)
   - Depth limit of 4: ~77.09% (11,024/14,300)
   - Depth limit of 5: ~76.38% (10,923/14,300)

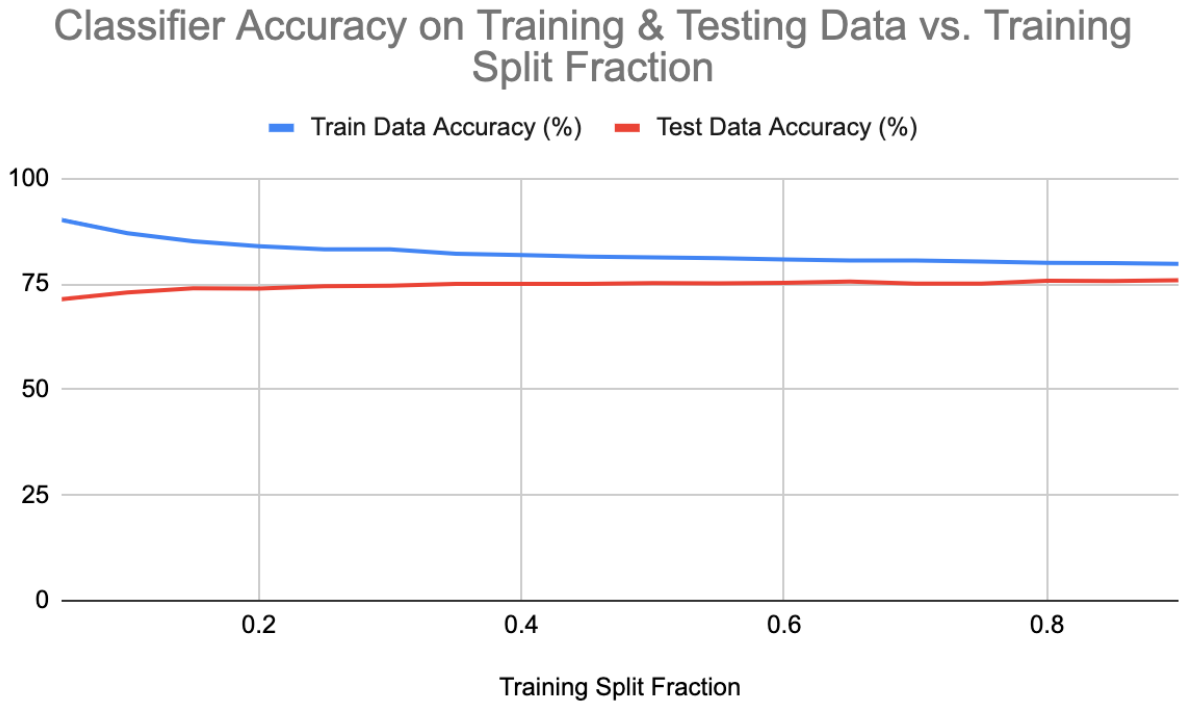
- Depth limit of 6: ~75.31% (10,770/14,300)
- Depth limit of 7: ~75.44% (10,788/14,300)
- Depth limit of 8: ~74.92% (10,713/14,300)
- Depth limit of 9: ~75.84% (10,845/14,300)
- Depth limit of 10: ~75.08% (10,736/14,300)

4. *Do we see overfitting with this data set? Repeat the experiment from question 3 with increasing depth (0, 1, ..., 10) and calculate the accuracy this time on both the testing data (like before) and the training data. Create a graph with these results and then provide a 1-2 sentence answer describing the graph:*



In this graph, we can see that after we reach a depth limit size rule of more than 2, our classifier begins to classify training data examples approximately 5% more accurately, even though it classified test data examples approximately the same for depth limits 0, 1, and 2. This is indicative of the decision tree classifier model potentially being “overfit” to the test data when compared to real world examples for higher depth limits.

5. *How does the amount of training data affect performance? To answer this, do the same calculations as above (average 100 experiments), but start with splits of 0.05 (5% of the data used for training) and work up to splits of size 0.9 (90% of the data used for training) in increments of 0.05. For these experiments use full depth trees, i.e. trees without any depth limit. Create a graph with these results and then provide a 1-2 sentence answer describing the graph.*



In this graph, we can see that training on fewer examples leads to more accurately classifying the examples used in training and less accurately classifying test data examples. As the training split fraction increases, the gap in classification accuracy between training and testing data examples narrows to ~5% with training data accuracy lowering and testing data accuracy growing.

6. *What does the training data size experiment tell us?*

   This experiment tells us that we can expect a better model when we leverage a higher portion of available data for training a decision tree model. Of course, appropriate precautions should be taken during training to mitigate “overfitting” against a training data set split on any ratio.