Assignment 1: Bayesian Network

Francesco Penasa

December 5, 2019

1 Introduction

The first assignment of the Machine Learning course consist in a comparison of different learning methods on Bayesian Network. In this assignment we will compare NPC, Greed Score-And-Set and Naive Bayes Structure on the given dataset 'leukemia.dat'. In order to do this we will use the programming language Python3 and the tool Hugin Lite. We used Hugin lite to train the Bayesian Network with different types of algorithm and to test them while Python3 granted us the ability to split the dataset in train set and in test set, such action has been done with the piece of code showed below.
In the following sections we explore the processes followed for each algorithm and the results obtained with the usage of each of them.

split_dataset.py

```
import pandas as pd
# import csv
data = pd.read_csv('leukemia.dat')

# random split in 80% train, 20% test
train_sample = data.sample(frac=0.8, random_state=42)
test_sample = data.drop(train_sample.index)

# export csv
train_sample.to_csv('leukemia_train.dat', index=False)
test_sample.to_csv('leukemia_test.dat', index=False)
```

2 NPC

2.1 Learning

For the learning phase we have used the file `leukemia_train.dat` containing 58 cases. Following the Learning Wizard in Hugin Lite and using the default values with three iteration we obtained the following values: Log-likelihood: -164.207, AIC: -186.207, BIC: -208.872.

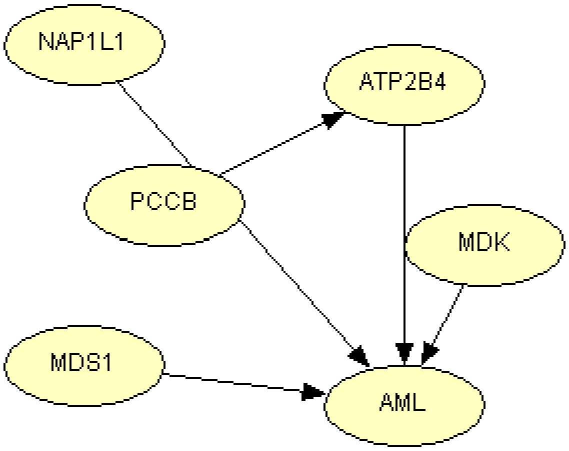

Figure 1: NPC Network

2.2 Results

Using the Analysis Wizard in Hugin Lite to predict the labels of the file `leukemia_test.dat`, containing 14 cases, we receive five true positive, seven true negative, two false positive and zero false negative. Such results from the confusion matrix can be rewritten as $accuracy = 0.85714$, $precision = 0.71429$, $recall = 1$.

Table 1: Confusion Matrix of test set with NPC.

Prediction	**Yes**	**No**	*Actual*
Yes	5	2	
No	0	7	

1. Number of cases: 14
2. Error rate: 14.29
3. Avg. Euclidian distance: 0.23222
4. Avg. Kulbach-Leibler divergence: 0.35152
5. Accuracy: 0.85714
6. Precision: 0.71429
7. Recall: 1

3 Greedy Search-And-Score

3.1 Learning

For the learning phase we have used the file `leukemia_train.dat` containing 58 cases. Following the Learning Wizard in Hugin Lite and using the default values with three iteration we obtained the following values: Log-likelihood: -158.429, AIC: -169.429, BIC: -180.761.

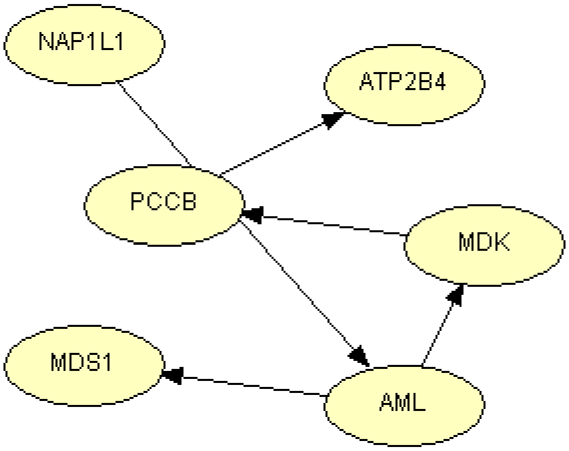

Figure 2: Greedy Search-And-Score Network

3.2 results

Using the Analysis Wizard in Hugin Lite to predict the labels of the test set we receive five true positive, seven true negative, two false positive and zero false negative. Such results from the confusion matrix can be rewritten as $accuracy = 0.71429$, $precision = 0.55556$, $recall = 1$.

Table 2: Confusion Matrix of test set for Greedy Seach-And-Score.

Prediction	**Yes**	**No**	*Actual*
Yes	5	4	
No	0	5	

1. Number of cases: 14
2. Error rate: 28.57
3. Avg. Euclidian distance: 0.32067
4. Avg. Kulbach-Leibler divergence: 0.44493
5. Accuracy: 0.71429
6. Precision: 0.55556
7. Recall: 1

4 Fixed Naive Bayes Structure

4.1 Learning

For the learning phase we have used the file `leukemia_train.dat` containing 58 cases. We have built the basic structure of the Bayesian Network assuming the features independent of each other given the label. Then we used the EM-Learning Wizard in Hugin Lite on the training set initializing the experience table to 1/Pa(x) for all the nodes of the network. After three iteration we obtained the following values: Log-likelihood: -161.329, AIC: -172.329, BIC: -183.661.

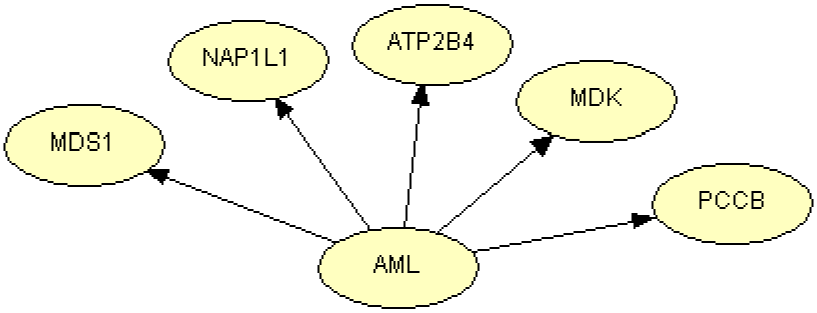

Figure 3: Fixed Naive Bayes Structure Bayesian network

4.2 Results

Using the Analysis Wizard in Hugin Lite to predict the labels of the file `leukemia_test.dat`, containing 14 cases, we receive five true positive, seven true negative, two false positive and zero false negative. Such results from the confusion matrix can be rewritten as $accuracy = 0.85714$, $precision = 0.71429$, $recall = 1$.

Table 3: Confusion Matrix of test set with Naive Bayes Structure.

Prediction	**Yes**	**No**	*Actual*
Yes	5	2	
No	0	7	

1. Number of cases: 14
2. Error rate: 14.29
3. Avg. Euclidian distance: 0.24563
4. Avg. Kulbach-Leibler divergence: 0.38629
5. Accuracy: 0.85714
6. Precision: 0.71429
7. Recall: 1

5 Conclusion

The table displayed below compares the results obtained with the three methods. Given the small dataset used both in training and testinig we encounter a similarity between the NPC and Naive Bayesian Structure, while the quality of the predictions given by the Greed Search-And-Score method are lower in respect to the other two.

	NPC	**GreedySAS**	**NBS**
Accuracy	0.85714	0.71429	0.85714
Precision	0.71429	0.55556	0.71429
Recall	1	1	1
Avg. Euclidian distance	0.23222	0.32067	0.24563
Avg. Kulvach-Leibler divergence	0.35152	0.44493	0.38629