CSCI 720 - Big Data Analytics

Nihal Surendra Parchand

np9603@rit.edu

Homework 6 Decision Tree

a. Describe the structure of your final program (the resulting classifier).
The resulting classifier program first imports the pandas library for reading the test data and store it in a dataframe called data_df. A result list is initialized to store the results from the classification on the test data. Then, we iterate through each row in the dataframe and store the corresponding values for each column i.e. (Flour, Sugar, Oils, Proteins) in different variables. The if-else statements are generated from the decision tree trainer program. For every depth, best split/threshold is calculated by minimizing the weighted entropy. The resulting classification result is appended for each row and the target attribute (RecipeType) is identified by going through the nested if-else ladder. We append 0 for Muffin and 1 for Cupcake. This result list is then iterated and each value is written in another file called HW06_Parchand_Nihal_MyClassifications.csv. Each line contains either a 0 or 1 as classification result for every row in the test data.

b. What was the accuracy of your resulting classifier on the training data?
The resulting classifier was able to achieve an accuracy of ~99% on the training data.

c. What was the hardest part of getting all this working?
The hardest part of this assignment was to break down the whole problem into sub divisions and work on it independently. After figuring out the first best split, I had to recursively find the best splits for building the entire decision tree. At first, I was using nested for loops which caused the time complexity to go beyond $O(n^3)$. Debugging that code was very time consuming and it took almost 5 minutes to run the whole program for just finding the first best split. So, I decided to split data and work on each column values independently. Another problem that I faced was in deciding how I will store the decision tree results.

d. Did anything go wrong?

I was stuck because of a wrong indentation in my code due to which I was getting Flour as the best split attribute for a lot of threshold values. Debugging helped me to quickly fix that issue.

e. Discussion:

What does the number 23 have to do with anything?

What does this have to do with math, statistics, or decision making?

Or, did the professor pull this number out of a hat?

The Bible has mentioned about a verse 10.23 which is also called as Decision Making.

Jeremiah 10:23 O Lord, I know that the way of man is not in himself: it is not in man that walketh to direct his steps.

f. Conclusions

After working on this assignment, I learned that storing dataframe values into numpy arrays are one of the most efficient ways to work for data with large number of records. Also, writing my own decision tree algorithm from scratch cleared a lot of my concepts like there can be splits based on the same columns in different depths as well. Writing a program to write another program also tested me with how to properly write nested if else statements. The indentation part was a bit confusing at the beginning.