Expert-driven approximation of MPE

Subtitle: Reinventing the World

Master's Thesis submitted to the
Faculty of Informatics of the *Università della Svizzera Italiana*
in partial fulfillment of the requirements for the degree of
Master of Science in Informatics
Artificial Intelligence

presented by
Thomas Francesco Tiotto

under the supervision of
Prof. Alessandro Facchini
co-supervised by
Prof. Alessandro Antonucci

September 2019

I certify that except where due acknowledgement has been given, the work presented in this thesis is that of the author alone; the work has not been submitted previously, in whole or in part, to qualify for any other academic award; and the content of the thesis is the result of work which has been carried out since the official commencement date of the approved research program.

Thomas Francesco Tiotto
Lugano, Yesterday September 2019

To my beloved

Someone said . . .

Someone

Abstract

This is a very abstract abstract.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et

vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Acknowledgements

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras

ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Contents

Figures

Tables

Chapter 1

Introduction

1.1 Context

state the general topic and give background of what your reader needs to know to understand the problem outline the current situation evaluate the current situation (advantages/ disadvantages) and identify the gap

In genere mancano referenze

While neural networks and Artificial Intelligence (AI) - as a field - have existed for nearly seventy years, the concept of artificial intelligence dates back to at least Ancient Greece. In ancient times, artificial intelligence embodied in mechanical men was part of the domain of myth; in the twentieth century, of that of science. During this last decade, Artificial Intelligence can't anymore be described by a limited set of terms, as it has materialised out of Man's imagination, broken out of laboratories and has been given lease to act in the world at large.

What do you mean? Why "reductive"?

No sector of our economy has been left untouched by the recent and rapid rise of machine learning that has been enabled by the rediscovery of deep neural networks, the availability of Big Data and cheap parallel computing power. Fields as diverse and as critical as are government, healthcare, finance and bioinformatics have been revolutionised and the possibility has been set for new ones - such as self-driving vehicles - to be born. The ever increasing reliance of our society on ever more complex machine learning-driven algorithms can only make us worry ever more about the ethical dilemmas posed by such a situation. Our society has only very recently been confronted with the dilemma of assigning blame when a driverless car causes the death of a person but this moral problem is only the tip of the iceberg, even when focusing only on the automotive industry. For example, how should a self-driving car behave when confronted with a real-world analogous of the classic Trolley Problem - a situation where each course of action is liable to cause harm? On what basis should a person be denied a mortgage, access to university or a job interview? How can we be sure that there is no bias in the system? How do we even define if the system is behaving morally? Would it currently be feasible for a person that feels they have been harmed by such a decision to appeal it, as prescribed by the recent EU General Data Protection Regulation (GDPR)? As more and more decisions are made in an automated way, with many of them significantly impacting both individuals and society at large, it comes natural to stop and wonder what are the characteristics we would want the systems making these decisions to have.

Explainable AI (xAI) is the sub-field of AI that rests at the intersection between Computer

Science, Social Sciences and Philosophy and whose aim is to define our desiderata of artificially intelligent systems and machine learning algorithms from the point of view of their explainability. The basic idea is that the prerequisite for the evaluation of the ethical and moral implications of a machine's decision is for the system to be "interpretable" or "explainable". Within the xAI community, there is currently no unanimously agreed upon definition of which these desiderata should be or of the best way to implement them in real systems. There is also no common, agreed-upon, definition of what is meant by the phrase "understanding a system": some authors equate it to having a *functional understanding*, void of the low-level details, while others decline it into the concepts of *interpretation* and *explanation*, the former indicating the output of a format that a human user can comprehend and the latter a set of features that have contributed to generating the system's decision.

The difficulties start even in trying to define what interpretability really is. Does it mean to gain the trust the system's user? Of type of user in particular? Does trust stem from some property of the decisions the system makes or from some other inherent characteristic of the machine? A common approach to solving the difficulty in defining interpretability is to try and define it post-hoc by categorising systems into ontologies, based on their perceived interpretablility; unfortunately this seems like a circular way of approaching the problem: the classification of system models is being done utilising the same criterion that is trying to be uncovered by doing so. For reference, a commonly used classification is the following:

- **Opaque systems**: these are systems that offer no insight into the mapping between inputs and outputs; all closed-source algorithms fall under this definition;

- **Interpretable systems**: this is the vastest category, as the characteristic of these systems is *transparency* i.e. their inner workings are accessible but the onus of comprehensibility falls completely onto the user. The classical example is that of neural networks where the mapping from inputs to outputs (the *weights*) is inspectable by the user who can, theoretically and depending on her skill, interpret them;

- **Comprehensible systems**: systems falling into this category emit additional symbols together with their outputs with the explicit intent of giving the user the means to interpret and understand the automated decisions; the additional symbols may be visualisations, natural-language text or any other means of demystifying the output. These extra symbols would need to be graded based on the user's expertise, as comprehension is a property that involves both man and machine but materialises on the human side.

Some authors propose to classify systems as *non-interpretable*, *ante-hoc interpretable/transparent* and *post-hoc interpretable*; this roughly corresponds to the ontology presented above.

What I hope can be gleamed from this brief introduction to the field of Explainable Artificial Intelligence, is that many of the problems it aims to tackle are hard *per-se* and may not have a unique optimal solution. This is because these issues are not only engineering problems, but exist at the intersection between man and machine and as such can't be tackled using only the methods of Computer Science. There is no way to satisfactorily investigate the human element of the situation without resorting to the well-established methods of the Social Sciences. There is little hope to know in which direction to procede without the guiding force that can only come from philosophy, because of its millennia-long tradition in thinking about ethical and high-level issues. It should be clear that when the human - and particularly the ethical - domain are part of the equation, it is impossible *by definition* to find an optimal and unique solution.

references to articles ?

explain better what follows here, since it is portant

what does functional mean?

what are low level details

??? explain better. gives examples, help th reader

does not sound well expressed here. also w has trust to do with interpretability? expla intuitions, motivations

why are you speaking about ontologies her not simply of categories etc?

in general. explain what a system, input a put are

References!!!

for instance? which methods? example?

you have to link what you just said with what you are gonna do later. for instance, illustrate what you are saying with examples. use one close to what you are going to do. also, you speak about NN but never about graphical methods, such as BN. you have to speak about them here too.

1.2 Problem and Significance

identify the importance of the proposed research - how does it address the gap? state the research problem/ questions state the research aims and/or research objectives state the hypotheses

again, references!

AI has a trust problem. The bigger problem with AI is not anymore its utility, as that has mostly been solved by deep neural networks, but its capacity to elicit the trust of the users. To be truly useful, an automated system should be able to make itself be trusted in a manner proportional to the criticality of its application. Unfortunately, the explainability and, by extension, the "trustability" of machine learning models are inversely proportional. There are many examples of modern methods - such as boosted trees, random forests, bagged trees, kernelized-SVMs - that show this tendency, but it is best exemplified by *deep neural networks* (DNN). Deep Neural Networks are machine learning models constructed by stacking many layers of artificial neurons, these systems are currently state of the art on a variety of tasks but are among the least easily interpretable systems due to the fact that they represent information in an implicit and distributed manner among their network weights. Some older methods, like decision trees or rule-based methods, are inherently more interpretable due to their simplicity and the fact that they can explicit state their reasoning steps, but are less accurate and flexible than more modern techniques.

ok but come on, there is not only NN in th world

The runaway success obtained by modern Machine Learning in a variety of domains, on a spectrum that goes from engineering to social work, has created the desire to also start applying these methods to mission-critical and traditionally more entrenched fields. A perfect example of a field exhibiting both these characteristics is that of medicine. The first successful artificially intelligent systems date back to the 1970s and '80s and were based on *symbolic methods* integrated with *knowledge-bases*. These systems were by design capable of providing an explanation for their reasoning and were thus accepted by the medical community in an implementation known as *expert systems* that aimed to perform functions similar to those of a human expert. The deficiency of modern AI methods in being able to provide causal links for their reasoning process has held back their acceptance in the field of medicine, regardless of their superior performance and accuracy.

In a high-stakes domain such as the medical one, it would be unthinkable for a doctor to trust the predictions of an AI system a priori; any decision with profound moral implications - such as prescribing or interrupting the treatment of a patient - would have to first be validated by a human. The possibility of carrying out this validation and its quality are dependent on the degree of interpretability of the model that made the decision. Unfortunately, as has been repeated many times, the best performing models are often also the most opaque to inspection.

Explainability is not a necessary condition only for the verification of the system which, as we have just discussed, is a presupposition for it to be applied in mission-critical domains, but also for the extraction of knowledge from data. The amount of information that a machine learning model can process is many orders of magnitude greater than that inspectable by any human; this may let a computer spot new patterns in the data that aren't immediately apparent or are latent given only a moderate amount of samples. Being able to turn this information into

not clear. you were speaking about "valida before. is this taken as a synonymous of "v ification"? if yes, why two words for the sa concept? if not, then explain better

why?

new knowledge implies the system having the ability to output human-interpretable symbols that are capable of communicating it in a comprehensible and effective way.

so, what is knowledge? why explainability essary to create knowledge? you do not ex this here, imho

There has recently been much research carried out on trying to explain and extract knowledge from deep neural networks together with attempts to marry the connectionist and symbolic approaches to artificial intelligence - a subfield known as *neuro-symbolic computation* while also reconsidering mixed approaches such as *Bayesian Networks*. A Bayesian Network is a graphical and computationally efficient way of representing dependencies between random variables. The graphical component is immediate as in the model each random variable is represented by a node of a Directed Acyclic Graph (DAG), with the edges connecting them standing for their dependencies. The efficiency stems from the fact that the graph structure imposes a factorisation of the joint probability space and thus lets each variable be calculated using only the values of its parents.

again, a lot of nice talk about NN, but then little about BN and thus what you are gonna do

1.3 Response

outline the methodology used - outline the order of information in the thesis - a roadmap - Maximum 2500 words.

The work carried out in this thesis concentrates on explainability in the medical domain and presents both a practical part, with the implementation of a Bayesian network-based system focused on *knowledge-extraction*, as defined at the end of the previous section, and a theoretical one, regarding the definition and validation of desiderata for an artificially intelligent system using the aforementioned system.

The implemented system aims at supporting medical decision making through the instauration of a dialogue with the user/domain expert. To this end, the information implicit in the data is used as basis for a constructive dialogue with the user; this starts with the expert informing the system of which knowledge is certain i.e. a variable's value that has been observed in a specific patient, and continues via a process where the next most probable *(variable, state)* pair is proposed, with the expert having the choice of accepting it or refusing it, if she believes that the variable under examination doesn't adequately explain the accumulated evidence. Each accepted variable is added to the evidence set, as the system gives priority to the domain expert's judgement. The result of the dialogue is an *explanation tree* whose nodes represent *(variable, state)* pairs and are organised into branches, depending on the flow of the dialogue; more specifically, there will always be a *main branch* corresponding to the choices of the user and none or more *alternative branches* whose role is to inform the expert of the possible alternative outcomes to his decisions.

This software system was developed and tested tested in collaboration with *Istituto Cantonale di Patologia*, a medical institute in Locarno, Ticino, Switzerland that specialises in the analysis of tissue samples received from hospitals, clinics and private doctors. Its main activity is to characterise the samples by using *histo-cytopathologic techniques*, with particular focus on the diagnosis of cancer and tumoural diseases in general.

The theoretical part of this thesis aims to understand how an

Chapter 2

Literature review

What is explainability? How is it defined? By whom? When? Why is it important? Notable works in the field

Suggestions when collecting references:

Copy everything in one document (with references !), but do not use it directly when writing the text

Copy it in the document (with references !) and color it , but do not use it directly when writing the text

What is the main point of the sentence / paragraph / article ?

What do I want to convey ?

Read a few references / paragraphs and only then write it down

2.1 Explainability

2.2 Importance of Explainability

2.3 Notable Works

2.4 Explaining the Most Probable Explanation

Butz et al. [2018]

controllare se qualcuno ha lavorato nell'est
i metodi del paper

2.5 A Progressive Explanation of Inference in "Hybrid" Bayesian Networks for Supporting Clinical Decision Making

Chapter 3

Methodology

The methodology section of a research paper answers two main questions:

1. How was the data collected or generated?
2. Which methodology was used to analyze it?

In writing the methodology chapter, explain the methodology you are using and why you chose it. Your reader needs to know the method you used to get your data because it effects your findings, and how you interpret your findings. It can be helpful to think of this piece as a "manual" - write so that someone can pick up your paper and do the same work you have done. Be direct and precise.

Your methodology is crucial to the quality of your research. It is important that the methodology is appropriate and your reasons for selecting it are clear. You should follow accepted practices; there is no need to re-invent a method. Your literature review can help you determine the method to use.

What goes into Methods chapter ?

It is all about HOW you will do the study What is the method ?

What does NOT go into Methods chapter ?

Results Implementation

introduzione con outline del capitolo

riassunto della fine del capitolo

3.1 Mathematical Background

Bayesian Networks (BN) are a class of Probabilistic Graphical Models that are used to represent systems under conditions of uncertainty. To give a formal definition we will first need a few basic concepts from probability and graph theory.

3.1.1 Probability Theory

Probability distributions

Definition 3.1 *A probability distribution is a function* $\mathbb{P} : \mathscr{S} \to \mathbb{R}$ *with* $\mathscr{S}$ *a set of events of interest. To be a valid probability distribution* $\mathbb{P}$ *must satisfy:*

- $\mathbb{P}(\sigma) \geq 0 \quad \forall \sigma \in \mathscr{S}$

- $\sum_{\sigma} = 1 \quad \forall \sigma \in \mathscr{S}$
- $\alpha, \beta \in \mathscr{S} \wedge \alpha \cap \beta = \emptyset \Rightarrow \mathbb{P}(\alpha \cup \beta) = \mathbb{P}(\alpha) + \mathbb{P}(\beta)$

Each event $\sigma \in \mathscr{S}$ must have a probability $\mathbb{P}(\sigma) \in [0, 1]$ and the sum of all these must equal 1. An event with $P(\sigma) = 0$ is deemed *impossible* while one with $\mathbb{P}(\sigma) = 1$ is *certain*.

There is some discord regarding how to actually *interpret* the probability of an event. What I believe to be the initially commonly held view is the *frequentist* one, that views the probability of an event as the ratio of times it would occur over a great number of trials. So, for example, saying that obtaining a heads has probability 0.5 when tossing a coin would mean that over repeated throws we would observe heads half the time.

Another, commonly held view is the *Bayesian* (from the 18th century mathematician Thomas Bayes) one in which probabilities are viewed as the *subjective* degree of belief attributable regarding the manifestation of an event. In this interpretation, stating that a coin has 0.5 probability of landing on heads simply means that the person making the claim personally believes that the chances of seeing heads of tails are the same. This is obviously a "softer" definition compared to the frequentist one but it is nonetheless useful in that it lets one characterise certain events that haven't come about yet or are liable to happen only once or a few times.

Philosophically, Bayesian inference assigns a probability to a hypothesis (a *prior*) while the frequentist method tests a raw hypothesis empirically before assigning it any probability. As Bayesian inference naturally embraces and deals with uncertainty, it is an enormously useful tool to model and reason about the real, stochastic world we live in.

Random Variables

Definition 3.2 *A random variable is a function that associates every outcome in $\mathscr{S}$ with a value.*

Random variables are a way of bringing to the fore the attributes of interest of events while dealing with them in a clean, mathematical way. The values that a random variable can take are a function of the events in sample space $\mathscr{S}$, each of these is assigned a value by the random variable function. I will only be dealing with *categorical random variables* i.e. those who's codomain is a discrete set of values. Every random variable has a probability distribution induced by the cardinality of the subsets of its values; in the case of categorical-valued one, such a distribution is *multinomial*.

If we were to take a Bayesian point of view, we would consider a random variable as simply representing the subjective degree of belief we would have over a set of outcomes we believed possible.

Conditional Probabilities

After having defined the basic notion of probability, we can construe one the basic building blocks of Bayesian Networks: the concept of *conditional probability*

Definition 3.3 *The conditional probability, "the probability of event β given event α" is:*

$$\mathbb{P}(\beta \mid \alpha) = \frac{\mathbb{P}(\beta \cap \alpha)}{\mathbb{P}(\alpha)} \tag{3.1}$$

That is, the relative proportion of event β compared to event α; this intuitively represents the probability of β *knowing* that α has already occurred.

Equation 3.1 can be easily manipulated to obtain another basic element of Bayesian Networks: what is called the *chain rule of conditional probabilities*:

$$\mathbb{P}(\beta \cap \alpha) = \mathbb{P}(\beta \mid \alpha)\mathbb{P}(\alpha) \tag{3.2}$$

This can be generalised to any number of events:

$$\mathbb{P}(\alpha_1 \cap \ldots \cap \alpha_n) = \mathbb{P}(\alpha_n \mid \alpha_1 \cap \ldots \cap \alpha_{n-1}) \ldots \mathbb{P}(\alpha_1 \mid \alpha_2)\mathbb{P}(\alpha_1) \tag{3.3}$$

Intuitively, it means that we can decompose joint probabilities as products of conditional probabilities. As we'll see, this is how the values in a Bayesian Network are calculated.

Independence

Now, we have just seen in Equation 3.1 that, in general, $\mathbb{P}(\beta|\alpha) \neq \mathbb{P}(\alpha)$ because $\mathbb{P}(\beta \cap \alpha) \neq \mathbb{P}(\beta)\mathbb{P}(\alpha)$

Definition 3.4 *Two events α and β are unconditionally independent $A \perp B$ - or simply independent - when:*

$$\mathbb{P}(\beta \mid \alpha) = \mathbb{P}(\beta) \Leftrightarrow \beta \perp \alpha \tag{3.4}$$

This means that knowing that α took place doesn't change our beliefs around β happening. In the real world it is hard, or actually impossible if we consider existence at a fine-enough level to involve Chaos Theory, to find two such perfectly non-interacting events. Thus, a more useful concept is that of *conditional independence* where two previously dependent event become independent when also conditioned on a third one.

Definition 3.5 *Two events α and β are conditionally independent $(\beta \perp \alpha \mid \gamma)$ when:*

$$\mathbb{P}(\beta \mid \alpha \cap \gamma) = \mathbb{P}(\beta \mid \gamma) \Leftrightarrow (\beta \perp \alpha \mid \gamma) \tag{3.5}$$

Correlation

Correlation, as defined by Stolp et al. [2006], is a measure of the degree to which two random variables are linearly dependent. The most used measure of such a dependence is the *Pearson Correlation Coefficient* or *bivariate correlation*.

Definition 3.6 *The correlation coefficient ρ of random variables X and Y is given by:*

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.6}$$

$$= \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \tag{3.7}$$

$$= \frac{\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} (x-\mu_X)(y-\mu_Y) p_{XY}(x,y)}{\sqrt{\sum_{x \in \mathcal{X}} (x-\mu_X)^2 p_X(x)} \sqrt{\sum_{y \in \mathcal{Y}} (x-\mu_Y)^2 p_Y(y)}} \tag{3.8}$$

with p_{XY} the joint probability mass function of X and Y and p_X and p_Y the marginal distributions of X and Y, respectively.

That is, the bivariate correlation coefficient for random variables X and Y is given by the *covariance* of X and Y divided by the product of their standard deviations.

The covariance is the *first centred moment* of the *joint distribution* of X and Y while the *standard deviation* is the square root of the *second centred moment* of the marginals.

ρ_{XY} is normalised so its values vary in the interval $[-1, 1]$; the correlation coefficient represents the degree of linear association between the two variables with $\rho_{XY} = -1$ being called *perfect anticorrelation* and $\rho_{XY} = +1$ *perfect correlation*. The two correspond to the cases where the linear equation perfectly describes the relationship between X and Y; the sign indicates the slope of the regression line describing the relationship i.e. if an increase in one of the two variables corresponds to an increase in the other in the pair, or viceversa. The closer ρ_{XY} tends to 0, the feebler the relationship between X and Y with the case $\rho_{XY} = 0$ indicating that the two variables are *independent*.

Mutual Information

Another way of characterising the interrelatedness of two variables is through the concept of *mutual information*, as defined in Cov [1967], that is closely linked to entropy, see Eq. 3.10.

Definition 3.7 *The mutual information of two random variables X and Y is given by:*

$$I_{XY} = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p_{XY}(x, y) \log\left(\frac{p_{XY}(x, y)}{p_X(x) p_Y(y)}\right) \tag{3.9}$$

I_{XY}, intuitively, measures the amount of information that X and Y share that can also be seen as the degree to which one variable is informative of the other. If X and Y are independent then they share no mutual information and knowing one of the two gives no information about the other. if X and Y are perfectly correlated ($\rho_{XY} = \pm 1$) then they both convey the same amount of information and I_{XY} is equal to the entropy $H(X) = H(Y)$.

3.1.2 Information Theory

The birth of the field of *information theory* is usually traced back to the seminal paper "A Mathematical Theory of Communication" (Shannon et al. [1949]) where Claude Shannon set the mathematical basis for the quantification of the amount of *information* transmissible over a noisy channel. In his words "The fundamental problem of communication is that of reproducing at one point, either exactly or approximately, a message selected at another point." The concepts of field are broad enough to have influenced practically every other scientific discipline and deep enough to have enabled the "digital age", for example by enabling the creation of ever more complicated coding schemes for the compression, reconstruction and obfuscation of digital data.

Entropy

In classical mechanical statistics, entropy can be seen as a measure of the uncertainty, or randomness, of a physical system. This concept was reapplied by Shannon to measure the amount of randomness in a random variable.

Definition 3.8 *Given a random variable X with probability distribution $\mathbb{P}(X)$, its entropy $H(X)$ is defined as the expected amount of information content carried by X (Schneider [2005]):*

$$H(X) = \mathbb{E}(I(X)) = \mathbb{E}(-\log(\mathbb{P}(X)) = -\sum_{i=1}^{n} \mathrm{P}(x_i) \log_b \mathrm{P}(x_i) \tag{3.10}$$

The base b of the logarithm defines the unit of measure. Shannon used $b = 2$ as he was dealing with the transmission of digital, binary-coded data; in this case the unit of measure are $bits$.

The simplest example of how information entropy characterises a random variable X, is in imagining X to model a coin and the task being to predict the probability of the outcome of a throw being heads. If the coin is fair, we will not be any more surprised to see the outcome being heads than tails; the entropy is maximum as there is maximum uncertainty regarding the outcome. However, if the coin is not fair and tails is more probable the we will be more surprised than not to see the outcome being heads. The entropy is sub-maximal because there is less uncertainty regarding the outcome: tails is more probable than heads. If one of the outcomes is impossible, for example if the coin has two heads, then the entropy of the coin is 0 as there is no uncertainty regarding the result of a toss.

Normalised Entropy

Plain entropy is not a good choice when trying to characterise random variables with different cardinalities of their sample space. Let us suppose that the objective is to find the variable with the least "entropic" distribution and we suppose that their values have all been generated by the same process, say Gaussian. Simply calculating their entropies and ordering them according to this criterion will bias the selection process towards the variables with smallest cardinality. This is because we supposed them to be distributed in the same way so there will naturally be less uncertainty when there are fewer possible outcomes. This can easily be understood by imagining the distributions to all be random uniform.

To obviate to this problem we need to *normalise* the entropy so that different-sized variables can be directly compared to each other. To achieve this, we can look at a measure of *normalised entropy* or *efficiency*:

$$\eta(X) = -\sum_{i=1}^{n} \frac{p(x_i) \log_b (p(x_i))}{\log_b(n)} \tag{3.11}$$

From Eq. 3.11 it can be seen that $\eta(X) \in [0, 1]$; it is thus normalised and comparable among distributions. This ratio expresses the amount of entropy found in the distribution compared to the maximum possible entropy when using n symbols, corresponding to the uniform distribution:

$$\mathrm{H}\left(\underbrace{\frac{1}{n}, \ldots, \frac{1}{n}}_{n}\right) = -\sum_{i=1}^{n} \frac{1}{n} \log_b \left(\frac{1}{n}\right) = -n \cdot \frac{1}{n} \log_b \left(\frac{1}{n}\right) = -\log_b \left(\frac{1}{n}\right) = \log_b(n) \tag{3.12}$$

3.1.3 Graph Theory

Many problems in Machine Learning (ML) don't involve classification or prediction of single data points in isolation, but of set of entities that may present a more, or less, complex relation with each other. Most real-world phenomena fit into the latter framework. Graphs are one of

the most powerful tools for the modelling of this class of problems, as their structure naturally captures the wide variety of relations that may exist between entities. These range from the atomical structure of a molecule to a social network of friends. In both these examples graphs help in reasoning, visualising and making inferences and predictions.

Graphs

Definition 3.9 *A graph is a tuple*

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}) \tag{3.13}$$

with $\mathcal{V} = \{v_1 \dots v_n\}$ *the set of vertices and* $\mathcal{E} = \mathcal{V} \times \mathcal{V}$ *the set of edges.*

For our scopes, we will only be considering the case where every element in $\mathcal{E}$ is a pair either of the form (v_i, v_j) or (v_j, v_i) with $i \neq j$. That is to say that the class of graphs presently of interest for us are those where there can be at most a single directed edge between any node in $\mathcal{V}$ and no self-loops. We are also interested in enforcing that there be no *cycles* in the graph, i.e. sequences of nodes of the form $v_i \to v_j \to \cdots \to v_i$. The resulting graph possessing only directed edges and no cycles is commonly called a *directed acyclic graph*, or DAG for short. This data structure is of paramount importance as it's the fundamental graphical representation used for Bayesian Networks.

Polytrees

We now have all elements to be able to formally define a Bayesian Network. I will also define polytrees and trees because these are a fundamental concept for the work carried out in this thesis.

Definition 3.10 *A loop is a trace* $v_i, v_j \dots v_i$ *of nodes obtained by following edges regardless of their direction*

Definition 3.11 *A directed graph containing no such loops is called a polytree.*

Definition 3.12 *A tree is a particular case of polytree where each node has at most one parent.*

D-separation

Dependence-separation or *d-separation*, as the name entails, is a concept relating to the conditional dependence between variables. It was first presented by Pearl and Dechter [1988] To define it, we first have to clarify when two sets of nodes X and Y are causally connected. This is so if $Z = \emptyset$ and they are part of one of the following three structures, called *v-structures* in this context:

- $X \to Z \to Y$
- $X \leftarrow Z \leftarrow Y$
- $X \leftarrow Z \to Y$

This means that knowing something about X also tells us something new about Y. X and Y are causally independent if they appear in the following v-structure:

- $X \to Z \leftarrow Y$

Such a configuration is called a *collider* and it blocks the flow of information from X to Y. If $Z \neq \emptyset$ the cases are reversed so colliders are open and the other three structures are blocked.

Definition 3.13 *Given disjoint subsets $X, Y, Z \subset \mathcal{X}$, X and Y are d-separated if:*

- *$Z \neq \emptyset$: no path between X and Y presents a collider*
- *$Z = \emptyset$: there is a collider on every path between X and Y*

The independencies between variables are encoded in the structure of the DAG so every distribution whose BN has the same connections between nodes also has the same independencies, regardless of the values of the variables.

An example using on the real network used in this thesis is shown in Fig. 3.1. Here we can see how the network's topology and the nodes chosen to be in the observed set Z define the resulting separations. In this case $X = \{\text{mut17q21}\}$, $Z = \{\text{differenziazione, pN}\}$ and $Y = V \setminus X \setminus Z$; the resulting set of nodes that are d-separated from X contains {c erB 2, FISH, loss 17, situ, pT, M, lateralita}. Basically, we are asking for the set of all nodes in the DAG that are d-separated from X, given evidence Z. The reason for this can easily be given by enumerating all paths through the v-structures in the network and applying the definitions for causal connections given above:

- mut17q21 → eta arrotondata: are directly connected so the cannot be independent
- mut17q21 → recettori estrogeni: like above
- mut17q21 → recettori estrogeni → morfologia: recettori estrogeni is not in Z so the path is open and mut17q21 is d-connected to morfologia
- mut17q21 → recettori estrogeni → recettori progestinici: like above
- mut17q21 → recettori estrogeni → ki67: as above
- mut17q21 → recettori estrogeni → differenziazione: as above
- mut17q21 → recettori estrogeni → differenziazione → c erbB 2: differenziazione is in Z so this path is closed, as there are no other paths between mut17q21 and c erbB 2, the two nodes are d-separated
- mut17q21 → recettori estrogeni → differenziazione → FISH: the only path to mut17q21 is through c erbB 2, that is already d-separated from mut17q21, so the two nodes are d-separated
- mut17q21 → recettori estrogeni → differenziazione → FISH → loss 17: as above
- mut17q21 → recettori estrogeni → differenziazione → pN: as recettori estrogeni, differenziazione and pN are arranged in a *collider* and differenziazione is in Z, then the path is open
- mut17q21 → recettori estrogeni → differenziazione → pN → situ: pN is in Z and differenziazione, pN, situ are arranged as a *fork*, thus the path is closed
- mut17q21 → recettori estrogeni → differenziazione → pN → pT: as above, M is also d-separated from mut17q21 so there are no open paths between mut17q21 and pN

- mut17q21 → recettori estrogeni → differenziazione → pN → M: differenziazione, pN, M are arranged as a *chain* and pN is in Z, the path is closed
- lateralita: lateralita is not connected to any other node and is thus d-separated from all others a priori

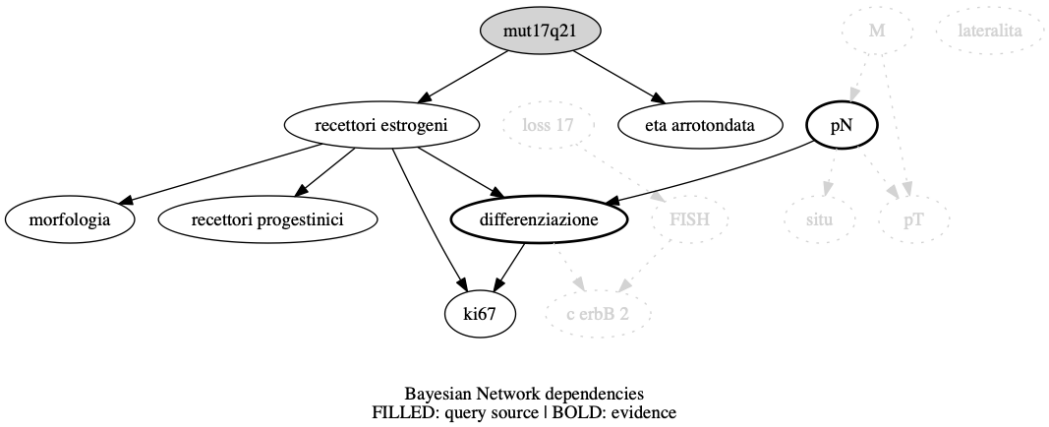


Figure 3.1. D-Separations in the provided data set (see Sec. 3.4) generated using the `plot_model` function (see Subsec. 3.3.3)

3.1.4 Bayesian Networks

Definition 3.14 *A Bayesian Network (BN) is a probabilistic graphical model represented by a DAG where each vertex corresponds to a random variable X_i and the edges model the dependencies among these.*

Such a model is basically a way of compactly representing an explicit joint distribution $\mathbb{P}(X_1 \cap \ldots \cap X_n) = \mathbb{P}(X_1)\ldots\mathbb{P}(X_n)$, that is factorised into $\mathbb{P}(X_n \mid X_1 \cap \ldots \cap X_{n-1})\ldots\mathbb{P}(X_2 \mid X_1)\mathbb{P}(X_1)$. The way this compactness is achieved is in exploiting the independencies that exist among the random variables:

$$\forall X_i : (X_i \perp \neg Desc(X_i) \mid Pa(X_i)) \tag{3.14}$$

with $Pa(X_i)$ the set of nodes that are parents of X_i and $Desc(X_i)$ the nodes that are not *descendents* of X_i. That is to say, every random variable X_i, given its parent nodes, is independent of all other nodes in the Bayesian Network that are not descended from it. Also, a BN gives the flexibility to drop the many weak dependencies that are bound to exist between variables thus leading to an even simpler model. A full probability table for a joint distribution of random variables obscures the independencies and requires an exponential number of entries for the representation. A Bayesian Network on the other hand can represent the same distribution using only a linear number of parameters. The way that Bayesian Networks can be used to reduce the storage requirements for uncertain information is by taking advantage of the conditional independencies embedded in the underlying distribution being modelled. The power of BNs comes from the additional information encoded in their structure and this was first explicitly described in its entirety by Pearl and Dechter [1988] who defined the concept of dependence separation (see Subsec. 3.1.3) and applied it to Bayesian Networks.

One nice characteristic of BNs is that they very naturally model the type of mixed causal and stochastic processes that we find in all of Nature. Imagine we want to represent the process modelled by joint distribution $\mathbb{P}(B,A) = \mathbb{P}(B)\mathbb{P}(A)$; using the chain rule for conditional probabilities (Eq. 3.3) we can write this as $\mathbb{P}(B \mid A)\mathbb{P}(A)$. A BN modelling this process would be composed of two nodes A and B with an edge from the former to the latter $A \rightarrow B$, A is called the "parent" of B. Each of these two nodes would have its own probability table, with $\mathbb{P}(A)$ representing the *prior* distribution over A and $\mathbb{P}(B \mid A)$ the *conditional probability distribution* of B given A.

We can now see why these types of models are named *Bayesian* Networks: the inference process is based in a given prior distribution/belief and evolves through a parent $\rightarrow$ child relationship to constantly yield an updated *posterior* belief. The BN DAG encodes a generative sampling where each variable's value is determined stochastically by Nature, based on the value of its parents. This process is also highly compatible with our view of causality and this is one of the reason that makes BNs highly interpretable. The prior $\mathbb{P}(A)$ can be seen as the result of some stochastic process caused by a series of latent (unmodelled) variables while the posterior $\mathbb{P}(B \mid A)$ is stochastically, causally determined by A. As I have mentioned in the previous paragraphs, there are probably no truly "prior" distributions in the Universe, at the modelling scale we are usually interested in. Only on arriving on the quantum particle level may we find "pure" stochastic, uncaused processes due to quantum collapse.

A good example of how BNs are well compatible with our notion of causality may be to imagine A as the random variable modelling the predisposition to having a certain disease and B to actually developing the symptoms for it. *First*, genetic and epigenetic factors such as the environment stochastically contributed to having the predisposition and *then* the development of the symptoms was stochastically determined by the degree of predisposition. Adding an extra time dimension certainly helps us in dealing with this class of models.

3.1.5 Bayesian Networks Structure Learning

In many probabilistic models initialisation is fast but then fitting the data is slow (ex. k-means). For Bayesian Networks the converse is true: fitting is fast as only sums of the counts in the data are needed but identifying the correct graph structure can take super-exponential time. Learning the Bayesian Network structure from data is commonly known as the Bayesian Network Structure Learning (BNSL) problem. The methods to solve this problem can be roughly categorised into one of three types.

Search and Score

This is the most naïve method as it does a brute force search over all the possibile graph structure space - i.e. all DAGs with the same number of variables as the input data - and scores all these depending on some cost function. This process is super-exponential but though the use of dynamic programming and heuristic search algorithms it can become sub-exponential. Nonetheless, solving the exact BNSL is only feasible up to 30 variables.

Constraint Learning

Methods of this type calculate some measure of correlation to identify the presence and direction of edges between nodes. A typical test is to iterate over all triplets while testing for conditional

independencies. Thanks to the d-separation properties outlined in Sub-Section. 3.1.4, this test is able to identify the correct edges. The algorithm is quadratic in time in the number of vertices.

Approximations

Several heuristical approaches have been developed to be able to find good network structures in an efficient manner. Examples of these are:

- Chow-Liu, that builds a tree approximation of the probability distribution
- Greedy hill-climbing, that adds/removes/flips an edge at a time
- optimal reinsertion, that iteratively calculates the optimal $Markovblanket$ (the subset of all nodes that are sufficient to determine the value of another subset) of an ever-smaller subset of nodes

3.1.6 Bayesian Networks Updating

All the types of inference presented are instances of *diagnostic reasoning*, also known as *abductive reasoning*. This type of explanation can either be modelled as a conditional probability or a MAP query and is of fundamental importance in many important problems of machine learning including medical diagnosis, that is of particular interest to us.

Conditional Probability Query

The *updating* problem is the process of updating the probabilities of nodes in the BN based on the observation of the values of other vertices. This process of conditioning on observed information is also called *data propagation*.

The following algorithm was described by Normand and Tritchler [1992] and applies to our case where the random variables follow a multinomial distribution. What we want, is to calculate the conditioned probability $\mathbb{P}(B \mid D)$ i.e. the updated probability of node B based on observed evidence E.

Definition 3.15 *The conditional probability query for variable B given evidence E is:*

$$\mathbb{P}(B \mid E) = \alpha\pi(B)\lambda(B) \tag{3.15}$$

with $\pi(B)\lambda(B)$ analogous to the prior and likelihood of B, respectively.

The likelihood of B depends only on the weighted likelihoods of its children $C_1, \ldots, C_k$:

$$\lambda(B) = \prod_l \lambda_{C_l}(B) \tag{3.16}$$

$$\lambda_{C_l}(B) = \sum_{C_l} \lambda(C_l)\, P(C_l \mid B) \tag{3.17}$$

and its prior similarly depends only on the information received from its parents A:

$$\pi(B) = \sum_A P(B \mid A)\pi_B(A) \tag{3.18}$$

$$\pi_B(A) = \alpha\pi(A) \prod_{S_B} \lambda_{S_B}(A) \tag{3.19}$$

The information is propagated down if any variable observed is above B while up if any variable observed lives in the tree rooted in B. Initially all leaf nodes' likelihoods are set at 1 and the priors of root nodes are assumed to be observable.

Maximum a Posteriori Query

Another common type of question we might ask a BN is the following: "given evidence E which is the most likely assignment of a subset of variables Y?". This is know as *Maximum a posteriori (MAP)* inference and is a much harder problem that a conditional probability query. We are trying to solve the an optimisation problem.

Definition 3.16 *As defined by Koller et al. [2007a]. Given evidence/observed variables $E = e$, $E \subseteq \mathcal{X}$ and sets $Y \subseteq \mathcal{X} - E$ and $Z = \mathcal{X} - E - Y$, with $\mathcal{X}$ the set of all variables in the BN, the MAP query for Y is the assignment of values $Y = y$ that has maximum probability:*

$$MAP(Y = y \mid E = e) = \underset{y}{argmax} \sum_{z} \mathbb{P}(Y = y, Z = z \mid E = e) \tag{3.20}$$

The MAP problem is hard to solve efficiently; that is it is part of the *NP-hard* complexity class, as proved by Shimony [1994]. Calculating it in a brute-force way would mean elencating all the possible variable-value tuples and computing their joint probabilities; as these are exponential in the number of variables, the problem is evidently untractable. Moreover, this is true even in a Bayesian Network. Such a model may possess a linear number of parameters but the underlying distribution is still exponential. Explicitly calculating the MAP defeats the very purpose of the BN, that is computational efficiency. For this reason, there exist a host of approaches to optimising MAP: elimination algorithms, gradient methods, simulated annealing and other stochastic local searches, belief propagation and integer linear programming.

Most Probable Explanation Query

A special case of MAP is the *Most probable explanation (MPE)* that,

Definition 3.17 *As defined by Koller et al. [2007a]. Given evidence/observed variables $E = e$, $E \subseteq \mathcal{X}$ and $W = \mathcal{X} - E$, the MPE query for W is the assignment of values $W = w$ that has maximum probability:*

$$MPE(W = w \mid E = e) = \underset{w}{argmax}\, \mathbb{P}(W = w \mid E = e) \tag{3.21}$$

This is an easier problem than MAP, as can be seen by comparing Eq. 3.20 with Eq. 3.21; MAP presents both a summation and a maximisation and as such is part conditional probability query, part MPE query. All algorithms for the computation of MAP obviously apply to MPE too, but there exist efficient approximate algorithms for MPE that do not generalise to MAP such as Loopy Belief Propagation

3.2 Methods

3.2.1 Libraries

Pomegranate

pomegranate (Schreiber [a]) is an open-source probabilistic models package for python. Its core philosophy is that every probabilistic model, from Hidden Markov to Bayesian Network, can be seen as a probability distribution and, as such, can be flexibly composed into hierarchical mixture models (Schreiber [2017]). The package implements:

- Probability Distributions
- General Mixture Models
- Hidden Markov Models
- Bayes Classifiers and Naïve Bayes
- Markov Chains
- **Bayesian Networks**
- Factor Graphs

This package was chosen among others for its good implementation of Bayesian Networks, its clear API and its performance. The package is written in cython and natively supports multi-core parallelism and out-of-core learning. Network structure learning from data, described in 3.1.5, appears to be particularly efficient, thanks to the implementation of prior knowledge into the graph selection process as described by Schreiber and Noble [2017]. The claim of this novel selection process is that it possesses the speed of a heuristic approach while yielding a far better quality estimate.

pomegranate currently only supports Discrete Bayesian Networks so the random variable of each node must have a categorical distribution.

Structure learning from data is achieved using the `from_samples` method of the `BayesianNetwork` class, with the default algorithm being the novel one described by Schreiber and Noble [2017]. The *probability* of a sample is calculated using the `probability` function of an object of `BayesianNetwork` type; the `predict_proba` function is used to return the probability of each variable in the model given some evidence. *Predictions* (described in detail in Sec. 3.1.6) are run by passing to the `predict` function of an object a matrix with `None` as placeholders for missing values . *Fitting* is done thought the `fit` function that uses MLE estimates to update each node's distribution in the model based on the input data.

A `BayesianNetwork` object can also be displayed graphically by calling its `plot` function. The output is a DOT file that is generated using the PyGraphviz package (PyGraphviz developer team), that is a python interface to the famous Graphviz (gra) graph visualisation software. An example of such an output is shown in Fig. 3.2.

Gurobi

Gurobi (gur) is a closed-source mathematical programming solver for Linear Programming, Quadratic Programming and Mixed-Integer Programming optimisation problems. It claims to

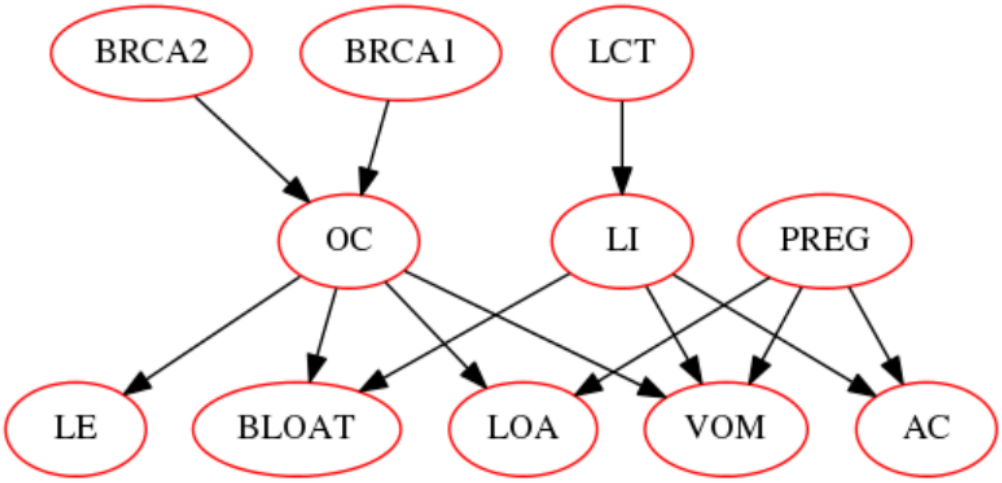


Figure 3.2. Example output of `plot` (Schreiber [b])

be the fastest solver available for these classes of problems. Gurobi offers object-oriented and matrix-oriented interfaces to, among others, Python, MATLAB and Excel.

Pandas

pandas (pan) is an extremely widely-used open-source python library that provides data structures and methods to aid in data analysis. The package excels in the manipulation of tabular data in the form of `DataFrame`, that is the analogous of R's `data.frame`. A `DataFrame` can be seen as a "general 2D, size-mutable structure with potentially heterogeneously-typed columns". The syntax for slicing is very close to R's as are many other functionalities; this is because one of Pandas' explicit goals is to offer all of CRAN's functionalities and be easily approachable by anyone already knowing the other language.

Pandas the default choice for this thesis' implementation because it is the *de facto* standard in data analysis applications. Its flexibility in reading Excel spreadsheets (the format the data set the project was built on, see Sec. 3.4) and in then manipulating the data confirmed that this was a good choice. Note that to read files in the Excel formats the additional `xlrd` package is needed.

Scikit-learn

scikit-learn (sci) aims at providing a unified API for basic Machine Learning; it does not include advanced paradigms such as Reinforcement Learning or graphical models for structured learning. The latter omission was the reason that lead me to select pomegranate as the basis for the implementation of a Bayesian Network. What is included are a stack supervised and unsupervised ML tools to prepare data sets, define machine learning models ranging from spectral analysis-based to ensemble methods to clustering and multiple evaluation and model selection utilities.

NumPy

NumPy (num) is another *de facto* standard package when doing scientific computing with python. Most scientific packages (including pandas, scikit-learn and TensorFlow) depend on NumPy for low-level operations; this is because NumPy is contains fast implementation of n-dimensional array objects together with powerful manipulation functions. In addition to this, NumPy implements linear algebra operations, Fourier Transform and random number generation. The closest parallel to NumPy - as R was for pandas, is MATLAB.

Networkx

NetworkX (net) is another widely-used package that is specialised in the creation and manipulation of graph-structured data. The main use for this package was in building the `MPEGraph` (

inserire referenza una volta scritto

) data structure that I used to build a graph representing the outcome of pseudo-MPE (see 3.3.3) applied to multiple initial evidences (see 3.3.3) .

Anytree

anytree (any) is a python package providing a lightweight implementation of a tree data structure and traversal methods. This package was used as a lightweight substitute for NetworkX when building a chain of reasoning using the pseudo-MPE algorithm (see 3.3.3). As the `MPETree` object is guaranteed to be a single chain, using a simpler resource than NetworkX was advantageous.

3.2.2 Algorithms

d-separation

A naïve implementation according to the definition (presented in Subsec. 3.1.4) to check for d-separation between node X and Y would have a complexity in the order of the number of trails between X and Y; this leads to an exponential in the size of the graph running time. Luckily, Koller et al. [2007b] present what is a linear time algorithm to solve the problem.

The `reachable` procedure takes as input the DAG representing the Bayesian Network $\mathcal{G}$, a source variable X and a set of observed variables Z; on exit it returns the set of variables R that are reachable from X. The procedure runs in two phases, traversing the graph twice: first bottom-up from leaves to roots, then viceversa. During the first run, the algorithm finds all nodes A that are ancestors of the evidence set Z. During the second, the procedure distinguishes the direction it visits each node in order to determine if it is traversable or not. Any node Y that is not in the evidence set is marked as reachable; if it is being visited in direction “up” it can be traversed as the v-structure is a chain. All the parents of Y are marked to be visited in the “up” direction (i.e. from below) and the converse is done for Y’s children. If Y is being visited in the “down” direction its children are again added to be visited in the “down” direction, because Y is traversable. Additionally, if Y happened to be in the set A, found in the first step, then Y’s parents are marked to be visited in the “up” direction then the collider is active and Y can be traversed (a collider is open iff. the central node or any of its descendants are observed).

The full procedure can be found in Koller et al. [2007b]; my implementation shown in 1 follows this pseudocode very closely but the procedure `d-separated`, instead of finding all nodes R that are d-connected to the input X, tests if a given target Y is d-separated from X or not. This gives some extra flexibility in how the function can be used. To find the set S of all nodes d-separated from X I simply iterate the `d-separated` test over all nodes V in the graph representing the BN.

Algorithm 1 d-separation algorithm

$separated\_list = \emptyset$
input: source X, evidence E, nodes V
for target $Y \in V$ **do**
  append $d-separated(X, Y, E)$ to $separated\_list$ {will return true or false}
end for
output: $separated\_list$

MPE

descrivere come ottenere MPE con Gurobi porter, UAI ecc...

3.3 Novel contributions

tengo qui o sposto nel cap 4?

3.3.1 Motivation

The inspiration for the work carried out in this thesis was the paper "Explaining the Most Probable Explanation" by Butz et al. [2018], that has been presented in detail in Sec. 2.4. This paper proposed a system that would build a Bayesian Network modelling a medical data set and, through the interaction with a medical expert, distill an explanation tree. This tree, deemed to represent the solution to the MPE query, could then be used to generate a natural language explanation that the authors claim would lead to the extraction of extra knowledge from the original data set.

The driving hypothesis of the paper was that Bayesian Networks and the solution to the MPE problem would be a powerful tool in helping medical experts gain insights into data. Unfortunately, the paper did not provide any indication that a such a system had ever been built and any validation of the method was left for future work. As of the finalisation of this thesis (July 3, 2019), there has been no work done in substantiating the conclusions by Butz et al. [2018]. As discussed in Chap. 1 and Chap. 2 there is an ever greater need for Machine Learning models and systems to be explainable, especially in mission-critical domains as is healthcare. Current machine learning systems are for the most part opaque and there is confusion regarding even what would constitute a good explanation of their working.

controllare altri autori

For these reason, I believe that building a proof-of-concept system whose logic was inspired by the method presented in the aforementioned paper and validating it with real medical experts would be an important step forwards in the direction of answering the following questions: Can the method of the paper be corroborated? Are Bayesian Networks a good ML model to bootstrap

an explanation from? How good an explanation does the proposed method give, as validated by a domain expert? What improvements are there to be made?

migliorare dopo aver fatto literature review perche avro piu idee

3.3.2 Theory

3.3.3 Algorithms

An important part of my work was developing the algorithms needed to adapt the ideas presented in the paper "Explaining the Most Probable Explanation" by Butz et al. [2018] and "A Progressive Explanation of Inference in 'Hybrid' Bayesian" by **?**. From the former, the construction of the probability tree through a constructive dialogue with the domain expert, the building of counterfactual explanation branches, the automatic generation of the most probable probability tree from initial evidence. From the latter, the generation of an "Inverse explanation". Finally, a simple procedure to output a natural language explanation was developed.

"Pseudo-MPE"

dove critico il fatto che non calcolano vera l'MPE come dicevano? qui o in cap.4?

eseguibile in modalita' esaustiva, d-separat thresholdata

Alternative Explanation Branches

"Pseudo-MPE" from Random Evidence

Inverse Explanation

da fare e trovare nome migliore

Natural Language Explanation

da fare e magari pensare anche a visual ex tions?

Pairwise Correlations

An interesting addition is an algorithm to measure the interrelatedness between pairs of variables.

Plot of network

3.4 Data set

As anticipated in Chap. 1 the work carried out in this thesis had a certain degree of collaboration with a third party, Istituto Cantonale di Patologia.

3.4.1 Istituto Cantonale di Patologia

Istituto Cantonale di Patologia (ICP) is an institute based in Locarno that is specialised in the histological analysis of tissue samples received from private patients, clinics and hospitals, mainly in support of cancer diagnosis. These tests are aimed at identifying the precise profile of the cancer cells and thus inform the clinician on the best treatment for the specific patient.

In addition to its clinical support activities, the Istituto also carries out scientific research aimed at better understanding certain types of cancers at a basic level. In the last ten years, the ICP has published more than 200 peer-reviewed papers and more than 100 works in non-peer reviewed journals and is active at a national and international level.

3.4.2 Motivation

My first contact with the ICP was during a meeting with Dr. Vittoria Martin (Martin [2012]), molecular citogenetist, in date 28/01/2019. The institute had expressed interest in bringing machine learning into their workflow in order to both augment their profiling capabilities for patients and to be able to extract new knowledge from their existing data. This knowledge-extraction may lead towards the confirmation of current scientific theories or may be the first step towards the formulation of novel ones.

chiedere a Vittoria martedÃň cosa si aspett vano/aspettano

My interest in collaborating with the Istituto stemmed from the desire to apply the methods described in Sec. 3.3 to a real-world case. Being the theoretical work being carried out in this thesis an expert-driven MPE approximation, collaboration with the institute has also provided the opportunity to implement a proof of concept using real histological data. The doctors and researchers of the Istituto have been able to validate the model software that I have developed from an Explainable AI point of view. That is to say, they have validated the capacity of the developed software to support clinical decision making and surface clarifying explanations of the data set.

da confermare con Vittoria piÃź avanti

3.4.3 Provided Data Set

The data set I was provided with consists of the histological records of 3218 breast cancer patients. The data set had been pre-processed by collaborators of the ICT with some of the values being compiled by hand. In Tab. 3.1 is a description of the measured variables, together with their clinical meaning. The value distribution of the data set is shown in Tab. 3.2.

The indications from Dr. Martin on how to further preprocess the data are shown in Tab. 3.3. Note that some variable names were simplified.

3.5 Validation Methodology

discutere con Vittoria cosa vogliono esattam come validare

Table 3.1. Data set variables

Variable	Meaning
Codice globale	Unique patient identifier
mut17q21	
loss 17	
età arrotondata	The age of the patient at diagnosis
Lateralità	The affected breast
Situ SUBGROUP MZ	The site code of the tumour
Morfologia SUBGROUP MZ	The morphology classification of the tumour
pT SUBGROUP MZ	Primary tumour in the TNM classification for breast cancer
pN SUBGROUP MZ	Pathologic in the TNM classification for breast cancer
M 8.2.96	Distant metastasis in the TNM classification for breast cancer
Differenziazione	
Recettori estrogeni percento 1.1.2003	
Recettori progestinici percento 1.1.2003	
c erbB 2 cod percento 1.1.2003	
Ki67 cod percento	
FISHRatio	

Table 3.2. Data set distribution

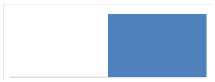
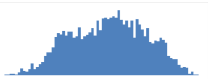
Variable	Unique values	Distribution
mut17q21	2	
loss 17	3	
eta arrotondata	74	
ateralità	3	
situ	5	
morfologia	5	
pT	23	
pN	6	
M	3	
differenziazione	5	
recettori estrogeni	40	
recettori progestinici	40	
c erbB 2	4	
Ki67	52	
FISH	5	

Table 3.3. Data set preprocessing

Variable	Action
Codice globale	Remove variable
mut17q21	Remove blanks
loss 17	Remove blanks
eta arrotondata	Bin into “< 40”, “$40-50$”, “≥ 50”
lateralita	Remove blanks and “sconosciuta”
situ	Remove blanks
morfologia	Remove blanks and “unuseful” if performance on classification is subpar
pT	Remove blanks and “unuseful”
pN	Remove blanks and bin into “0” and “$\neq 0$”
M	Remove blanks
differenziazione	Remove blanks and “Sconosciuto o non applicabile”
recettori estrogeni	Remove blanks and bin into “negativo” if ≤ 10, “debolmente positivo” if ≤ 50, “fortemente positivo” if > 50
recettori progestinici	Remove blanks and bin into “negativo” if ≤ 10, “debolmente positivo” if ≤ 50, “fortemente positivo” if > 50
c erbB 2	Remove blanks
ki67	Remove blanks and bin into “<14”, “14-20”, “20-30”, “>30”
FISH	Remove blanks

Chapter 4

Results

Chapter 5

Conclusions

5.1 Critique of MPE paper

Glossary

Bibliography

anytree. URL https://anytree.readthedocs.io/en/latest/.

Graphviz. URL https://www.graphviz.org.

Gurobi. URL https://www.gurobi.com.

NetworkX. URL https://networkx.github.io.

NumPy. URL http://numpy.org.

Pandas. URL https://pandas.pydata.org/about.html.

scikit-learn. URL http://scikit-learn.github.io/stable.

Elements of, pages 19–21. 1967. ISBN 9780471241959. URL http://staff.ustc.edu.cn/{~}cgong821/Wiley.Interscience.Elements.of.Information.Theory.Jul.2006.eBook-DDU.pdf.

Raphaela Butz, Arjen Hommersom, and Marko van Eekelen. Explaining the Most Probable Explanation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11142 LNAI, pages 50–63. 2018. ISBN 9783030004606. doi: 10.1007/978-3-030-00461-3_4. URL http://link.springer.com/10.1007/978-3-030-00461-3{_}4.

Istituto Cantonale di Patologia. Istituto Cantonale di Patologia. URL https://www4.ti.ch/dss/dsp/icp/chi-siamo/presentazione/.

Daphne Koller, Nir Friedman, Sašo Džeroski, Charles Sutton, Andrew McCallum, Avi Pfeffer, Pieter Abbeel, Ming-Fai Wong, David Heckerman, Chris Meek, et al. *Introduction to statistical relational learning*, pages 22–23. MIT press, 2007a.

Daphne Koller, Nir Friedman, Sašo Džeroski, Charles Sutton, Andrew McCallum, Avi Pfeffer, Pieter Abbeel, Ming-Fai Wong, David Heckerman, Chris Meek, et al. *Introduction to statistical relational learning*, pages 74–76. MIT press, 2007b.

Vittoria Martin. Curriculum vitae. 5:13–15, 2012. URL https://m4.ti.ch/fileadmin/DSS/DSP/ICP/pdf/Collaboratori/VM{_}CV.pdf.

Sharon Use Normand and David Tritchler. Parameter updating in a bayes network. *Journal of the American Statistical Association*, 87(420):1109–1115, 1992. ISSN 1537274X. doi: 10.1080/01621459.1992.10476266.

Judea Pearl and Rina Dechter. Identifying Independencies in Causal Graphs with Feedback 2 BAYESIAN NETWORKS VS . *Intelligence*, (1):43, 1988. URL http://www.cs.ucla.edu/pub/stat{_}ser/R243.pdf.

PyGraphviz developer team. Pygraphviz. URL https://pygraphviz.github.io.

Thomas D. Schneider. Molecular Information Theory Primer. (301), 2005. URL http://users.fred.net/tds/lab/paper/primer/primer.pdf.

Jacob Schreiber. Pomegranate, a. URL https://pomegranate.readthedocs.io/en/latest/.

Jacob Schreiber. Bayesian Network Structure Learning tutorial, b. URL https://github.com/jmschrei/pomegranate/blob/master/tutorials/B_Model_Tutorial_4b_Bayesian_Network_Structure_Learning.ipynb.

Jacob Schreiber. Pomegranate: fast and flexible probabilistic modeling in python. 18:1–6, 2017. URL http://arxiv.org/abs/1711.00137.

Jacob Schreiber and William Noble. Finding the optimal bayesian network given a constraint graph. 2017. doi: 10.7287/peerj-cs.122v0.2/reviews/2. URL https://peerj.com/articles/cs-122/.

C. E. Shannon, W. Weaver, and R. E. Blahut. The mathematical theory of communication (TODO check jahr 48/49). *Urbana: University of Illinois press*, 117(April 1928):379–423, 1949. ISSN 07246811. doi: 10.2307/3611062. URL http://math.harvard.edu/{~}ctm/home/text/others/shannon/entropy/entropy.pdf.

Solomon Eyal Shimony. Finding MAPs for belief networks is NP-hard. *Artificial Intelligence*, 68 (2):399–410, 1994. ISSN 00043702. doi: 10.1016/0004-3702(94)90072-8. URL http://www.cs.columbia.edu/{~}adrian/candp/Shi94-FindingMAPbeliefNetworksNPhard.pdf.

Chandler Stolp, Shirley Dowdy, and Stanley Wearden. *Statistics for Research*, volume 3, page 637. 2006. ISBN 047126735X. doi: 10.2307/3324586.