REPORT OF HEART DATA

Maurizio Franchi

June 2015

In this report, I consider the data file *heart.dat* that contain twenty-two Boolean features (x1, . . . , x22) and one Boolean label, named y that can be 0 or 1.

Using the program *awk* I created two sets of data named *test.dat* and *train.dat* with 100 examples. To obtain the same percentage of 0 and 1 in both sets it was divided the *heart.dat* in two file: one contains only 0 and the other only 1. The order of the data in the files are randomized to remove temporal correlations. After these procedures the same number of 0 and 1 were divided in the two sets (*test.dat* and *train.dat*) in this way we have that:

- the set *test.dat* contains 27 lines that have y = 0 and 73 lines that have y = 1
- the set *train.dat* contains 28 lines that have y = 0 and 72 lines that have y = 1.

The data are mixed to avoid a block of only 0 and a block of only 1.

In *Hugin Lite* I loaded as Data Source the file *train.dat*, that has comma as separator symbol. To learn the structure of the model for the specified data, I selected firstly *NPC (Necessary Path Condition)* and then *Greedy search-and-score* algorithms. For both cases, I decide leave the default values and then I initialized all Experience tables with user-defined value and set the Experience count to zero.

The Figure 1 is the network of NPC for train

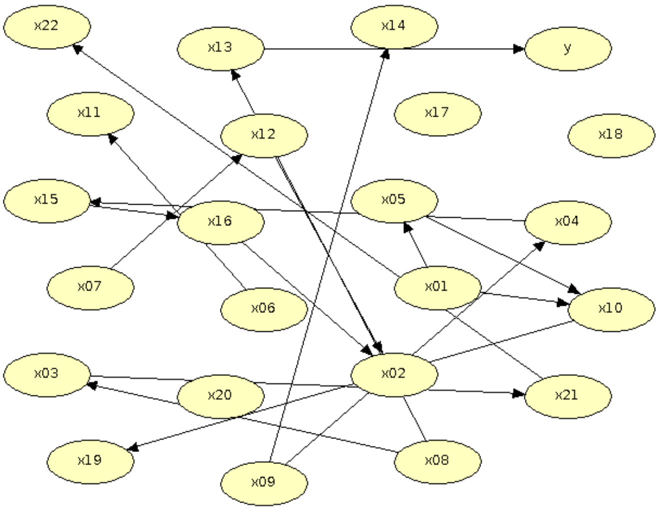

Figure 1: NPC algorithm for train

To generate the simulated cases, the number of cases was set to 100, because the *train.dat* contains 100 samples, and the percent missing value was set to the default value i.e. 5. After having simulated, I see in the Data Source the minimum and the maximum value of the conflict to see what was the range of conflicting data set. The the minimum value of the conflict is -9.83542 of the 34th case and the maximum value is 4.23422 of the 24th case. If I consider, for instance, the Data Accuracy of the node x01, I have all cases that have a value for the probability of evidence (P(x01 = “state in case”)).

To prediction if the state is ‘0’ or ‘1’, I used the specific state with a belief greater or equal the ROC (Receiver Operating Characteristic) cutoff threshold and used maximum belief.

If the predicted state is ‘1’ there are four cases where the actual value is ‘0’, so the probability, given x01 as observed, is less than 0.75; while there are fifty-five cases where the actual value is ‘1’, so the probability is more than 0.75. If the predicted state is ‘0’ there are two cases where the actual value is ‘1’, so the probability, given x01 as observed, is less than 0.96, while there are thirty-nine cases where the actual value is ‘0’, so the probability is more than 0.96.

Moreover, the error rate is 6.00, the average Euclidean distance is 0.10797 and the average Kulbach-Leibler divergence is 0.20870.

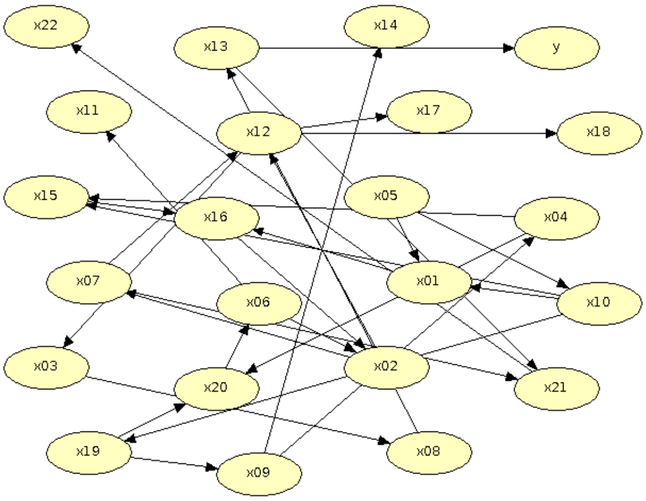

Figure 2: Greedy search-and-score algorithm algorithm for train

The Figure 2 is the network of Greedy search-and-score for train

To generate the simulated cases, the number of cases was set to 100, because the *train.dat* contains 100 samples, and the percent missing value was set to the default value i.e. 5. After having simulated, I see in the Data Source the minimum and the maximum value of the conflict to see what was the range of conflicting data set. The minimum value of the conflict is -14.75514 of the 34th case and the maximum value is 4.16563 of the 40th case. If I consider, for instance, the Data Accuracy of the node x01, I have all cases that have a value for the probability of evidence (P(x01 = “state in case”)).

To prediction if the state is ‘0’ or ‘1’, I used the specific state with a belief greater or equal the ROC (Receiver Operating Characteristic) cutoff threshold and used maximum belief. If the predicted state is ‘1’ there are four cases where the actual value is ‘0’, so the probability, given x01 as observed, is less than 0.56; while there are thirty-nine cases where the actual value is ‘1’, so the probability is more than 0.56. If the predicted state is ‘0’ there are two cases where the actual value is ‘1’, so the probability, given x01 as observed, is less than 0.88, while there are fifty-five cases where the actual value is ‘0’, so the probability is more than 0.88. Moreover, the error rate is 6.00, the average Euclidean distance is 0.09641 and the average Kulbach-Leibler divergence is 0.17077.

In *Hugin Lite* I loaded as Data Source the file test.dat, that has comma as separator symbol. To learn the structure of the model for the specified data, I selected firstly NPC (Necessary Path Condition) and then Greedy search-and-score algorithms. For both cases, I decide leave the default values and then I initialized all Experience tables with user-defined value and set the Experience count to zero.

The Figure 3 is the network of NPC for test

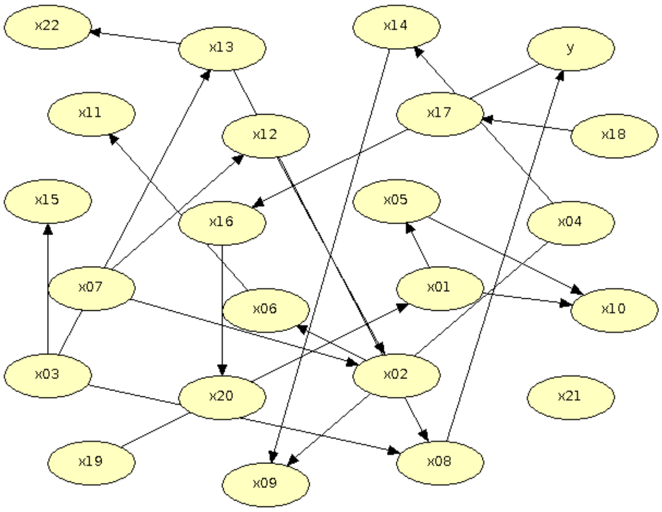

Figure 3: NPC for test

To generate the simulated cases, the number of cases was set to 100, because the *test.dat* contains 100 samples, and the percent missing value was set to the default value i.e. 5. After having simulated, in the Data Source, the minimum value of the conflict is -7.99885 of the 25th case and the maximum value is 2.35168 of the 28th case. If I consider, for instance, the Data Accuracy of the node x01, I have all cases that have a value for the probability of evidence (P(x01 = “state in case”)).

To prediction if the state is ‘0’ or ‘1’, I used the specific state with a belief greater or equal the ROC (Receiver Operating Characteristic) cutoff threshold and used maximum belief.

If the predicted state is ‘1’ there are two cases where the actual value is ‘0’, so the probability, given x01 as observed, is less than 0.81; while there are fifty-five cases where the actual value is ‘1’, so the probability is more than 0.81. If the predicted state is ‘0’ there are two cases where the actual value is ‘1’, so the probability, given x01 as observed, is less than 0.63, while I have thirty-nine cases where the actual value is ‘0’, so the probability is more than 0.63.

Moreover, the error rate is 5.00, the average Euclidean distance is 0.08700 and the average Kulbach- Leibler divergence is 0.16404.

The Figure 4 is the network of Greedysearch-and-score for test

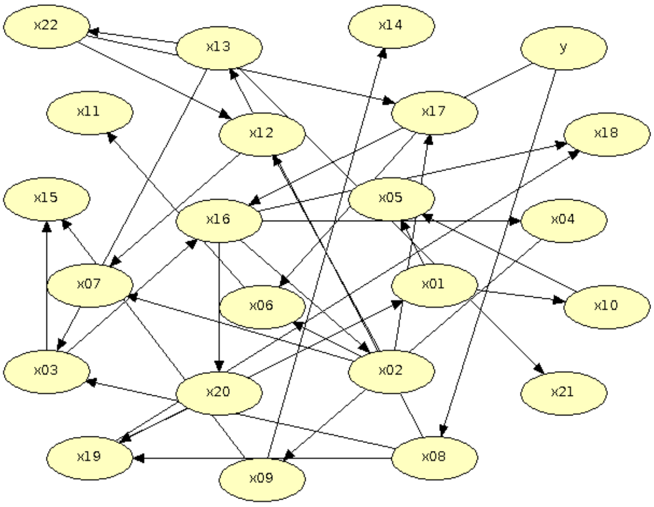

Figure 4: Greedy search-and-score algorithm algorithm for test

To generate the simulated cases, the number of cases was set to 100, because the *test.dat* contains 100 samples, and the percent missing value was set to the default value i.e. 5. After having simulated, in the Data Source, the minimum value of the conflict is -13.68413 of the 25th case and the maximum value is 1.37316 of the 28th case.

If I consider, for instance, the Data Accuracy of the node x01, I have all cases that have a value for the probability of evidence (P(x01 = “state in case”)).

To prediction if the state is ‘0’ or ‘1’, I used the specific state with a belief greater or equal the ROC (Receiver Operating Characteristic) cutoff threshold and used maximum belief. If the predicted state is ‘1’ there are two cases where the actual value is ‘0’, so the probability, given x01 as observed, is less than 0.89; while there are forty-three cases where the actual value is ‘1’, so the probability is more than 0.89. If the predicted state is ‘0’ there are three cases where the actual value is ‘1’, so the probability, given x01 as observed, is less than 0.57, while there are fifty-two cases where the actual value is ‘0’, so the probability is more than 0.57.

Moreover, the error rate is 6.00, the average Euclidean distance is 0.09641 and the average Kulbach- Leibler divergence is 0.17077.