UNIVERSIDAD
DE GRANADA

STATISTICAL MODELS WITH VARIATIONAL METHODS

LUIS ANTONIO ORTEGA ANDRÉS

End-of-Degree Project

Double Degree in Computer Science and Mathematics

Tutor

Serafín Moral Callejón

FACULTY OF SCIENCE

H.T.S. OF COMPUTER ENGINEER AND TELECOMMUNICATIONS

Granada, Friday 21st February, 2020

ABSTRACT

Some introduction about how important Variational methods are nowadays and what this project is about.

CONTENTS

Part I

BASIC CONCEPTS

In this chapter we will introduce the underlying concepts of probability and graph theory that we will need.

1

PROBABILITY

All our theory will be made under the assumption that there is a *referential set* Ω, set of all possible outcomes of an experiment. Any subset of Ω will be called *event*.

Definition 1. Let $\mathcal{P}(\Omega)$ be the power set of Ω. Then, $\mathcal{F} \subset \mathcal{P}(\Omega)$ is called a *σ-algebra* if it satisfies:

- $\Omega \in \mathcal{F}$.
- $\mathcal{F}$ is closed under complementation.
- $\mathcal{F}$ is closed under countable unions.

From these properties it follows that $\emptyset \in \mathcal{F}$ and that $\mathcal{F}$ is closed under countable intersections.

The tuple $(\Omega, \mathcal{F})$ is called a *measurable space*.

Definition 2. A *probability* P over $(\Omega, \mathcal{F})$ is a mapping $P : \mathcal{F} \to [0, 1]$ which satisfies

- $P(\alpha) \geq 0 \ \ \forall \alpha \in \mathcal{F}$.
- $P(\Omega) = 1$.
- P is countably additive, that is, if $\{\alpha_i\}_{i \in \mathbb{N}} \subset \mathcal{F}$, is a countable collection of pairwise disjoint sets, then

$$P\Big(\bigcup_{i \in \mathbb{N}} \alpha_i\Big) = \sum_{i \in \mathbb{N}} P(\alpha_i).$$

- For any sequence $\{\alpha_n\}_{n \in \mathbb{N}}$ such that $\alpha_i \subset \alpha_{i+1} \ \forall i \in \mathbb{N}$ and $\alpha_n \xrightarrow{n \to \infty} \Omega$, then

$$P(\alpha_n) \xrightarrow{n \to \infty} P(\Omega) = 1.$$

The first condition guarantees non negativity. The second one states that the *trivial event* has the maximal possible probability of 1. The third condition implies that given two mutually disjoint events, the probability of either one of them occurring is equal to the sum of the probabilities of each one. The last condition sets an upper semi-continuity.

From these conditions it follows that $P(\emptyset) = 0$ and $P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$.

The triple $(\Omega, \mathcal{F}, P)$ is called a *probability space*.

Definition 3. A function $f : \Omega_1 \to \Omega_2$ between two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ is said to be *measurable* if $f^{-1}(\alpha) \in \mathcal{F}_1$ for every $\alpha \in \mathcal{F}_2$.

Definition 4. A *random variable* is a measurable function $X : \Omega \to E$ from a probability space $(\Omega, \mathcal{F}, P)$ to a measurable space E.

The probability of X taking a value on a measurable set $S \subset E$ is written as

$$P_X(S) = P(X \in S) = P(\{a \in \Omega \mid X(a) \in S\}).$$

Where P_X is called the *probability distribution* of the random variable X.

We will adopt the following notation from now on: random variables will be denoted with an upper case letter like X and a set of variables with a bold symbol like $\boldsymbol{X}$ The meaning of $P(state)$ will be clear without a reference to the variable. Otherwise $P(X = state)$ will be used. We will denote by $P(x)$ the probability of X taking a specific value.

Also $P(x \text{ or } y) = P(x \cup y)$ and $P(x, y) = P(x \cap y)$.

Definition 5. The *cumulative distribution function* of a random variable X is the function given by:

$$F_X(x) = P(X \leq x)$$

where the right-hand side represents the probability of the random variable taking value below or equal to x.

Definition 6. When the image of a random variable X is countable, the random variable it is called a *discrete random variable*, its *probability mass function* p gives the probability of it being equal to some value.

$$p(x) = P(X = x)$$

If the image is uncountable then X is called a *continuous random variable* and it's *probability density function* f is a non-negative Lebesgue-integrable such that

$$F_X(x) = P(X \leq x) = \int_{-\infty}^{x} f(u)du$$

We will define some concepts regarding a joint probability $P(x, y)$, that is to say, the probability of both random variables X and Y.

Definition 7. A *marginal distribution* $P(x)$ of the joint distribution is the distribution of a single variable given by

$$P(x) = \sum_y P(x, y) \qquad\qquad P(x) = \int_y P(x, y)$$

We can understand this as the probability of an event irrespective of the outcome of the other variable.

Definition 8. The *conditional probability* of X given Y is defined as

$$P(x|y) = \frac{P(x, y)}{P(y)}$$

If $P(y) = 0$ then it is not defined.

With this definition the conditional probability is the probability of one event occurring in the presence of a second event.

Theorem 1. ***(Bayes' theorem)***. *Given two random variables X, Y, then*

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

Example 1. Consider a study where the relation of a disease d and an habit h is being investigated. Suppose that $P(d) = 10^{-5}$, $P(h) = 0.5$ and $P(h|d) = 0.9$. What is the probability that a person with habit h will have the disease d?

$$P(d|h) = \frac{P(d, h)}{P(h)} = \frac{P(h|d)P(d)}{P(h)} = \frac{0.9 \times 10^{-5}}{0.5} = 1.8 \times 10^{-5}$$

If we set the probability of having habit h to a much lower value as $P(h) = 0.001$, then the above calculation gives approximately $1/100$. Intuitively, a smaller number of people have the habit and most of them have the desease. This means that the relation between having the desease and the habit is stronger now compared with the case where more people had the habit.

Now suppose we have some observed data $\mathcal{D}$ and we want to learn about a set of parameters θ. Using Bayes' rule we got that

$$P(\theta|\mathcal{D}) = \frac{P(\mathcal{D}|\theta)P(\theta)}{P(\mathcal{D})} = \frac{P(\mathcal{D}|\theta)P(\theta)}{\int_\theta P(\mathcal{D}|\theta)P(\theta)}$$

This shows how from a *generative model* $P(\mathcal{D}|\theta)$ of the dataset and a *prior* belief $P(\theta)$, we can infer the *posterior* distribution $P(\theta|\mathcal{D})$.

Definition 9. We say that two random variables X and Y are *independent* if knowing one of them doesn't give any extra information about the other. Mathematically,

$$P(x, y) = P(x)P(y)$$

From this it follows that if X and Y are independent, then $P(x|y) = P(x)$.

Definition 10. Let X, Y and Z be three random variables, then X and Y are *conditionally independent* given Z if and only if

$$P(x, y|z) = P(x|z)P(y|z)$$

in that case we will denote $X \perp\!\!\!\perp Y \mid Z$. If X and Y aren't conditionally independent, they are *conditionally dependent* $X \not\perp\!\!\!\perp Y \mid Z$

Both independence definitions can be made over sets of variables $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ in a straight forward way.

2

GRAPH THEORY

Definition 11. A *graph* $G = (V, E)$ is a set of vertices or nodes V and edges $E \subset V \times V$ between them. If V is a set of ordered pairs then the graph is called a *directed graph*, otherwise if V is a set of unordered pairs it is called an *undirected graph*.

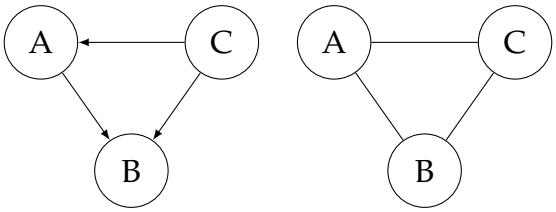


Figure 1: Example of directed and undirected graph, respectively.

Definition 12. In a directed graph $G = (V, E)$, a *directed path* $A \to B$ is a sequence of vertices $A = A_0, A_1, \ldots, A_{n-1}, A_n = B$ where $(A_i, A_{i+1}) \in E \ \forall i \in 0, \ldots, n-1$.

If G is a undirected graph, $A \to B$ is an *undirected path* if $\{A_i, A_{i+1}\} \in E \ \forall i \in 0, \ldots, n-1$

Definition 13. Let A, B be two vertices of a directed graph G. If $A \to B$ is a directed path and $B \not\to A$ (meaning there isn't a directed path from B to A), then A is called an *ancestor* of B and B is called a *descendant* of A.

For example, in the figure 1, C is an ancestor of B.

Definition 14. A *directed acyclic graph (DAG)* is a directed graph such that no directed path between any two nodes revisits a vertex.

As we can see in the figure 2, $A \to B \to C \to A \to B$ is a path from A to B that revisits A.

Now where are going to define some relations between nodes in a DAG.

Definition 15. The *parents* of a node A is the set of nodes B such that there is a directed edge from B to A. The same applies for the *children* of a node.

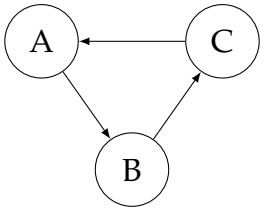


Figure 2: Example of graph which isn't a DAG.

The *Markov blanket* of a node is composed by the node itself, its children, its parents and the parents of its children.

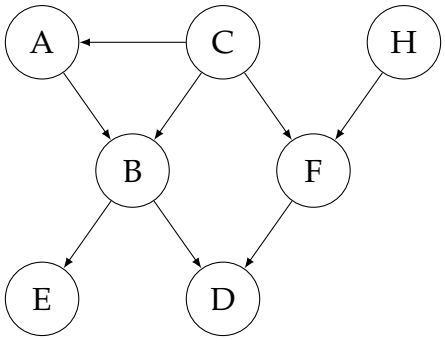


Figure 3: Directed acyclic graph

Definition 16. In a graph, the *neighbors* of a node are those directly connected to it.

We can use figure 4 to reflect on these definitions. The parents of B are $pa(B) = \{A, C\}$ and its children are $ch(B) = \{E, D\}$. Taking this into account, its neighbors are $ne(B) = \{A, C, E, D\}$ and its Markov blanket is $\{A, B, C, D, E, F\}$.

Part II

GRAPHICAL MODELS

A *graphical model* is a statistical model for which a graph expresses the conditional dependence structure between random variables.

Commonly, they provide a graph-based representation for encoding a multi-dimensional distribution representing a set of independences that hold in the specific distribution. The most commonly used are *Bayesian networks* and *Markov random fields*, which differ in the set of independences they can encode and the factorization of the distribution that they include.

3

BAYESIAN NETWORKS

Consider we have N variables with the corresponding distribution $P(x_1, \ldots, x_N)$. Let $\mathcal{E}$ be a set of indexes such as $\texttt{evidence} = \{X_e = x_e \mid e \in \mathcal{E}\}$. Inference could be made by brute force:

$$P(X_i = x_i \mid \texttt{evidence}) = \frac{\int_{j \notin \mathcal{E}, j \neq i} P(\texttt{evidence}, x_j, X_i = x_i)}{\int_{j \notin \mathcal{E}} P(\texttt{evidence}, x_j)}$$

The notation when using discrete variables is analogous replacing integration with summations.

Lets suppose all these variables are binary, this calculation will require $O(2^{N-\#\mathcal{E}})$ operations. Also, all entries of a table $P(x_1, \ldots, x_N)$ take $O(2^N)$ space.

This is unpractical when taking into account millions of variables. The underlying idea of belief networks is to specify which variables are independent of others, factoring the joint probability distribution.

Definition 17. Let $G = (V, E)$ be a graph where $V = \{X_1, \ldots, X_n\}$ is a set of random variables. We say that the joint probability $P(x_1, \ldots, x_n)$ *factorizes* according to G if and only if

$$P(x_1, \ldots, x_N) = \prod_{i=1}^{N} P(x_i | pa(x_i))$$

Definition 18. A *belief network or Bayesian network* is a pair $\mathcal{B} = (G, P)$ where P factorizes over G. It is a probabilistic graphical model that represents conditional dependencies of a set of variables $X_1, \ldots, X_n$.

Any probability distribution can be written as a Bayesian Network, even though it may end up been a fully-connected DAG. To set the specification of the Belief Network, we need to define all elements of the probability tables $P(x_i|pa(x_i))$. When the number of variables is large, this is still intractable so the tables are generally parameterized is a low dimensional manner.

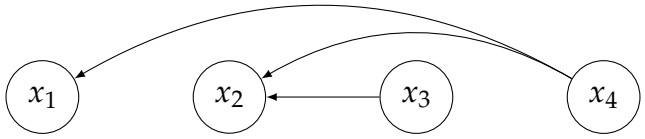


Figure 4: Bayesian Network factorizing $P(x_1, x_2, x_3, x_4) = P(x_1|x_4)P(x_2|x_3, x_4)P(x_3)P(x_4)$

Bayesian Networks are good for encoding conditional independence over the variables, but aren't for encoding dependence. For example, with the following network $P(x, y) = P(y|x)P(x)$ represented as $x \to y$ in a DAG. It may appear to encode dependence between both variables but the conditional $P(y|x)$ could happen to equal $P(y)$, giving $P(x, y) = P(x)P(y)$.

How could we check if two variables are conditionally independent given a Bayesian Network? For example in figure 4, $X_1 \perp\!\!\!\perp X_2 \mid X_4$ as[1]:

$$\begin{aligned} P(x_2|x_4) &= \frac{1}{P(x_4)} \int_{x_1,x_3} P(x_1, x_2, x_3, x_4) = \frac{1}{P(x_4)} \int_{x_1,x_3} P(x_1|x_4)P(x_2|x_3, x_4)P(x_3)P(x_4) \\ &= \int_{x_3} P(x_2|x_3, x_4)P(x_3) \end{aligned}$$

$$\begin{aligned} P(x_1, x_2|x_4) &= \frac{1}{P(x_4)} \int_{x_3} P(x_1, x_2, x_3, x_4) = \frac{1}{P(x_4)} \int_{x_3} P(x_1|x_4)P(x_2|x_3, x_4)P(x_3)P(x_4) \\ &= P(x_1|x_4) \int_{x_3} P(x_2|x_3, x_4)P(x_3) = P(x_1|x_4)P(x_2|x_4) \end{aligned}$$

Now we are going to define two central concepts to determine conditional independence in any Bayesian Network, these are *d-connection* and *d-separation*.

Definition 19. Let G be a DAG, and $x \in V$ be a vertice such that $\forall y \in ne(X), (y, x) \in E$, then x is called a *collider*.

For example in figure 4, x_2 is a collider.

Definition 20. Let G be a DAG where $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ are disjoint sets of vertices. We say that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *d-connected* by $\boldsymbol{Z}$ if and only if there exists an undirected path U from any vertex in $\boldsymbol{X}$ to any vertex in $\boldsymbol{Y}$ such that:

- For any collider C, itself or any it's descendants is in $\boldsymbol{Z}$
- No non-collider on U is on $\boldsymbol{Z}$

Definition 21. Let G be a DAG where $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ are disjoint sets of vertices. $\boldsymbol{X}$ and $\boldsymbol{Y}$ are *d-separated* by $\boldsymbol{Z}$ if and only if they are not d-connected by $\boldsymbol{Z}$ in G

For example, in figure 5 d d-separates a and c (b is a collider in the path that isn't in $\{d\}$), and $\{d, e\}$ d-connect them.

1 Continuous variable notation is used

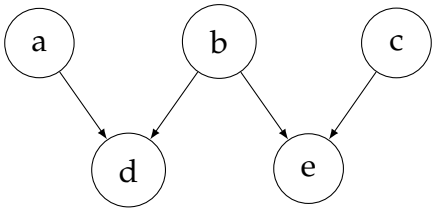


Figure 5: D-separation example

Theorem 2 (Verma and Pearl, 1988, Geiger et al., 1990 [2]). *Let G be a DAG where $\boldsymbol{X}, \boldsymbol{Y}$ and $\boldsymbol{Z}$ are disjoint sets of vertices. If $\boldsymbol{X}$ and $\boldsymbol{Y}$ are d-separated by $\boldsymbol{Z}$, then they are independent conditional on $\boldsymbol{Z}$ in all probability distributions that G can represent.*

The Bayes Ball algorithm [4] provides a linear time complexity algorithm that computes conditional independent using this theorem.

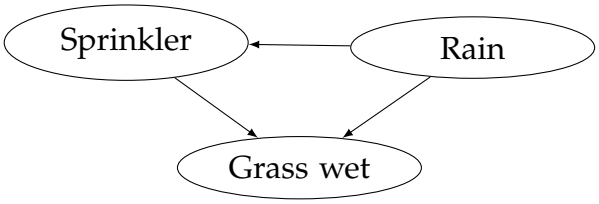
Example 2. In this example we are modeling three discrete random variables: Sprinkler (S), Rain (R) and Grass wet (G).

The joint probability function is:

$$P(s, r, g) = P(s|r)P(g|s, r)P(r)$$

The following DAG illustrates the Bayesian Network among with the probability tables we are using.

	Sprinkler	
Rain	T	F
F	0.4	0.6
T	0.01	0.99



Rain	
T	F
0.2	0.8

		Grass wet	
Sprinkler	Rain	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

This model can answer questions about the presence of a cause given the presence of an effect. For example, What is the probability that it has being raining given the grass is wet?

$$P(R = T|G = T) = \frac{P(G = T, R = T)}{P(G = T)} = \frac{\sum_s P(G = T, R = T, s)}{\sum_{r,s} P(G = T, r, s)}$$

Using the expression of the joint probability among with the tables we can compute every term. For example:

$$\begin{aligned} P(G = T, R = T, S = T) &= P(S = T|R = T)P(G = T|R = T, S = T)P(R = T) \\ &= 0.01 * 0.99 * 0.2 = 0.00198 \end{aligned}$$

Cites so the references appear (testing) [3, 1, 5]

BIBLIOGRAPHY

[1] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

[2] Rina Dechter Judea Pearl. Identifying independences in casual graphs with feedback.

[3] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models, Principles and Techniques*. The MIT Press.

[4] Ross D. Shachter. Bayes-ball: The rational pastime.

[5] Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families and Variational Inference*. Now Publishers Inc.