MACHINE LEARNING

LABORATORY SESSION 1: BAYESIAN NETWORKS

Pablo Rosales Rodríguez

Student Number: 213420

The initial dataset must be divided into two subsets: one for training, the other one for testing. To accomplish such a task, a python script has been created.

Given the two subsets, three different algorithms must be used for learning the Bayesian network, over the training set. Then there must be an evaluation phase using the testing subset.

NPC

According to the HuginExpert tutorial, NPC stands for Necessary Path Condition. A criterion for improving previous algorithms such as the PC one. As it is exposed in the mentioned source, *"the necessary path condition says that in order for two variables X and Y to be independent conditional on a set S, with no proper subset of S for which this holds, there must exist a path between X and every Z in S (not crossing Y) and between Y and every Z in S (not crossing X)".* NPC is a constraint-based algorithm: the potential dependencies between nodes are the starting point for building the structure. An important fact about this algorithm is that the user can specify the desired level of significance for the performance.

The learning method followed by the software is the Estimation Maximization. It needs a prior distribution in order to complete the estimation phase. Then, the algorithm maximizes the likelihood of the whole dataset.

At first, the learning phase is accomplished on the training set, which contains the 80% of the initial data. As a result, the following data and graph are obtained.

Log-likelihood	-146.9
AIC	-161.9
BIC	-177.2

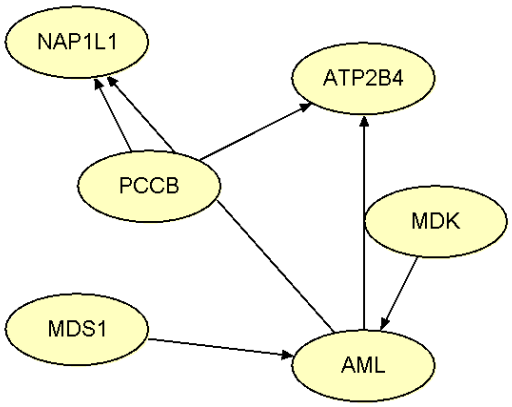

Figure [1]. Bayesian network obtained when using NPC algorithm.

When performing the testing phase, the results are the following:

Log-likelihood	-39.03
AIC	-54.03
BIC	-59.34

The ROC curve represents the true positive rate on the y axis and the false positive rate on the x axis. Obviously, having a curve closer to the top left corner will be preferable. The area under the ROC curve is a key indicator for comparing different algorithms. As greater this figure, better the prediction. To conclude, the ROC curve can be interpreted as a measure of the performance of the selected variable as a classifier.

For the estimated network, taking as classifier AML, the following ROC curve is obtained. It only provides three outliers.

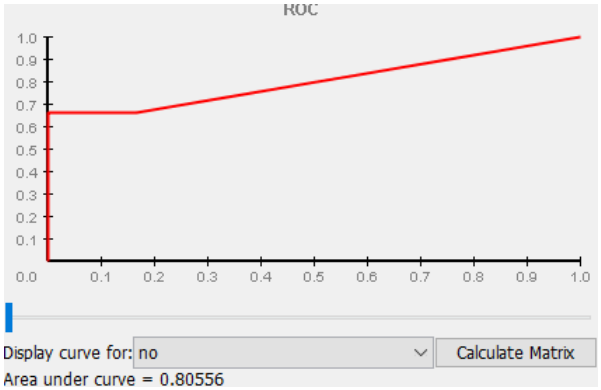

Figure [2]. ROC curve obtained when using NPC algorithm.

True \ Predicted	Yes	No
Yes	6	0
No	3	6

According to the obtained confusion matrix, the following performance measures can be calculated:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{12}{3 + 6 + 6} = 0.8 = 80\%$$

$$Precision = \frac{TP}{TP + FP} = \frac{6}{6 + 3} = 0.667 = 66.7\%$$

$$Recall\ or\ sensitivity = \frac{TP}{TP + FN} = \frac{6}{6} = 1 = 100\%$$

GREEDY SEARCH-AND-SCORE

It is a score-based model selection. The algorithm searches through the space of possible networks and assign to each of the models a score. The chosen structure is the one that maximizes the score.

The figure [3] represents the obtained graph, after performing the learning phase.

Log-likelihood	-150.13
AIC	-161.13
BIC	-172.36

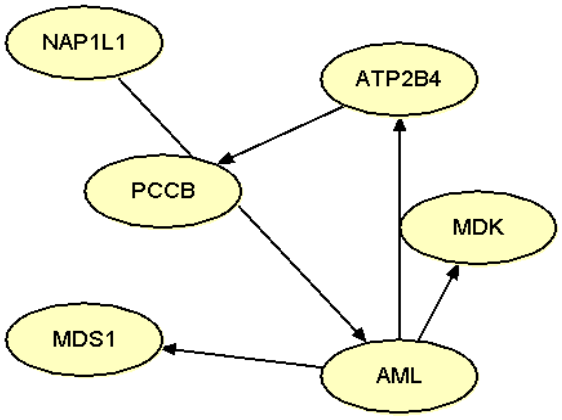

Figure [3]. Bayesian network obtained when using greedy algorithm.

The following values are obtained during the testing phase:

Log-likelihood	-37.18
AIC	-46.18
BIC	-49.36

Analyzing the AML ROC curve, the following figure is obtained:

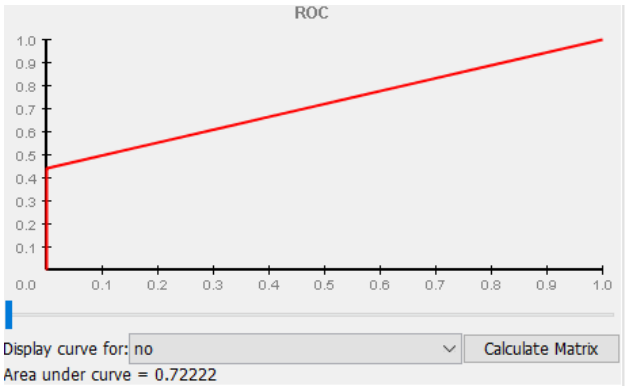

Figure [4]. ROC curve obtained when using greedy algorithm.

True \ Predicted	Yes	No
Yes	6	0
No	4	5

$$Accuracy = \frac{11}{6+4+5} = 0.733 = 73.3\%$$

$$Precision == \frac{6}{6} = 1 = 100\%$$

$$Recall\ or\ sensitivity = \frac{6}{6+4} = 0.60 = 60\%$$

The number of outliers has increased. The area under the ROC curve is smaller than in the previous case, so, in advance, this learning method could be considered to have a worse performance than the NPC one.

FIXED NAIVE BAYES STRUCTURE

The structure is manually modelled, fixing the constraints between the different nodes. The network is represented in figure [5]. Given AML, all the other variables are independent.

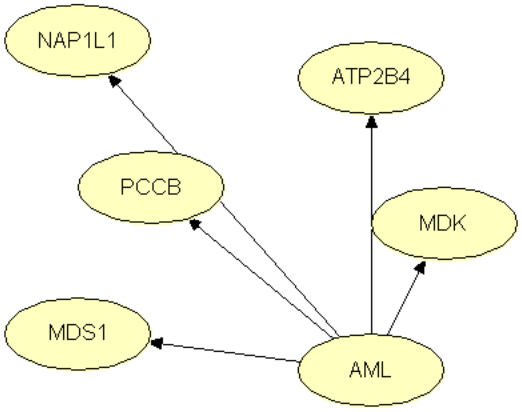

Figure [5]. Bayesian network obtained when using Naive Bayes algorithm.

When, directly performing the learning phase, these are the parameters obtained:

Log-likelihood	-45.17
AIC	-82.17
BIC	-95.27

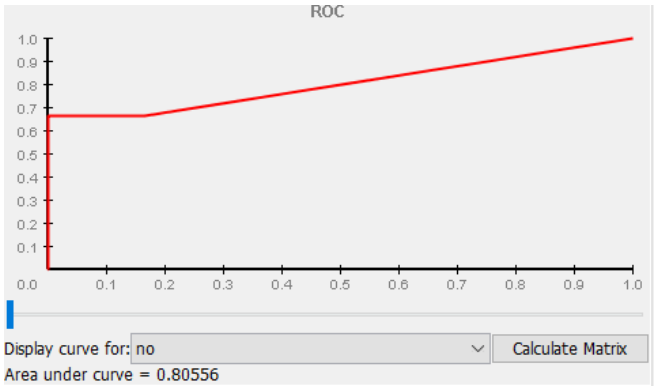

Figure [6]. ROC curve obtained when using Naive Bayes algorithm.

True \ Predicted	Yes	No
Yes	5	1
No	2	7

$$Accuracy = \frac{12}{5+1+2+7} = 0.8 = 80\%$$

$$Precision == \frac{5}{5+2} = 0.714 = 71.4\%$$

$$Recall\ or\ sensitivity = \frac{5}{5+1} = 0.833 = 83.3\%$$

Regarding the computed indicators and the area under the curve, the Naive Bayes method seems to offer a better performance than the previous one.

CONCLUSION

As a conclusion, the NPC algorithm offered the best performance of the three, followed by the greedy search-and-score algorithm. The worst option is the Naive Bayes algorithm.