Big Data Analytics

CSCI 720 - 02

Homework 5

Aditya Kalyan Jayanti
aj8582@rit.edu

Rohit Ravishankar
rr9105@rit.edu

**NOTE:- My laptop stopped functioning last Thursday, 18th October 2018, and because of this, I wouldn't have been able to submit this homework on time. I have informed Prof. Kinsman of the situation and sought his concurrence on completing this homework in collaboration with my classmate, Aditya Kalyan Jayanti. Please take due cognizance of the same.**

**- Rohit Ravishankar**

### 3.a. What stopping criteria did you use?

- When the purity of each node is greater than 95% we stop splitting.

### 3.b. Did you use any pruning or post-pruning?

- No

### 3.c. What splitting decision were you using?

- The splitting decision was based on the best weighted GINI index for the given subsection of the data.

### 3.d. What structure did your final decision tree classifier have? What was the if-else tree you got?

```
def my_classifier_function(file, columns_with_data):
    if columns_with_data["Sugar"] <= 19:
        if columns_with_data["Egg"] <= 12:
            if columns_with_data["Butter or Margarine"] <= 18:
                if columns_with_data["Baking Powder"] <= 2:
                    file.write('1\n')
                else:
                    if columns_with_data["cinnamon"] <= 11:
                        file.write('0\n')
                    else:
                        file.write('1\n')
            else:
                file.write('0\n')
        else:
            file.write('0\n')
    else:
        if columns_with_data["Canned Pumpkin_or_Fruit"] <= 26:
            if columns_with_data["Vanilla"] <= 6:
                file.write('0\n')
            else:
                file.write('1\n')
        else:
            file.write('1\n')
```

Aditya Kalyan Jayanti
aj8582@rit.edu
Rohit Ravishankar
rr9105@rit.edu

### 3.e. Run the original training data back through your classifier. What was the accuracy of your resulting classifier, on the training data?

- The accuracy was found to be 96.72%.
- There were two misclassifications, and given 61 records our accuracy would be $\frac{59}{61} * 100$

### 3.f. Did your program actually create the classifier program, or did it just generate the attribute list and thresholds for you to hand-code in

- Yes, it actually created the classifier program.

### 3.g. What else did you learn along the way here?

- We learnt how to write a program which writes another runnable program.
- We learnt on which attributes should we split, how to split on the attributes & when to stop splitting.
- We learnt how to build a decision tree, without hard coding the classes & values.
- We learnt how to calculate the accuracy of a given model.

### 3.h. What can you conclude?

- Based on the values of the ingredients for the recipe, we can conclude that if a recipe has certain values of ingredients, it can be classified as a cupcake or a muffin.
- If sugar <= 19, egg <= 12, butter/margarine <= 18 & baking_powder <= 2 or cinnamon > 11, it is a muffin.
- If sugar > 19 & vanilla > 6 & pumpkin/fruit > 26, it is a muffin.
- In all other cases, it is a cupcake.
- We can evidently see this from the aforementioned code section of the decision tree.

### 4. Accuracy: Classification accuracy (TP+TN) of validation data is counted toward part of your grade

- The accuracy was found to be 96.72%.
- There were two misclassifications, and given 61 records our accuracy would be $\frac{59}{61} * 100$