Lindsay Cagarli and Anna Kurchenko

CSCI 420- Intro to Data Mining

HW 05

1. Who did what roles during the assignment?

It was a collaborative effort to complete this assignment, but in particular, Anna had more say in the actual building of the decision tree, and Lindsay was responsible for more data and documentation handling.

2. What was the maximum call depth of your final classifier?

The maximum depth that we used is 8.

3. Describe the decisions of your final trained classifier program (the resulting classifier). What were the most important attributes (the ones at the top)? Inspecting it, what does it tell you about the relative importance of the attributes? What is the most important attribute?

The most important attributes were RoofRack (Depth: 0), HasGlasses (Depth: 1), SideDents (Depth: 2). Most of these attributes are related to the state of the car. For example ‘SideDents’ show the car has been banged up from previous accidents which might indicate a risky driver. Roof-racks are stereotypically known as a marker of an aggressive driver. The third attribute ‘HasGlasses’ is a bit of a curve-ball since it leads us to make assumptions about the driver. It may indicate a driver who is more cautious on the road since they have worse vision and would be more conscious of adhering to safety rules. These top attributes suggest that the condition of the vehicle is an important factor when trying to identify an aggressive driver. The most important attribute was having a roof rack.

4. What was your choice of a minimum number of records in a node?

Minimum records in a node are set to 5.

5. What was your choice for a maximum splitting depth for the code?

Maximum splitting depth is set to 8.

6. What was your choice for a maximum node purity to stop at?

Maximum node purity is when a node reaches 90% of purity as recommended.

7. What features did you create, if any?

We did not create any features.

We used entropy and information gain ratio to help determine which feature would be the most effective.

8. Did you discover any features which could be ignored?

We discovered that attributes longHair, Wears_Hat, and foreign could be ignored as they showed a low correlation coefficient.

9. Generate a confusion matrix for the original training data:
How many aggressive were classified as aggressive?

- 2105

How many aggressive were classified as non-aggressive?

- 1195

How many non-aggressive were classified as non-aggressive?

- 216

How many non-aggressive were classified as aggressive?

- 6384

10. What was the accuracy of your resulting classifier, on the training data?

86.6%

11. What was the hardest part of getting all this working? Did anything go wrong? Did anything go very well?

The hardest part was raising our accuracy after getting an initial decision tree code working. It was overwhelming to manage all the moving pieces trying to fine tune things to raise the accuracy. We kept jumping between modifying test sets and the information gain function, since both of these areas were hard to know for sure if we got right.

Getting started was also challenging since we had to incorporate so many pieces. It did really help to start with the previous HW_4, but we didn't realize that was recommended until an hour or two of trying to reinvent the wheel.

One thing that went well was when we finally balanced the data set, our accuracy shot way up. This was such an easy fix and gave us 30% right on top of our number.

One thing that went very wrong was when we were getting started trying to use the validation data. We did not originally realize that the intent column values were missing, and since we didn't print out our target values once we processed the data, we kept running into index/null pointer errors in our big data frame that were very frustrating.

12. Conclusions: How did you clean your data? What pre-processing did you do? Did you create any new features to use? What did you discover? Anything else you discovered along the way? Did you run into the accuracy paradox? Was the data completely separable? Write full paragraphs and full sentences. Show strong evidence of learning.

We proceeded to balance the set to have equal representation of "pull over" and "let pass" intent values. Balancing the data before training gave our classifier better odds at distinguishing the two categories and led to more accurate results. We also converted the intent values "pull over" and "let pass" into binary to simplify calculations. We additionally cleaned our data by flooring all values for greater precision.

The data was also not fully separable as there was no clear cut boundary between "pull over" and "let pass". There is still some fear that there may be some overfitting happening as that can be influenced by multiple factors such as depth and node purity.

We discovered that data pre-processing is really the lunch-pin to getting a classifier working. This point was reiterated by Dr. Kinsman in class multiple times but we got to find out the hard way after trying to skip ahead. Anxious to get started on the decision tree, we didn't really look at the data too much and our approach ended up taking way more time than it could have if we followed the writeup recommendations carefully from the get-go.