Università degli Studi di Milano

Data Science and Economics

Probabilistic Modeling

Bayesian Network Modeling

Application on Breast Cancer case-study

Authors:

Emanuele Morales - 941935

A.Y. 2020/2021

Contents

1 Introduction

The aim of this project is to implement a Bayesian Network on a breast cancer dataset, trying to obtain a flexible tool that can be used to determine and visualize the interaction among the variables, paying particular attention on the confounding factors.

This project collocates in the Nutritional Epidemiology context, the branch of epidemiology that wants to investigate the relationship between the diets of patients and their diseases. It starts from the results of a previous study, in which considering all the nutrients intakes of the patients, four dietary patterns were identified, representing the quantities, proportions, and combination of different food and the frequency with which they are habitually consumed, and a value that represents the adherence of the patients to each of these patterns.

Besides these dietary patterns, the dataset contains also different variables related with the patient (personal data, medical data, etc.). These data, if correlated both with the cancer outcome, both with the diet followed by the patient, could distort the correlation, causing spurious regression between diet and cancer.

So far, the correlation between three dietary patterns and the cancer was found, and also a set of confounding factors, correlated both with cancer and both with the diets, by using logistic regression approach.

In figure 1 is reported the output of the logistic regression of the previous analysis, where it can be observed the significant correlation between dietary patterns and cancer (adjusted for confounding).

These results are confirmed by the literature [1], and we can consider them as the "expert knowledge". Starting from them, a Bayesian network is built, and the structure is refined by the combination of different structure learning techniques combined with bootstrap, in order to obtain a more precise network explaining the strength of the arcs between variables.

The validation of the model is performed by applying cross-validation in order to obtain an estimate of the goodness of fit and to measure the predictive accuracy (k-fold cross-validation).

Finally some queries are executed on the Bayesian Network in order to observe how the probability of getting cancer changes by changing the diets, stratifying by other factors.

[1] In this project, the main reference is the paper: Edefonti, V. et al. *Nutrient dietary patterns and the risk of breast and ovarian cancers.*

```
Call:
glm(formula = V2 ~ X1 + X2 + X3 + X4 + V12 + GIN4 + V11 + FUM1 +
    ALC1, family = binomial, data = df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8388  -1.1304  -0.8208   1.1726   1.8118

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.656214   0.162649  -4.035 5.47e-05 ***
X1          -0.026122   0.029550  -0.884 0.376705
X2           0.065373   0.029227   2.237 0.025302 *
X3           0.132125   0.029752   4.441 8.96e-06 ***
X4           0.090980   0.028981   3.139 0.001693 **
V12          0.072488   0.008208   8.831  < 2e-16 ***
GIN4        -0.034134   0.033738  -1.012 0.311661
V11         -0.064463   0.021488  -3.000 0.002700 **
FUM1         0.023659   0.042001   0.563 0.573230
ALC1         0.169106   0.050267   3.364 0.000768 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 1: Adjusted logistic regression output that shows the correlation between diets and cancer.

2 Data-set description

The data for the analysis are derived from two case-control studies conducted in different cancer centers in Italy.

The data set contains the information about 5157 patients and the following categorical variables are considered:

- V2: 1 = case; 0 = control.
- AGE: Age of the patient (<40; >=40 and <60; >=60).
- EDU - Education of the patient (elementary; middle/high; university).
- GIN4 - Menopausal Status (pre-menopause; peri-menopause; post-menopause).
- BMI - Body Mass Index (underweight; normal weight; overweight; obesity).
- CHILD - Number of pregnancies (0; 1; 2+).
- SMOKING - Smoking status (non smoker; smoker for <=25y; smoker for >25y).
- ALCOHOL - Qty of weekly alcohol intake (0 unit/week; 1-5 unit/week; 5-10 unit/week; >10 unit/week).

- FIS3 - Physical activity [15-19 y.o.] (heavy; medium; standing; sedentary).
- X2 - Diet with high sugar and low vegetable fats consume: (1 low adherence; 2 low/medium adherence; 3 medium/high adherence; 4 high adherence).
- X3 - Diet with high starch and low animal product consume: (1 low adherence; 2 low/medium adherence; 3 medium/high adherence; 4 high adherence).
- X4 - Diet with high sodium and low beta-carotene consume: (1 low adherence; 2 low/medium adherence; 3 medium/high adherence; 4 high adherence)

From the previous analysis, it is known that dietary pattern X2, X3, X4 are correlated with cancer (variable V2) and that the set of variables {GIN4, CHILD, SMOKING, ALCOHOL, FIS3} are confounding factors. These information will be given as input for the structure learning of the network in the form of a white list. Other relationship between variables will be found by using the combination of different structure learning procedures.

The limit of the previous analysis was related to the fact that it is "only" possible to observe that there is a positive correlation between diets and cancer output, but logistic regression does not provide a flexible model and the possibility of stratifying the analysis.

Instead, with Bayesian network, beside having a graphical representation of the relationship among the variables, it is possible to perform stratified analysis, by observing the trend of the probability of getting cancer by varying the type of diets and the adherence to them, by fixing other conditions in the network. For example, we could be interested in knowing how the probability of getting cancer changes, by varying dietary pattern X2, stratified by the number of children, or the smoking status.

This kind of analysis assumes a particular interest if related to the probability of getting cancer, by varying diet, conditioned to the confounding variables.

3 Methodology

3.1 Structure Learning

The first step consists in finding a larger set of relationships among variables that enlarges the set of relationships given in input to the Network in the form of a white list.

This task is performed by applying bootstrap technique on different structure learning algorithms, considering a threshold as a value that represents the boundary between strong arcs and weak arcs.

For example, the first learning procedure applied is PC. Firstly, data are re sampled by using bootstrap; for each bootstrap sample a network is learned. Then it is calculated in terms of frequencies how often each possible arc appears in the network. The arcs that have a frequency greater than a threshold (t = 0.5) are included in the network in output from the specific learning procedure.

	from	to	strength	direction
1	AGE	EDU	0.20	0.5000000
2	AGE	GIN4	1.00	0.5000000
3	AGE	ANTRO	0.02	0.7500000
4	AGE	CHILD	0.57	0.8684211
5	AGE	SMOKE	1.00	0.5250000
6	AGE	ALCOHOL	0.87	0.5000000

Figure 2: Example of strength of the arcs after the application of bootstrap on the Hill-Climbing learning procedure. Variables AGE and EDU have a low strength value (<0.5), instead AGE and CHILD an high strength value (>=0.50). The element inserted in the white list have the maximum value of strength (=1).

This method is applied on different learning procedures. It follows the complete list:

- Constraint-based structure algorithms: PC, Grow-Shrink.
- Score-based structure algorithms: Hill-climbing, Tabu greedy search.
- Hybrid structure algorithm: Max-Min Hill Climbing.

For each of the bootstrapped-structure learning procedure applied, it is obtained the adjacent matrix correspondent to the structure of the network found. In this case five adjacent matrices are obtained, one for each algorithm applied.

Then it is calculated the sum of these five matrices in order to obtain a union-matrix, where each row and column represent the variables and the cells represent the number of times an arc connected a specific variable (row) with another variable (column) in the learning procedure applied (for example: the cell corresponding to the intersection of row EDU and the column ANTR0 contains the value 2. This means that this arc appeared in two of the five learning procedures applied).

It follows the union matrix:

	AGE	EDU	GIN4	ANTR0	CHILD	SMOKE	ALCOHOL	FIS4	X2	X3	X4	V2
AGE	0	0	5	0	2	2	2	0	5	5	5	0
EDU	0	0	0	3	2	0	0	2	5	5	0	5
GIN4	5	2	0	0	0	0	0	0	5	5	5	5
ANTR0	0	0	0	0	2	0	0	0	5	0	5	0
CHILD	0	1	0	2	0	0	0	0	5	0	0	5
SMOKE	0	2	0	0	0	0	0	0	5	5	5	5
ALCOHOL	2	0	0	0	0	0	0	0	5	5	5	5
FIS4	0	0	0	0	0	0	0	0	5	5	5	0
X2	0	0	0	0	0	0	0	0	0	0	0	5
X3	0	0	0	0	0	0	0	0	0	0	0	5
X4	0	0	0	0	0	0	0	0	0	0	0	5
V2	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3: Union matrix: each intersection of the matrix represents the number of times an arc appears between the correspondent row-columns variables.

As it can be observed, there are arcs that appear one, two, three and five times.

By excluding the arcs that appear just one time, three types of network can be derived from this matrix. One with the arcs that appear at least two times, one with the arcs that appear at least three times and one with arcs that appear in all the five learning structure algorithms. In the following paragraph will be discussed how to find the optimal threshold to select the number of arcs to include in the model.

3.2 Validation of the model

In order to choose the best number of arcs to be considered in our model, two aspects must be kept in consideration: the graphical aspect (so the easy interpretation of the network and the relationship between variables) and the predictive power.

Some considerations about the first aspect can be given by observing the plot of the Bayesian network, considering different threshold levels to prune the arcs:

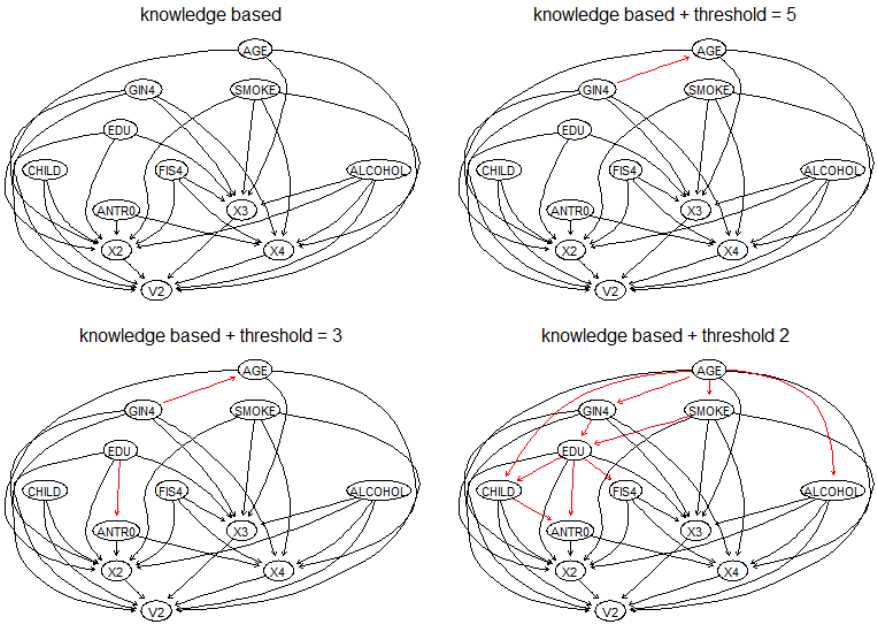

Figure 4: The first plot includes the arcs related to the "expert knowledge" network (arcs passed by the white list). The second plot includes also the arcs that appear in at least two learning procedure models, the third the arcs that appear in at least three learning procedure models and the last the arcs that appear in all the models.

All the four plots are not easy to interpret due to the large number of relationships between the variables. By decreasing the threshold of admitted arcs, the networks become more and more difficult to be interpreted.

About the goodness of fit of the models, it is possible to measure the predictive accuracy of the models by applying k-fold cross-validation as an indicator of the goodness of fit of the model, observing how the loss-function changes by varying the number of arcs included in the network. This technique consists in partitioning data in k subsets (in this case 10). Each subset is used in turn to validate the model fitted on the remaining k-1 subsets.

We can compare the resulting sets of loss values by plotting them as box plots:

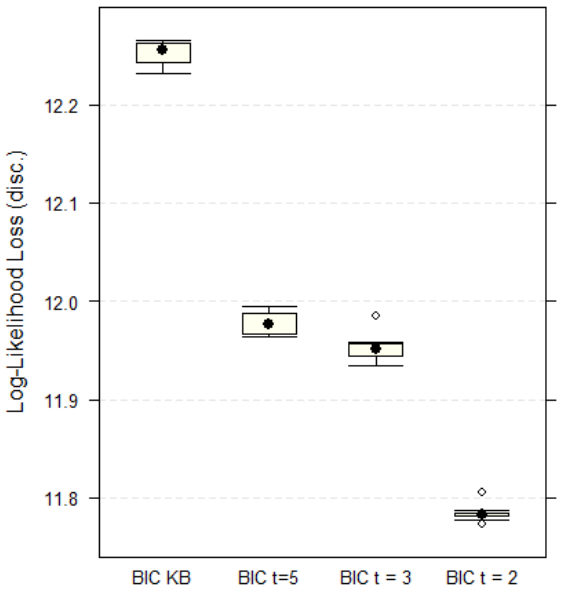

Figure 5: Box plots representing the change in the BIC loss function by varying the number of arcs included in the network

As expected, by increasing the number of arcs, the accuracy of the model improves. In order to observe how the probability of getting cancer varies conditioned to other variables will be chosen the model with more arcs (threshold = 2), even if it the network is more difficult to read.

4 Results

Considering the model in output from the previous paragraph, it is possible to study the conditioned probability distribution of the variable V2 by varying the adherence of patients to a specific diet, stratifying the analysis by the confounding factors and other variables. This task can be performed by performing queries on the Bayesian Network. In Figure 6, the plots of the conditioned and stratified probability can be observed.

From the charts we can infer that:

- It is possible to consider just the probability of getting cancer by varying diets, from the black lines in the plots. Particularly, the more significant trend identified by this model is related to the diet X3, in which it can be observed an upward trend. X3 is a diet characterized by the high consume of starch. We can find a confirmation of this result also in literature. For example, in the paper *"Nutrient dietary patterns and the risk of breast and ovarian cancers" - Edefonti et al.*, it is shown that *"starch-rich dietary pattern are directly associated with breast cancer"*.
- the probability of getting cancer increasing the adherence to diet X2 seems to be independent from the level of alcohol consumption (until the third quartile, the lines have similar trends)
- the probability of getting cancer increasing the adherence to diet X2 improves for obese patient and reduces for under-weighted patients.
- the probability of getting cancer increasing the adherence to diet X3 has a common trend independently from the level of education, but it can be observed that it is much lower in the case of an high level of instruction.
- the probability of getting cancer improving the adherence to diet X3 seems to slightly increase for patients in post-menopause (upward trend in the yellow line)
- the probability of getting cancer seem to have an upward trend in the case of increasing adherence to diet X4 for patients smoking for more than 25 years.
- there is not evidence of a significant change in the probability by changing the adherence to the diet X4 and changing the physical activity of the patients in age 19-25.

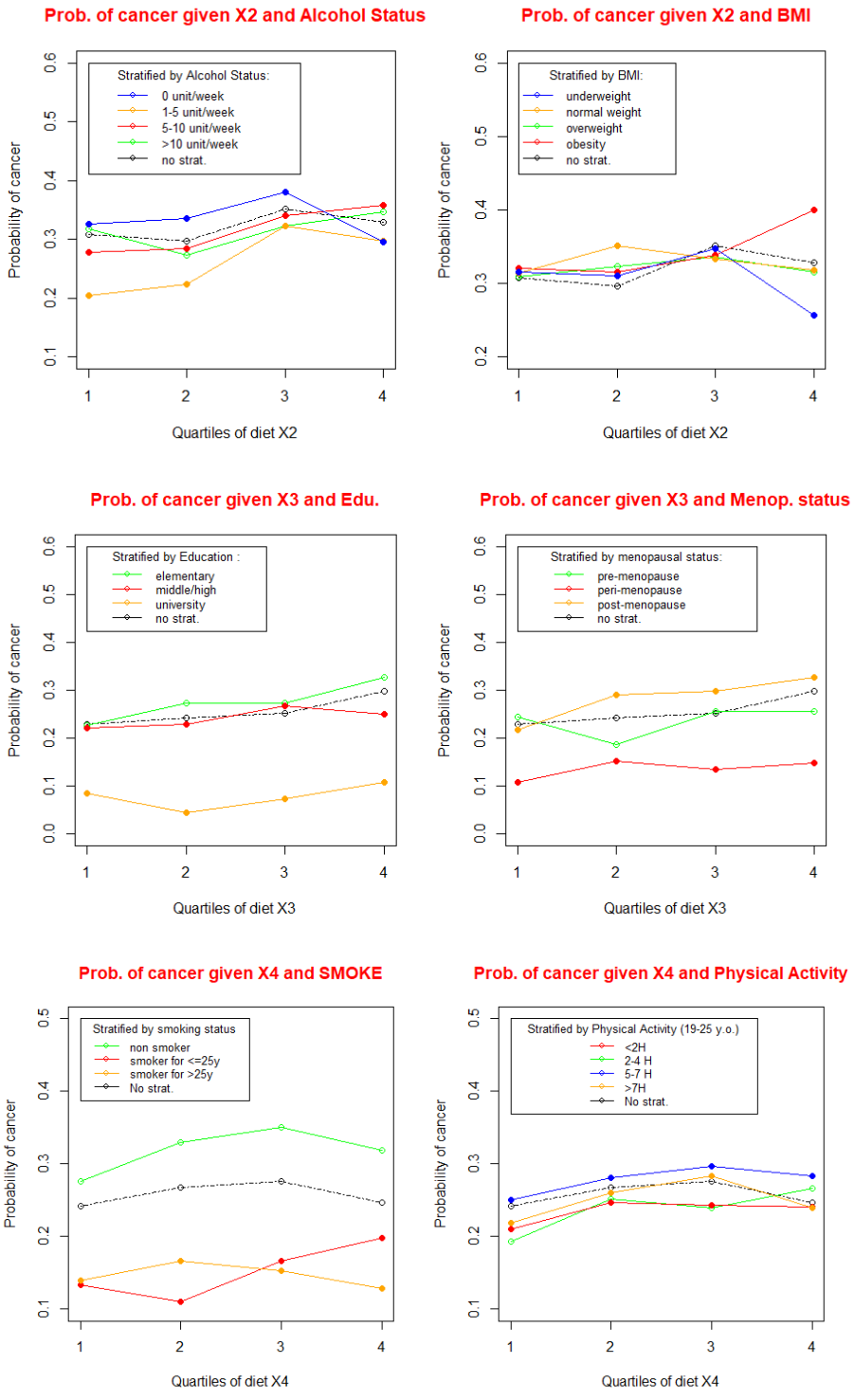

Figure 6: Example of queries executed on the Network.