HW5 Questions

1) Based on the plot in step 4 part 1, can you determine if friends\_count and

followers\_count can be used for classification? Is there a pattern? Consider

zooming in on the plot to get a better picture.

There is a pattern that is quite obvious from the plotted data. All of the bots have near if not 0 friends and 0 followers whereas the normal users have ranges from 0-2 for friends and 0-1 for followers. After zooming in it becomes more apparent that the bots hug the axis of having a lot of followers or a lot of friends, but they do not have a balance whereas the regular users occupy the enter space of the graph or in other words have both of these not just one or the other.

2) What was the training accuracy? What was the testing accuracy? Based on this

can you conclude that the decision tree is a good Machine Learning model for

your dataset?

Training Accuracy: 0.9524064171122995

Testing Accuracy: 0.9526184538653366

Based off of these numbers I would sa that the decision tree is a great model for the dataset as there is a not any significant gap between the training and testing accuracy. This shows how the training from the model reflects into the testing and it is quite accurate when it comes to finding the bots

3) What percentage of the training and test data are bots? If you had a different

dataset where only 10% of the rows were bots, would 85% accuracy be good?

95% of the training and the test data were bots. If we were to have a test dataset with 10% of the rows being bots 85% accuracy would not be too good. This is as even without any sort of machine learning model the data set will represent 90% of the data being bots. So, in that case 85% is actually worse than what will be initially provide with the data.

4) When you split the data into train and test in Step 6, you might have assigned the

first 70% of the data to the training data and the last 30% of the data as test data.

Can you see any problems with doing this? Can you think of any better

alternatives?

There is one major issue with this in that the data then gets splits unevenly, after looking at the file for the data I noticed that the first 1200+ users are real and the rest are all bots. By splitting 70% and 30% this can possibly sway the data and not give the most accurate information. The best alternative would be to resort and clean up the information given in a way that when you split it 70 – 30 it is a more random chance of getting a bot versus getting a real user. By doing so it should provide a more accurate solution.

5) The Decision Tree Classifier learned this data set quite effectively. Can you give

an example of a classification task for which decision tree would perform poorly?

One example that I would think is too difficult for the decision tree classifier would be a problem where the data is linearly separated. This is as after reading it seems that decision tree classifiers are not good at handling problems in which the boundary is complex and not axis assign, along with when there is small data sets they tend to overfit, and they may not capture true patterns and get mixed up. My example would be represented by the flower example from the sides in class in which the data was divided by a negative y = x line shifted on the graph.

6) Often in data science you are just given a data set and a task (e.g., learn a

classifier for detecting bots). The instructions in the assignment made a lot of

decisions for you. List 5 examples of these decisions that the assignment made

for you. (In other words, if you were trying to solve this problem without this

homework guiding you, what would you have to decide yourself in processing

your data and designing your classifier?)

The first decision made in the instructions was through the data cleaning in which the instructions said to remove columns with 35% or more of missing values. The next was through the features that were selected for the decision tree classified (friend count, followers count, etc.). Along with deciding the data split as well of 70% training and 30% testing. Also the instructions stated the exact instructions or building features such as the bag of words given. Lastly was through the model choice and saying how we need to use a decision tree classifier.