# Assignment 1 Report

Tianyu Zhang Z5135099

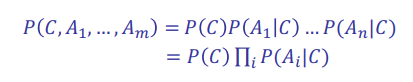
The main difference between Bayesian networks and Naive Bayesian networks lies in the assumption of independence between variables. Bayesian network uses arrows to describe the correlation between features, while naive Bayesian assumes that all features are independent, and all features are only related to the outcome variable.

Bayesian network of the dataset:

Diagram

Description automatically generated

With the independent assumption, each feature node only has one parent node (the output variable), therefore the joint probability of all variables becomes:



For both algorithms, the time complexity of MPE calculation is O(np), where p is the row number of the largest conditional distribution table. In Naïve Bayesian network, each variable only has BC as parent node, so the largest table is

|outcome\_space(variable)| \* |outcome\_space(BC)|

For normal Bayesian network, each variable has multiple parents, so the largest conditional distribution table will be:

Therefore, the Naïve Bayesian network has less runtime, in task 6, the normal Bayesian network took about 7 second to classify all data, the Naïve Bayesian network only took about 3 second in task 10 to perform the same task.

Furthermore, since the Naïve Bayesian network need less parameters in the joint probability calculation, it requires less memory comparing to the normal Bayesian network. These two algorithms have similar coding complexity since they just use the same Bayesian and chain formula on different networks, hence they have similar code.

The prediction accuracy of the Naive Bayesian network is influenced by the naive independence assumption because the variables may not actually be independent. To test the independence between features, we conducted chi-square independence tests for features other than BC.

The following figure shows the Chi-square test result. Numbers in the image are the p-value of the hypothesis that two features are independent. As a result, most features have p-value below 0.05, which means most features are dependent with each other.

A picture containing chart

Description automatically generated

As a result, the training accuracy of normal Bayesian network was about 0.8423, the cross-validation accuracy was about 0.85. On the other hand, affected by the independence assumption, Naïve Bayesian network had lower performance of 0.7926 training accuracy and 0.8 cross-validation accuracy.

In conclusion, comparing to normal Bayesian network, Naïve Bayesian network has lower accuracy, shorter runtime, need less memory, and has similar code complexity.