Bayesian Network applied to the Heart failure data-set

Lirida Papallazi
lirida.papallaz i@studenti.unimi.it

November 4, 2020

Contents

List of Figures

Chapter 1

Introduction

The aim of this project is to analyze the application and the results of a Bayesian network to the Heart Failure dataset available on kaggle.[1]

R is the programming language that will be extensively used and in particular the bnlearn package, acronym of "Bayesian network structure learning", which was created for learning the graphical structure of Bayesian networks, estimate their parameters and perform inferences.

Chapter 2

Statistical and Mathematical background

In order to facilitate the understanding of the analysis conducted and the results obtained, the reader will be provided with a brief mathematical and statistical explanation of the concepts underlying the concept of Bayesian networks.

A Bayesian network in particular, is a probabilistic graphical model that represents a set of variables $X = X_1, .., X_n$ and their conditional dependencies via a directed acyclic graph (DAG) $G = (V, A)$, where each node $v_i \in V$ corresponds to a random variable X_i.

2.1 Graph Theory

Graph theory is a branch of mathematics that studies graphs, mathematical structures used to represent relations between objects.

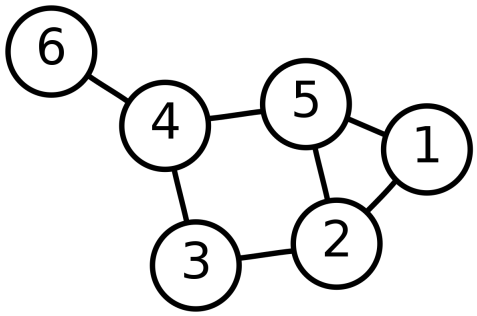

Figure 2.1: Example of undirected graph

2.1.1 What is a Graph

A graph $G = (V, A)$ is a combination of vertices V which are connected by arcs or edges A. In particular, when the arc is defined as an ordered pair of nodes the graph will be said to be directed, otherwise it will be undirected. As previously stated, a Bayesian network represents the relations between its variables via a directed acyclic graph, that is: each arc is directed from one vertex to another and it is not possible

that starting from any vertex v, and following the directed sequence of edges, to loop back to v again.

An alternative explanation would require the introduction of a few important concepts to define what a cycle is.

Walk

A walk can be a finite or infinite sequence of edges that connect a sequence of vertices.

Trail

A trail is a walk in which all arcs are distinct, namely edges are not repeated.

Path

A path is a trail in which all vertices and as a consequence all arcs, are distinct, namely there is no repeated vertex.

If two nodes $u = v$ then the u, v-walk and the u, v-trail are closed. A closed trail is a circuit and a circuit with no repeated vertex is called a cycle.

2.1.2 The structure of a graph

Real world graphs are typically sparse or dense, but never saturated which means that it is very difficult to observe a phenomenon for which every node is connected to every other node. This aspect can also be observed when representing a finite graph in via Adjacency Matrix. An adjacency matrix is a square matrix whose elements indicate whether pairs of vertices are connected by and edge or not in the graph. In real life these matrices will be mostly sparse and when the corresponding graph is acyclic (as in the case of Bayesian networks), namely there are no cycles nor loops, the diagonal will be composed only by zeros.

2.2 Bayesian Statistics

As the name says, Bayesian Networks have their foundations on Bayesian Statistics, which in turn involves on the concept of conditional probabilities.

2.2.1 Conditional Probability

The conditional probability of an event A is the the probability of A given event B, and it can be calculated using the Bayes rule.

Let A and B be two sets of events; the formula for computing A when B has already happened will be:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \tag{2.1}$$

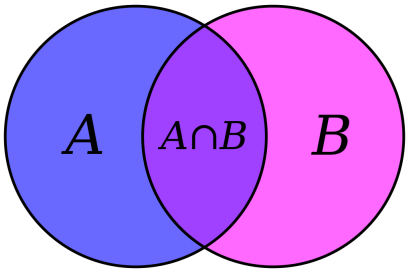

Figure 2.2: Venn Diagram of Intersection

while the formula for computing the probability of B when A has already happened will be:

$$P(B|A) = \frac{P(B \cap A)}{P(A)} \tag{2.2}$$

It is possible to write the first equation as a function of the second:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{2.3}$$

This is the conditional probability of event A when event B has already happened and it is also the simplest expression of the Bayes' rule for two events.

2.2.2 Bayes Theorem

Given a set of alternatives $A_1, ..., A_n$ partitioning the event space Ω such that $A_i \in A_j = \emptyset \ \forall i \neq j$ and $\sum_{i=1}^{n} \cup A_i = \Omega$ the conditional probability will be given by the following equation:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_i)P(A_J)} \tag{2.4}$$

Where:

- $P(A)$ is the prior probability or marginal probability of A, and prior means that it does not take into account any information about B;
- $P(A|B)$ is the conditional probability of A, given B and is the so called posterior probability and changes when empirical evidence comes in;
- $P(B|A)$ is the conditional probability of B, given A;
- $P(B)$ is the prior probability of B and it acts as a normalizing constant to ensure the value of $P(A_j|B)$ is a valid probability, namely a number between 0 and 1.[2]

Due to the complexity and the length of the topic we refer the reader to other sources of information more focused on Bayesian statistics and now the focus will switch to Bayesian Networks (BNs).

2.3 Bayesian Networks

The edges of the Directed Acyclic Graphs implemented by Bayesian Networks are used to represent direct dependencies among the variables, which makes BNs particularly ideal for predicting the likelihood that any of the several possible known causes of an occurred event, can be considered contributing factors. In particular, an arrow that starts from node v_i and goes into node v_j means that v_i influences v_j, thus the former is referred to as the "parent" of the second, and the latter as its child.

2.3.1 Basic concepts

Bayesian networks are defined by a pair $BN = \langle G, \Theta \rangle$:

- G is the network structure, namely the DAG and
- Θ denotes the set of parameters of the network, i.e. the local probability distributions associated with each variable X_i.

Factorization definition

The main role of the network structure is to express the conditional independence relationship among the variables specifying the factorization of the global distribution and it is able to do this if its joint probability density function can be written as a product of the individual density functions, conditional on their parent variables.[3]

$$P(X) = \prod_{i=1}^{n} P(X_i|\Pi_{X_i}) \tag{2.5}$$

where Π_{X_i} is the set of the parents of X_i.

I-map

This type of representation allows to define a correspondence, called Indepencency Map (I-map), between the absence of an arc between two nodes and probabilistic independence.[4]

Local Markov property

BN is a Bayesian network with respect to the graph G if it satisfies the local Markov property, namely each variable is conditionally independent of its non-descendants given its parent variables.[5]

Markov blanket

The Markov blanket of a node v_i is the set of nodes consisting of its parents, its children and any other parents of its children. As a consequence of this BN is a Bayesian network with respect to G if every node is conditionally independent of all other nodes in the network, given its Markov blanket.[6]

2.3.2 Purposes of Bayesian Networks

Bayesian networks are mainly used for three purposes.

Probabilistic Inference

Bayesian Networks can be used to answer probabilistic queries on the relationships existing between the variables by computing the posterior distribution of variables given evidence.

Structure Learning

Typically the network is specified by an expert but when its construction is too complicated the structure and the parameters must be learned from the data. One of the methods developed for structure learning requires a scoring function, typically posterior probability of the structure given the training data, such as BIC; but an alternative could be implementing Markov Chain Monte Carlo (MCMC) algorithms.

Parameter Learning

Parameter learning is the process of using data to learn the distributions of the variables composing a Bayesian network. There are many approaches to learn the conditional distributions but typically either maximum likelihood or expectation-maximization algorithm.

Chapter 3

Data Analysis

Extending the concepts explained of the previous chapter to the data-set analysed, the aim was that of first of all building the Bayesian Network structure and therefore the relations between the variables, and afterward estimate the impact that each variable and combination of variables had on the ultimate target, namely the patient's heart failure.

3.1 Data-set and variables

The data-set is composed by 13 variables and 299 observations:

- age = numerical variable;
- anaemia, boolean variable = equal to 1 if patient is anaemic, 0 otherwise;
- creatine phosphokinase (cpk), numerical variable = is an enzyme that catalyses the conversion of creatine and uses adenosine triphosphate (ATP) to create phosphocreatine (PCr) and adenosine diphosphate (ADP). This CK enzyme reaction is reversible.[7]
- diabetes, boolean variable = equal to 1 if patient is anaemic, 0 otherwise;
- ejection fraction, percentual variable = Ejection fraction is a measurement of the portion of total blood ejected from a chamber with each contraction and it is commonly measured via echocardiography.
- high blood pressure (hypertension), boolean variable = equal to 1 if patient is anaemic, 0 otherwise;
- platelets count in the blood (kiloplatelets/mL);
- serum creatinine, numerical variable = creatinine is a breakdown product of creatine phosphate from muscle and protein metabolism and it depends on muscle mass. Serum creatinine is an indicator of kidney health.
- serum sodium, numerical variable = it measures the level of sodium in blood.
- sex, boolean variable = equal to 1 if patient is male, 0 if female;
- smoking, boolean variable = equal to 1 if patient is a smoker, 0 otherwise;

- follow up period, numerical variable = monitoring period;
- death event, boolean variable = equal to 1 if patient died during the follow up period, 0 otherwise.

3.2 Pre-processing

As it is possible to notice the data-set is "mixed", namely it contains both numerical and categorical variables, which would result in a so called Hybrid network.

Given the nature of the data-set it is possible to modify all the continuous variable by simply setting some thresholds for each variable. In particular:

- age:
  - below 65 = adult;
  - above 65 = elder.
- creatine kinase (cpk):
  - below 120 = normal;
  - above 120 = high.[8]
- ejection fraction:
  - below 45% = heart failure;
  - between 45 and 70% = normal;
  - over 75% = high;[9]
- platelets:
  - below 150000 = thrombocytopenia;
  - between = normal;
  - above 4500000 = thrombocytosis.[10]
- follow-up (time):
  - below 95 days = short;
  - between 95 and 190 = medium;
  - above 190 = long time
- serum creatinine:
  - below 0.75 = muscle disease;
  - between 0.75 and 1.2 = normal;
  - above 1.2 = kidney failure.[11]
- serum sodium:
  - below 135 = Hyponatremia;

– between 135 and 145 = normal;
– above 145 = Hypernatremia (involves dehydration).[12]

And lastly the boolean variables are converted from 0-1 values to their respective meaning (no and yes, or for the sex variable, female and male), so that the final data-set will look as follows.

	anaemia	diabetes	hypertension	sex	smoking	death	age	cpk	ejection_fraction	platelets	serum_creatinine	serum_sodium	time
1	no	no	yes	male	no	yes	elder	high	heart failure	normal	high	Hyponatremia	short
2	no	no	no	male	no	yes	adult	high	heart failure	normal	normal	normal	short
3	no	no	no	male	yes	yes	adult	high	heart failure	normal	high	Hyponatremia	short
4	yes	no	no	male	no	yes	adult	normal	heart failure	normal	high	normal	short
5	yes	yes	no	female	no	yes	adult	high	heart failure	normal	high	Hyponatremia	short
6	yes	no	yes	male	yes	yes	elder	normal	heart failure	normal	high	Hyponatremia	short

Figure 3.1: First six rows of the final dataset

On a side note: some of the thresholds were adapted to the data, for example the age variable the age variable should have been divided into three intervals:

- below 30 = young;
- between 30 and 60 = adult;
- above 60 = elder/old.

However the youngest patient in the whole data-set is 40, making such division impractical.

3.3 Dependencies

The choice for the joint probability distribution in this case is that of a multinomial distribution therefore, the next step involves the definition of the dependencies between the variables represented through arcs: variables that are not linked by an arc are conditionally independent.

It is also important to say that the analysis was performed without considering two variables: the serum creatinine and the serum sodium due to lack of information on the dependencies with the other variables.

variable	age	sex	smoking	diabetes	anaemia	hypertension	time	death	cpk	ejection fraction	platelets
age	0	0	0	0	0	0	0	0	0	0	0
sex	0	0	0	0	0	0	0	0	0	0	0
smoking	0	0	0	0	0	0	0	0	0	0	0
diabetes	1	0	1	0	0	0	0	0	0	0	0
anaemia	0	1	1	0	0	0	0	0	0	0	0
hypertension	1	1	1	1	0	0	0	0	1	1	0
time	1	0	1	1	1	1	0	0	1	1	1
death	1	0	1	1	0	1	1	0	0	1	1
cpk	0	0	1	1	0	0	0	0	0	0	1
ejection fraction	1	1	1	1	1	0	0	0	0	0	0
platelets	1	1	1	0	0	0	0	0	1	0	0
if dependent = 1											
if independent = 0											

Figure 3.2: Variables' dependencies table

As it is possible to notice from Figure 3.2, the table basically represents the adjacency matrix of the graph.

3.4 Structure Learning

In order to define the Bayesian Network different techniques were implemented.

3.4.1 Pre-defined structure

This method required the definition of the variables and of the dependencies between each of them. The authors decided to start off considering only 6 variables: age, sex, smoking, anaemia, hypertension and death and then adding one by one the remaining variables.

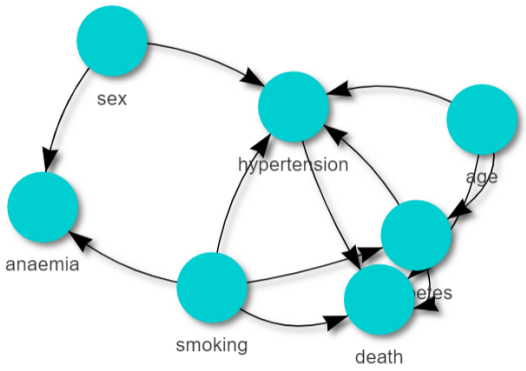

Figure 3.3: Bayesian network with only 6 variables: age, sex, smoking, anaemia, hypertension and death

However this method this process results in a partially directed graph when adding the platelet count therefore and it is therefore necessary to resort to an alternative.

When the network structure is not pre-specified the alternative would be resorting to an algorithm. Several algorithms have been presented in literature and three approaches can be identified:

- constraint-based algorithms;
- score-based algorithms;
- hybrid algorithms.

We will start off with the first type of algorithms, and then an explanation for the second and the third types will follow.

3.4.2 Constraint-based Algorithms

Constraint-based algorithms are based on Pearl's Inductive Causation (IC) algorithm which provides a framework for learning the DAG structure of BNs using conditional independence tests.[13] However the original algorithm cannot be applied to any real-world problem due to the exponential number of possible conditional independence relationships, which lead to the development of alternative algorithm:

- PC;
- Grow-Shrink (GS): based on the Grow-Shrink Markov blanket algorithm;
- Incremental Association (IAMB);
- Fast Incremental Association (Fast-IAMB;
- Interleaved Incremental Association (Inter-IAMB).

With the exception of PC, all these algorithms first learn the Markov blanket of each node in order to reduce the number of conditional independence test and consequently the overall complexity. The main tests used for discrete data (such in this case) are: mutual information (G^2), Pearson's χ^2 and mutual information (shrinkage) and the null hypothesis of independence can be tested using: the asymptotic $\chi^2_{(R1)(C1)L}$, the Monte Carlo permutation approach, the Sequential Monte Carlo permutation or the semiparametric χ^2 distribution.

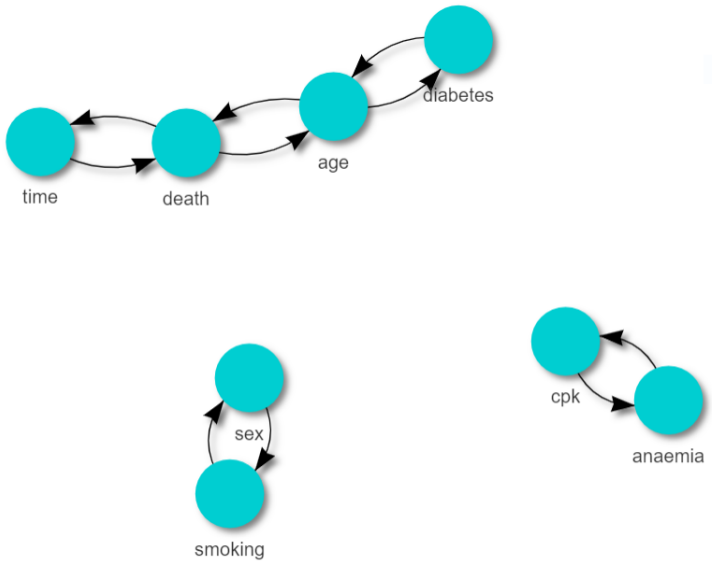

Figure 3.4: Grow-Shrink algorithm graphic result

```
Bayesian network learned via Constraint-based methods

model:
  [undirected graph]
nodes:                                  11
arcs:                                   5
  undirected arcs:                      5
  directed arcs:                        0
average markov blanket size:            0.91
average neighbourhood size:             0.91
average branching factor:               0.00

learning algorithm:                     Grow-Shrink
conditional independence test:          Mutual Information (disc.)
alpha threshold:                        0.05
tests used in the learning procedure:   206
```

Figure 3.5: Grow-Shrink algorithm result

As it is possible to see the Grow-Shrink algorithm seems to confirm the lack of dependence between anaemia and death, but it seems to be connected to the creatine phosphokinase level. This problem can be solved by using the whitelist and blacklist arguments and obtain the correct DAG.

3.4.3 Score-based Algorithms

[13] Score-based learning algorithms implement heuristic optimisation techniques to solve the problem of learning the structure of the BN by assigning a network score that maximizes the fit. The most famous are:

- greedy-search algorithms (such as Hill-climbing or Tabu search);
- genetic algorithms;
- simulated annealing.

In particular the application of the Hill climbing algorithms, with the specification of the sole whitelist (namely the positive dependencies between variables), of the sole blacklist (namely the independencies between variables) and lastly of both the lists resulted in the production following graphs with a score of, respectively: -2827.026, -2191.557 and -2840.302.

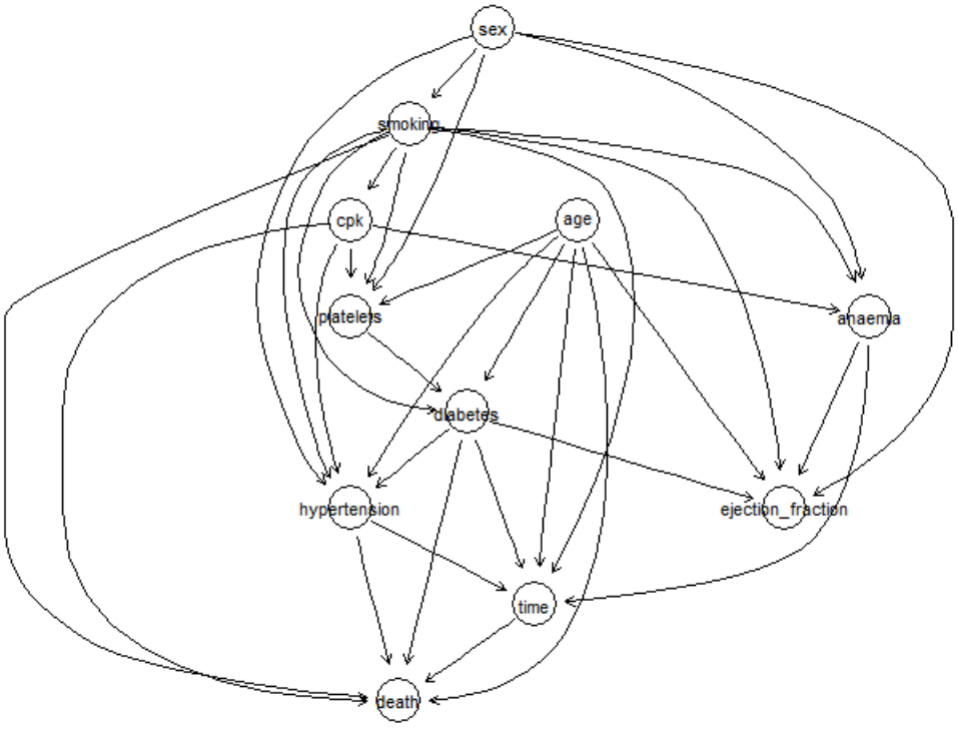

Figure 3.6: Hill-climbing algorithm with whitelist specification (AIC)

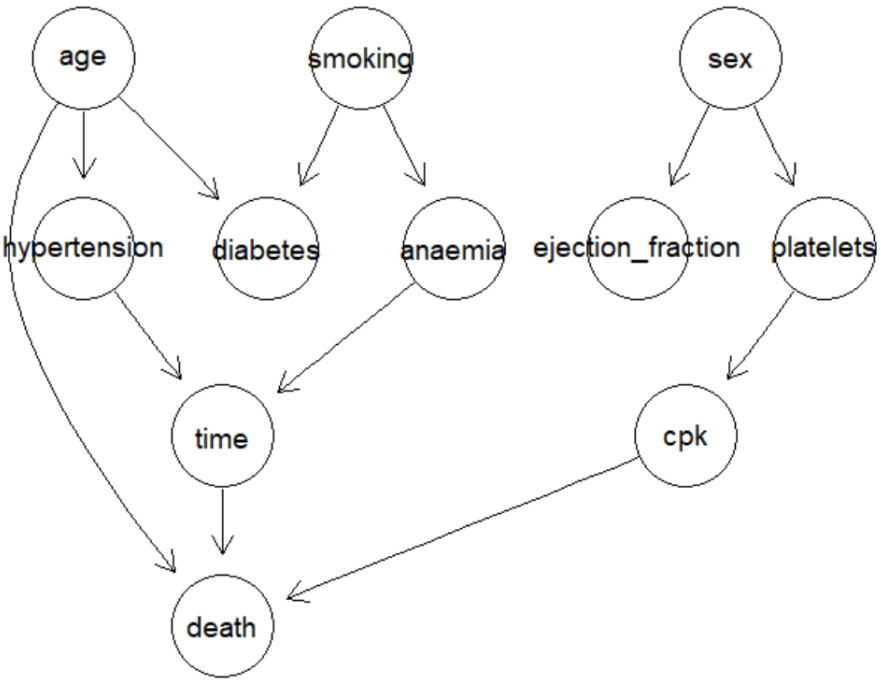

Figure 3.7: Hill-climbing algorithm with blacklist specification (AIC)

3.5 Parameter Learning

Once the structure of the BN has been learned it is possible to proceed with the estimation of the parameters of the global distribution. It is quite obvious that the structure of the network is assumed to be known.

The main approaches used are:

- maximum likelihood estimation;
- Bayesian estimation.

In particular, for the result of the structure built via Hill-climbing algorithm in the case of specification of the sole whitelist and then the sole blacklist the parameters estimated applying mle and the Bayesian estimation are the following.

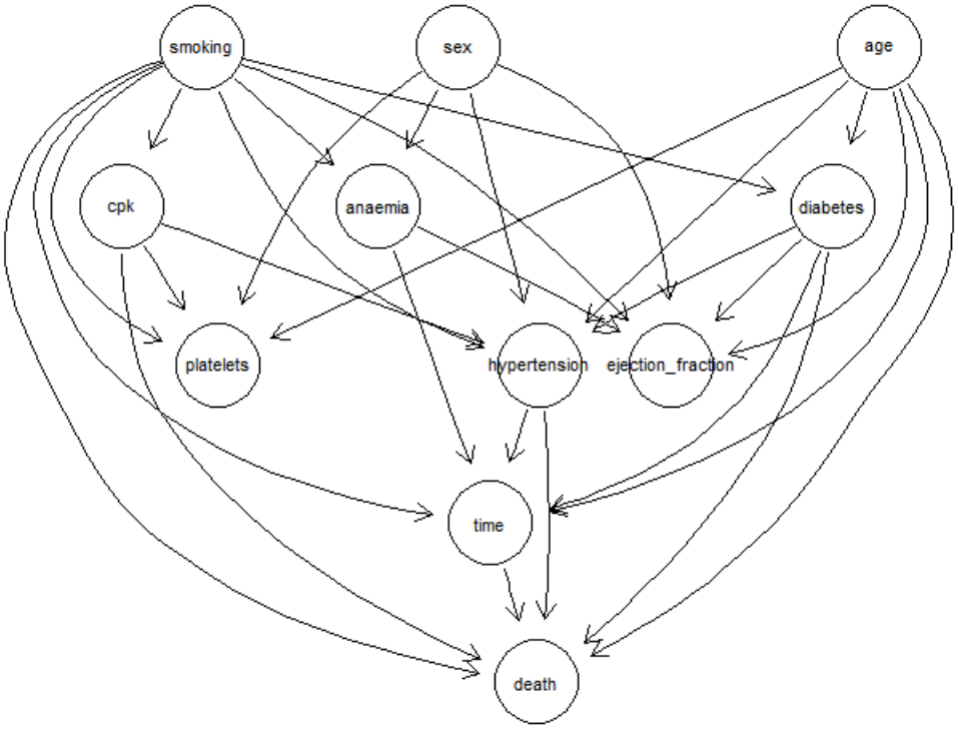

Figure 3.8: Hill-climbing algorithm with both whitelist and blacklist specification (AIC)

```
, , diabetes = yes, hypertension = yes, smoking = yes, age = elder

        anaemia
time             no        yes
  short  0.75000000 1.00000000
  medium 0.25000000 0.00000000
  long   0.00000000 0.00000000
```

Figure 3.9: Hill-climbing structure with whitelist specification, mle

```
, , diabetes = yes, hypertension = yes, smoking = yes, age = elder

        anaemia
time             no         yes
  short  0.746770026 0.979797980
  medium 0.250645995 0.010101010
  long   0.002583979 0.010101010
```

Figure 3.10: Hill-climbing structure with whitelist specification, Bayesian estimation

[MLE 1]

```
Conditional probability table:

, , hypertension = no

        anaemia
time            no       yes
  short  0.3628319 0.3580247
  medium 0.2300885 0.3827160
  long   0.4070796 0.2592593

, , hypertension = yes

        anaemia
time            no       yes
  short  0.4035088 0.6458333
  medium 0.3684211 0.1875000
  long   0.2280702 0.1666667
```

[Bayes 2]

```
Conditional probability table:

, , hypertension = no

        anaemia
time            no       yes
  short  0.3627667 0.3579487
  medium 0.2303164 0.3825641
  long   0.4069169 0.2594872

, , hypertension = yes

        anaemia
time            no       yes
  short  0.4032023 0.6442142
  medium 0.3682678 0.1882556
  long   0.2285298 0.1675302
```

Figure 3.11: Comparison between parameter learning via mle and Bayesian estimation

Chapter 4

Conclusion

The topic of Bayesian networks is extremely broad and in this project we have limited ourselves to scratching the surface although the practical applications and usefulness in various fields are still evident even from this small analysis.

In general fitting the network and querying the model is only the first part of Bayesian networks. In our analysis for example, by knowing that the chances of having a heart failure are influenced by combination of causes such as the ejection fraction and the platelet count, especially when the patient is diabetic then doctors and general practitioners can increase the follow-up time in order to avoid the death of the patient.

Moreover the conditional probabilities estimated after defining the structure can be used to create what-if scenarios or in optimization problems.

Then a way to extend the bootstrap in regression models can be the following:

The original sample $s = \left[\mathbf{z}_1^{'}, \mathbf{z}_2^{'}, \ldots, \mathbf{z}_n^{'}\right]$ can be resampled with replacement B times, obtaining bootstrap samples $s_b^* = \left[\mathbf{z}_{b1}^{'}, \mathbf{z}_{b2}^{'}, \ldots, \mathbf{z}_{bn}^{*'}\right], b = 1, \ldots, B$.

For each of this bootstrap samples, perform an OLS estimation and obtain regression coefficient estimates $\boldsymbol{\beta}^b = [\beta_{b0}^{;}\beta_{b1}^{;} \ldots, \beta_{bp}^{]}{}^{'}$.

Use these B estimates to compute biases, standard errors, confidence intervals, etc. of regression coefficient estimates.

Bibliography

[1] Larxel, Heart Failure Prediction, https://www.kaggle.com/andrewmvd/heart-failure-clinical-data

[2] Review of basic statistical inference and probability: random variables, G. Manzi, Advanced Multivariate Statistics course (2019/2020)

[3] M. Scutari, Modelling Survey Data with Bayesian Networks (18/05/2015)

[4] R. Nagarajan, M. Scutari, S. Lèbre, Bayesian Networks in R with Applications in System Biology, Springer, 2013

[5] R. Nagarajan, M. Scutari, S. Lèbre, Bayesian Networks in R with Applications in System Biology, Springer, 2013

[6] Markov blanket, Wikipedia, https://en.wikipedia.org/wiki/Bayesian_network Restrictions_on_priors

[7] Creatine Kinase, Wikipedia,https://it.wikipedia.org/wiki/Creatinchinasi

[8] M.A. Chen and D. Zieve, Creatine phosphokinase test, https://www.mountsinai.org/health-library/tests/creatine-phosphokinase-test: :text=Normal%20Results,per%20liter%20(mcg%2FL)

[9] American Heart Association, Ejection Fraction Heart Failure Measurement, https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement: :text=What %20is%20%E2%80%9Cejection%20fraction%E2%80%9D%3F,pushed %20out%20with%20each%20heartbeat.)

[10] A. Frosi, Piastrinopenie (trombocitopenia), https://www.idoctors.it/patologia-piastrinopenie–trombocitopenia–26747: :text=La%20normale%20sopravvivenza %20delle%20piastrine,definisce%20trombocitosi%20(o%20piastrinosi).

[11] C. Haldeman-Englert, L. Cunningham, R. Turley, Creatinine (Blood), https://www.urmc.rochester.edu/encyclopedia/content.aspx?ContentTypeID =167ContentID=creatinine$_serum$

[12] J.L. Lewis, Hypernatremia (High Level of Sodium in the Blood), https://www.msdmanuals.com/home/hormonal-and-metabolic-disorders/electrolyte-balance/hypernatremia-high-level-of-sodium-in-the-blood: :text=In%20hypernatremia%2C%20the%20level%20of,%2C%20kidney %20dysfunction%2C%20and%20diuretics.

[13] M.Scutari, J.B. Denis, Bayesian Networks With Examples in R