Bayesian network model for diagnosis of psychiatric diseases

Fundamentals of Artificial Intelligence and Knowledge Representation

Module 3

Alex Costanzino

alex.costanzino@studio.unibo.it

GitHub

Xiaowei Wen

xiaowei.wen@studio.unibo.it

GitHub

Academic year 2020-2021

Abstract

Due to the numerous possible causes involved, it is not easy for general physicians to identify the precise reason of the psychiatric diseases, hence, to decide the correct treatment [1].

Bayesian networks are recognized as efficient graphical models with significant capabilities for investigating biomedical data, either to obtain relationships between variables, either for medical predictions.

The development of this Bayesian network led to the identification of most significant factors that affect some important diseases and their correlations.

1 Model

Bayesian networks are a probabilistic graphical models that measure the conditional dependence structure of a set of random variables based on the Bayes theorem:

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

Bayesian networks are graphical models that contain information about causal probability relationships between variables and are often used to aid in decision making [4].
The causal probability relationships in a Bayesian network can be suggested by experts or updated using the Bayes theorem and new data being collected.
The inter-variable dependence structure is represented by **nodes** (variables) and **directed arcs** (conditional relationships) in the form of a *directed acyclic graph* (**DAG**):

- Each node corresponds to a random variable, which may be discrete or continuous;
- A set of directed links connects pairs of nodes. If there is a link from node X to node Y, X is said to be *parent* of Y. The graph has no directed cycles
- Each node X_i has a conditional probability distribution $P(X_i|\text{Parents}(X_i))$ that quantifies the effect of the parents on the node.

In our model, only a sub-net of the original network was considered. The involved variables are:

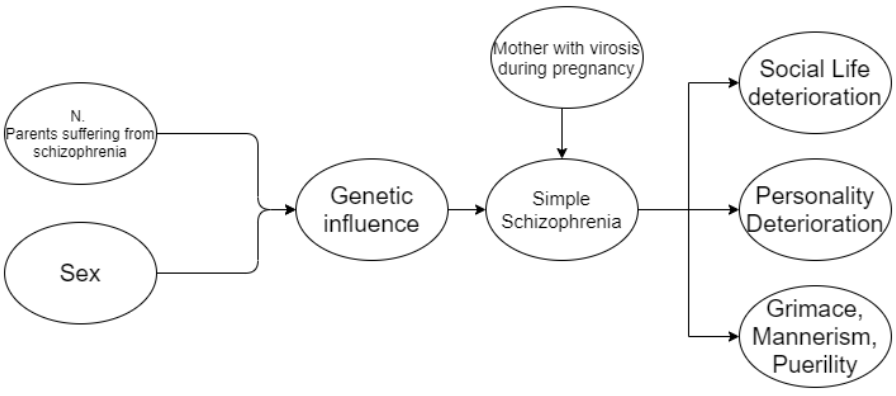
- **Simple schizophrenia**: the main variable of interest. It is a discrete variable: is the disease present or not?
- **Sex**: sex of the patient. It is known that women are more liable to have or to develop a psychiatric disease than men. It is a discrete variable: male or female?
- **Number of parents suffering from schizophrenia**: family history is one of the main risk factor for schizophrenia. It is a discrete variable: none, one or both parents have the disease as well?
- **Genetic influence schizophrenia**: it gathers all the others genetic factors and it is influenced by sex and family history. It is a discrete variable;

- **Mother with virosis during the pregnancy**: there is a chance of schizophrenia for those whose mothers had some kind of virosis during pregnancy. It is a discrete variable: did the mother had a virosis during pregnancy?

- **Personality deterioration**: it can be one of the consequences of schizophrenia. It is a progressive decline in an individual sense of personal identity, self-worth, motivational forces, and emotional life to the point at which the patient appears to be a changed person or even a non-person. It is a discrete variable: is the personality deteriorated?

- **Social life deterioration**: another main consequence of schizophrenia. It is a progressive decline of the willing of interaction of the patient with the environment and other people. It is a discrete variable: is the social life deteriorated?

- **Mannerism**: habits are repetitive movements performed by normal people when they are bored, anxious, self-conscious, or tired. Some people carry out normal actions in a peculiar fashion, usually in an attempt to call attention to themselves. These are referred to as mannerisms. Mannerisms are particularly common in patients with schizophrenia. Compulsions are repetitive, often ritualistic actions carried out in response to an obsession, to reduce anxiety, or to avoid a future dreaded outcome [3]. It is a discrete variable: does the patient exhibit mannerism?



2 Conditional probability distribution tables

Conditional probability distributions (**CPDs**) are a form of tabular representation suitable for discrete variables. Each row in a CPD contains the conditional probability of each node value for a conditioning case. A conditioning case is just a possible combination of values for the parent nodes.

The following tables are the CPDs for the nodes **Sex** and **Genetic Influence**:

Sex (Male)	0.5
Sex (Female)	0.5

Table 1: CPD of **Sex**

Sex	Parents suffering from schizophrenia	Genetic influence (True)	Genetic influence (False)
Male	Zero	0.01	0.99
Male	One	0.05	0.95
Male	Both	0.30	0.70
Female	Zero	0.01	0.99
Female	One	0.06	0.46
Female	Both	0.46	0.54

Table 2: CPD of **Genetic influence**

For conciseness sake, other CPDs definition can be found in the notebook.

3 Independence

A and B are independent, denoted $P \models (A \perp B)$, iff either:

- $P(A|B) = P(A)$;
- $P(B|A) = P(B)$;
- $P(A, B) = P(A)P(B)$.

X and Y are conditionally independent, given Z, denoted $P \models (X \perp Y, Z)$, iff:

$$P(X, Y|Z) = P(X|Z)P(Y, Z)$$

In our network, there can be made 642 valid independence assertions, with respect to the all possible given evidence. The main local independencies are:

- $P \models$ (Sex $\perp$ Mother with virosis during pregnancy, Parents suffering from schizophrenia);
- $P \models$ (Parents suffering from schizophrenia $\perp$ Mother with virosis during pregnancy, Sex);
- $P \models$ (Mother with virosis during pregnancy $\perp$ Parents suffering from schizophrenia, Sex, Genetic influence);
- $P \models$ (Genetic influence $\perp$ Mother with virosis during pregnancy | Parents suffering from schizophrenia, Sex);
- $P \models$ (Simple schizophrenia $\perp$ Parents suffering from schizophrenia, Sex | Mother with virosis during pregnancy, Genetic influence);
- $P \models$ (Social life deterioration $\perp$ Mother with virosis during pregnancy, Sex, Parents suffering from schizophrenia, Mannerism, Personality deterioration, Genetic influence | Simple schizophrenia);
- $P \models$ (Mannerism $\perp$ Mother with virosis during pregnancy, Social life deterioration, Sex, Parents suffering from schizophrenia, Personality deterioration, Genetic influence | Simple schizophrenia);

- $P \models$ (Personality deterioration $\perp$ Mother with virosis during pregnancy, Social life deterioration, Parents suffering from schizophrenia, Mannerism, Sex, Genetic influence | Simple schizophrenia).

3.1 Markov Blanket

The **Markov blanket** of a given node (in *red*) consists in the set (in *yellow*) of:

- Its parents;
- Its children;
- Its children's other parents.

Each node is conditionally independent of all others, given its Markov Blanket. The followings are the Markov blankets for the nodes **Sex** and **Genetic Influence**

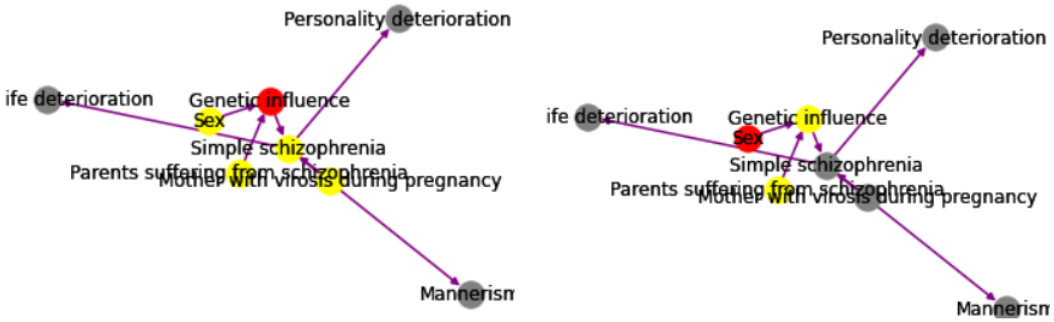


Figure 1: Markov blankets of **Sex** and **Genetic influence**

For conciseness sake, other Markov blankets can be found in the notebook.

4 Queries

4.1 Exact inference with variable elimination

The **variable elimination** is one of the most widespread method for exact inference. It's an improvement of the **inference by enumeration**, where the repeated sub-expression of the query are evaluated only once (*dynamic programming*).
Different orderings of the variables cause different intermediate factors to be generated during the calculation, but every choice yields a valid algorithm.
In general, the time and space requirements of variable elimination are dominated by the size of the largest factor constructed during the operation of the algorithm. This in turn is determined by the order of elimination of variables and by the structure of the network. It is intractable to determine the optimal ordering, but several good heuristics are available.

4.1.1 Some interesting queries

- The apriori probability of having a simple schizophrenia is (fortunately) kinda low:

Simple schizophrenia	Φ(Simple schizophrenia)
True	0.0207
False	0.9793

Table 3: P(Simple schizophrenia)

- If we start considering genetic causes and congenital issues, we will see that the probability will start to rise:

Simple schizophrenia	Φ(Simple schizophrenia)
True	0.025
False	0.75

Table 4: P(Simple schizophrenia|Genetic influence $= True$, Mother with virosis $= True$)

- The probability of having a simple schizophrenia, given the virosis during the pregnancy, gets higher as the number of parents with schizophrenia gets higher. That makes sense with the concept of familiarity of the disease.

Simple schizophrenia	Φ(Simple schizophrenia)
True	0.0817
False	0.9183

Table 5: $P(\text{Simple schizophrenia}|\text{Parents} = Zero, \text{Mother with virosis} = True)$

Simple schizophrenia	Φ(Simple schizophrenia)
True	0.0893
False	0.9106

Table 6: $P(\text{Simple schizophrenia}|\text{Parents} = One, \text{Mother with virosis} = True)$

Simple schizophrenia	Φ(Simple schizophrenia)
True	0.1446
False	0.8554

Table 7: $P(\text{Simple schizophrenia}|\text{Parents} = Both, \text{Mother with virosis} = True)$

4.2 Approximate inference

Exact inference is carried out with posterior probabilities that are not always tractable. Approximate inference methods address this issue by sampling from the untractable posterior (*stochastic methods*) or by approximating the posterior with a tractable distribution (*deterministic methods*).

To study this way of inference, we used two inference method [2]:

- **Rejection sampling**: is a general method for producing samples from hard-to-sample distribution given an easy-to-sample distribution. In its simplest form, it can be used to compute conditional probabilities. The algorithm first generates samples from the prior distribution specified by the network. Then, it rejects all those that not match the evidence. Finally, the estimate is obtained by counting how often the evidence occurs in the remaining samples;

- **Likelihood inference**: it avoids the inefficiency of rejection sampling by generating only events that are consistent with the evidence. The algorithm fixes the values for the evidence variables and samples only non-evidence variables. This guarantees that each event generated is consisted with the evidence. Nevertheless, not all events are equal. Before tallying the counts in the distribution of the query variable, each event is weighted by the likelihood that the event accords to the evidence, as measured by the product of the conditional probabilities for each evidence variable, given its parents. Intuitively, events in which the actual evidence appears unlikely should be given less weight.

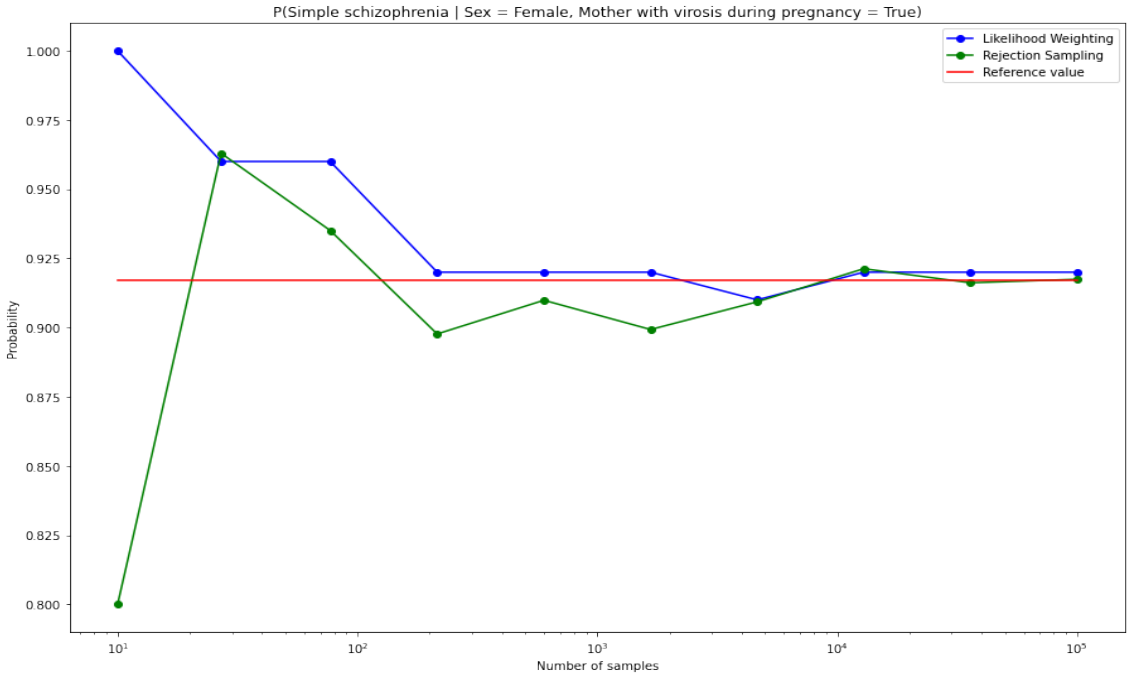


Figure 2: Approximate inference performance comparison

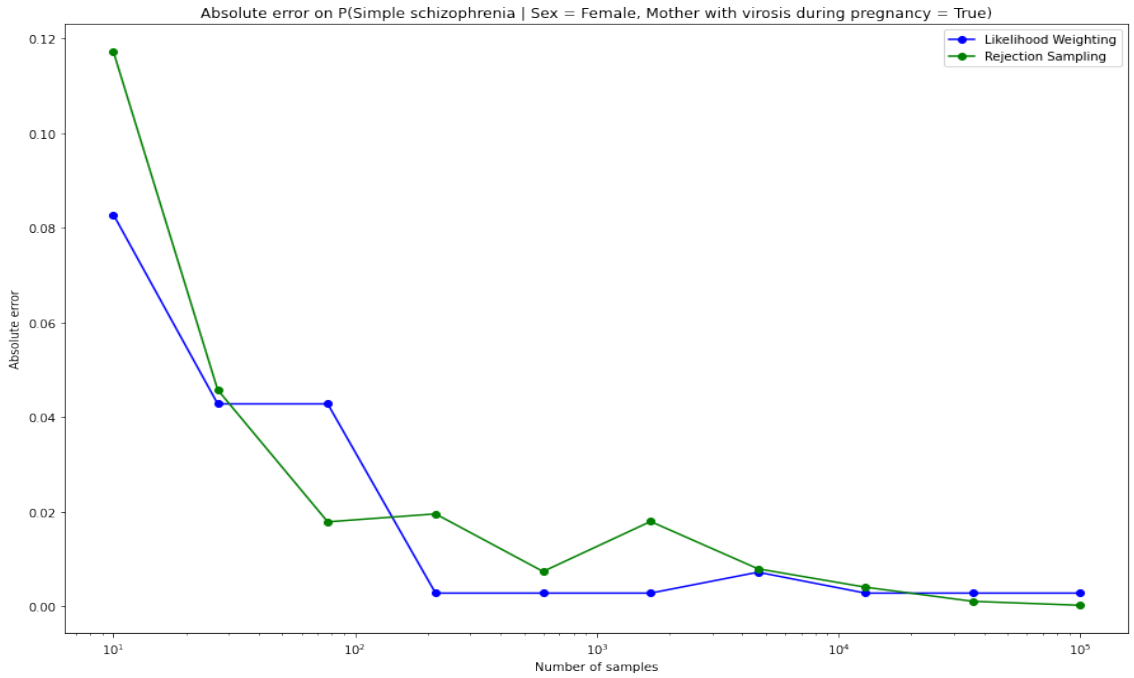


Figure 3: Absolute error of approximate inference

The graphs are showing that,for our model, the approximate methods requires a lot of computation w.r.t. the exact inference (10^5 samples). Despite of this computational effort, the result is good because we can see that both approximating method are converging toward the reference value which is computed by an exact inference.

5 Conclusion and future developments

During the development of the model, due to the lack of data, we could not develop the whole network proposed by the paper, but only a subgraph, whose data were present in the paper. Indeed, one of the future development could be the collection of the missing data and implement all other nodes, making this network more compliant for diagnosis purposes.

Then, we can also inquire more specifically about the type of causes, such as the timing of the virosis, and how they influence other nodes: could we use a continuous variable or a discrete to represents time? Can different types of virus influence the probability of other nodes as well? Does it depend on the RNA or DNA basis?.

References

[1] Daniel-Ioan Curiac et al. "Bayesian network model for diagnosis of psychiatric diseases". In: *Proceedings of the ITI 2009 31st International Conference on Information Technology Interfaces*. 2009, pp. 61–66. DOI: 10.1109/ITI.2009.5196055.

[2] Stuart Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 3rd ed. Prentice Hall, 2010.

[3] Roger M. Kurlan. "Habits, Mannerisms, Compulsions, and Stereotypies". In: *Hyperkinetic Movement Disorders*. Oxford, UK: Oxford University Press, July 2015. ISBN: 9780199375479. DOI: 10.1093/med/9780199925643.003.0007. URL: https://oxfordmedicine.com/view/10.1093/med/9780199925643.001.0001/med-9780199925643-chapter-7.

[4] Shen Liu et al. *Computational and Statistical Methods for Analysing Big Data with Applications*. English. Netherlands: Elsevier, 2016. ISBN: 9780128037324. DOI: 10.1016/C2015-0-00198-8.