Artificial Intelligence

Assignment 5

Team Members-

Jayakrishna Anvesh Paidy 112675567

Shubham Bagi 112672171

### Clickstream Mining with Decision Trees

The project is based on a task posed in KDD Cup 2000. It involves mining click-stream data collected from Gazelle.com, which sells legware products. Your task is to determine: Given a set of page views, will the visitor view another page on the site or will he leave?

We have run it for different threshold. The values are:

1. For Threshold = 0.01

Number of Nodes: 71

Accuracy reported: 74.82%

1. For Threshold value = 0.05

Number of Nodes – 76

Accuracy Reported: - 75.01%

1. For Threshold value =1

Number of Nodes - 1371

Accuracy Reached - 74.52%

Observation found - Number of Nodes increases when the Threshold increases.

So, we can see that there were many nodes which were expanded when the Threshold value was set to 1. This means that almost all the nodes in the tree was expanded and hence it took a long time to run. But for the case of threshold value = 0.01 and 0.05 we noticed that there was quite a lot of pruning happened and hence the time taken to traverse over the whole tree was less.

Hence, we can say that we get an accuracy score of 74.82%.

### 2. Spam Filter

The dataset we will be using is a subset of 2005 TREC Public Spam Corpus. It contains a training set and a test set. Both files use the same format: each line represents the space-delimited properties of an email, with the first one being the email ID, the second one being whether it is a spam or ham (non-spam), and the rest are words and their occurrence numbers in this email. In preprocessing, non-word characters have been removed, and features selected like what Mehran Sahami did in his original paper using Naive Bayes to classify spams.

Approach:

1. We were given training data where we had to learn and classify the data according to ‘SPAM’ and ‘HAM’ using Naïve Bayes Classifier.
2. Initially, we computed the prior probabilities of SPAM and HAM and the likelihoods of each word (conditional probability also)
3. So, the conditional probability was calculated depending on the count of SPAM and HAM counts and the total mail counts.
4. Once this was done, we now had to classify the test data according to its category.
5. The labels were separated as this is useful to compute the accuracy of our prediction.
6. Finally, as we parsed through the test data, we had to classify the data with the help of the conditional probability.
7. Once classified as SPAM or HAM we need to check the accuracy with the label stored from the test data.
8. Recall, Precision was also calculated accordingly.

There were cases were some of the words found in the TEST DATA was not present in the dictionary of words got from the TRAINING SET. Hence as we need to classify it accordingly we used the concept of **LAPLACE SMOOTHING** to overcome this problem.

An addition of 1 was added if the word was not found in the dictionary. But there was no change in accuracy which was observed.

Important Step which gave us a better accuracy.

1. We did calculate the mean and standard deviation of the word count so that we could remove the **OUTLIERS** from the data. So, we keep the data only which is within **2 STANDARD DEVIATIONS** of the distribution. By doing this we got an **increase** in the accuracy.

Accuracy before outlier was added: 89.3%

Accuracy after outlier was added: 90.9%

Precision before outlier was added: 85.67%

Precision after outlier was added: 86.43%