CSE537: Artificial Intelligence, Fall 2018

Assignment 5

(Due: 08 December 2018)

Submission Details			
Name	SBU ID	Email Id	% Contribution
Shobhit Khandelwal	112074908	shobhit.khandelwal@stonybrook.edu	50 %
Vivek Bansal	112044493	vivek.bansal@stonybrook.edu	50 %

Answer 1) Clickstream Mining with Decision Trees

Approach:

We first use the training data to build the decision tree using ID3 algorithm. Also we use chi-square criterion to cut off/prune the tree at a node.Following are the results obtained by running the classifier for various values of p-value

TestCases:
1. P-Value = 0.05

Results
Num Nodes = 356
Tree prediction accuracy: 0.73452
Output file prediction accuracy: 0.73452
Tree prediction matches output file

2. P-Value = 0.01

Results
Num Nodes = 236
Tree prediction accuracy: 0.73536
Output file prediction accuracy: 0.73536
Tree prediction matches output file

3. P-Value = 1 (0.999)

Results
Num Nodes = 26626
Tree prediction accuracy: 0.59336
Output file prediction accuracy: 0.59336
Tree prediction matches output file

(2) Explain which options work well and why?

From the results obtained, it can be seen that lower p-value gives good accuracy and creates less number of tree nodes. This can be attributed to the fact that pruning the tree reduces the over fitting problem and thus gives higher accuracy with the test data. Also the time it takes to build the tree with p-value close to 1 is very high because the complete tree has to be built. Therefore p-value = 0.01 which is builds a decision tree that is neither to over fit nor to under fit and hence gives the best predictions.

Answer 2) Naive Bayes for Spam Detection

Approach:

To precisely classify mails as spam, I have created a dictionary of unique words from our training set. Then I calculated total number of mails available and number of spam mails and ham mails. This is done to get the probability P (spam | D) and probability P(ham | D). Also I have calculated P(word | spam) for each of the words in the spam mail. Similarly I calculated P (word | ham) for each of the words in the ham mail. Then, I applied naive Bayes to get the estimate of mail type.

Output:

Spam Detection Precision : 95.6896551724
Ham Detection Precision : 87.1428571429