COMP 4106 Assignment 3

Luke Harper 100886836

Artificial Data Generation

The artificial data generation was generated in a dependent manner for 4 classes. First a tree was implemented with a 10 nodes arranged in a balance form (Figure 1) representing the dependent relationship between features. Second, a random generation of 10 pairs of probabilities from 0 to 1 were performed for each class. This resulted in 40 random number pairs. These random numbers represented the probability that a feature would be 0 given the parent's value of 0 for the first element in the pair or a parent's value of 1 given the second element in the pair. These random pairs were then assigned to each feature in the tree to set the dependence tree probabilities. Finally the features for the data were generated by running through the dependence tree.

Classification Algorithms

All algorithms used 5-fold validation for training and testing.

Independent Bayesian classification was the first algorithm used to classify the data. The training method involved getting the probability that each feature is in a classification algorithm, and then checking the probability a set of features is present in each classification algorithm for the testing data. The classification with the highest probability is chosen.

Dependence tree classification was the second classification algorithm. This involved creating a complete graph with all of the features. The graph was then assigned edge weights using the Expected Mutual information Measure. A Maximum spanning tree was created using Prim's algorithm and these edge weights to determine a dependence tree. For training, the generated dependence was duplicated for each class. The dependence tree was then used to calculate the probability of a given feature given the parents feature.

The final classification algorithm used was decisions trees. A tree was pre generated by recursively determining the feature that had the greatest information gain given the previous features. This was done through entropy calculations. For the entropy's positive and negative portions were two different classes. A tree was generated for every combination of classes. Once the tree was generated the testing data was run through a pairwise comparison of trees to determine the most likely classification

Artificial Data Generation Results

The artificial data generation classification produces some interesting results. The decision tree produces the best results at nearly 80% accuracy. The Independent Bayes classification produces the next best accuracy at around 75%. Finally the dependence tree was the least accurate around 72%. These results are attributed to the balances nature of the initial tree and the random probabilities / data. The classifications were not different enough to produce very different results. This led to a very different tree being generated by the dependence tree (Figure 2).

For the real data, the results are more expected. The dependence tree produces the best result, followed by the decision tree and finally the independent Bayes was the least accurate. The thresholding mechanism used was taking the average value for each feature in each class. If the data has a larger value than threshold it was assigned a 1, otherwise it was assigned a 0. This thresholding produced very similar data sets for each class. Additionally the limited number of samples used resulted in a very poor accuracy. 37% for dependence trees, 27% for decision trees and 22% for independent Bayes.

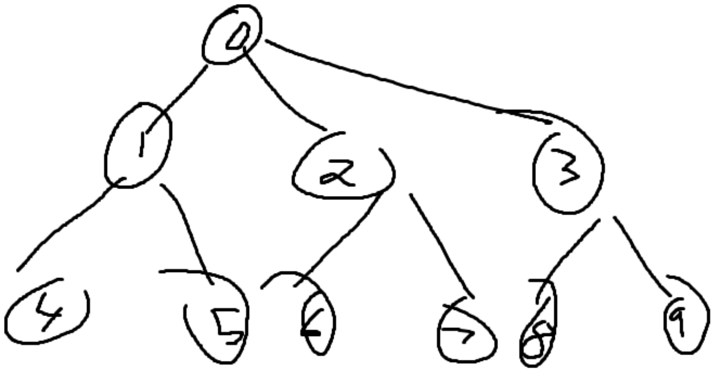

Figure 1

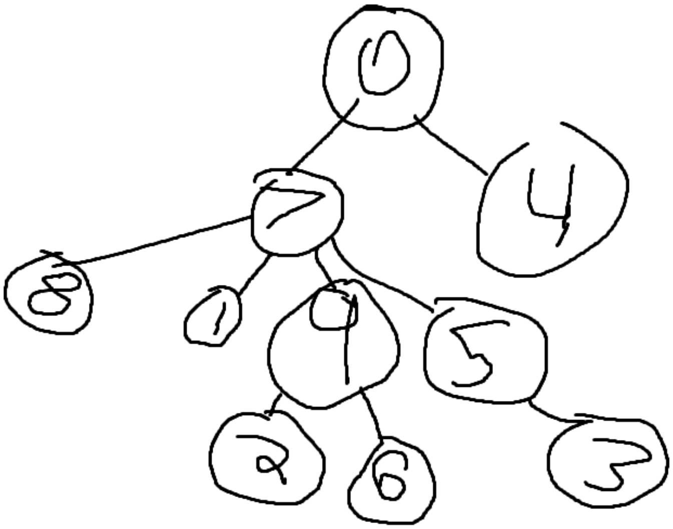

Figure 2