



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, Sunday 16th February, 2020

ABSTRACT

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

CONTENTS

I	BASIC CONCEPTS	4
1	PROBABILITY	5
2	GRAPH THEORY	9
II	GRAPHICAL MODELS	11
3	BAYESIAN NETWORKS	12

Part I

BASIC CONCEPTS

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

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 (Ω, \mathcal{F}) is called a *measurable space*.

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

- $P(\alpha) \geq 0 \quad \forall \alpha \in \mathcal{F}$.
- $P(\Omega) = 1$.
- 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

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

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

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

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 (Ω, \mathcal{F}, P) is called a *probability space*.

Definition 3. A function $f : \Omega_1 \rightarrow \Omega_2$ between two measurable spaces 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 $X : \Omega \rightarrow E$ from a probability space (Ω, \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

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

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 \mathbf{X} . The meaning of $P(\text{state})$ will be clear without a reference to the variable. Otherwise $P(X = \text{state})$ 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 $P(x, y) = P(x \cap y)$.

We will define some concepts regarding a joint distribution $P(x, y)$, that is to say, the probability of both random variables x and 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 is called *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 its *probability density function* f is a non-negative Lebesgue-integrable such that

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

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

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

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 disease. This means that the relation between having the disease 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 \mathcal{X}, \mathcal{Y} and \mathcal{Z} .

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*.

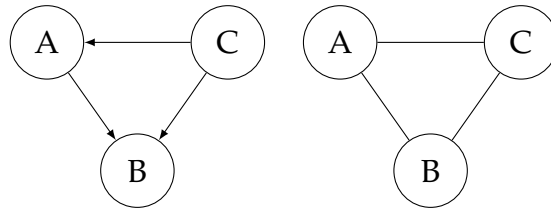


Figure 1: Example of directed and undirected graph, respectively.

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

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

Definition 13. Let A, B be two vertices of a directed graph G . If $A \rightarrow B$ is a directed path and $B \not\rightarrow A$ (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, $A \rightarrow B \rightarrow C \rightarrow A \rightarrow B$ is a path from A to B that revisits A .

Now we 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.

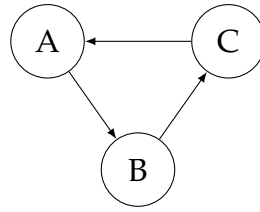


Figure 2: Example of graph which isn't a DAG.

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

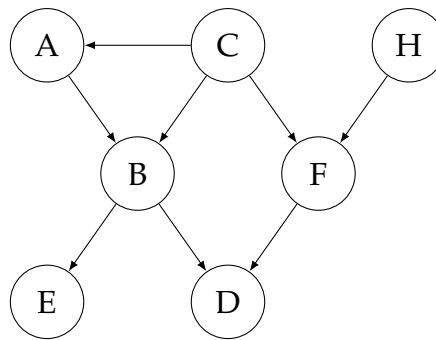


Figure 3: Directed acyclic graph

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.

BAYESIAN NETWORKS

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

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

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 $P(x_1, \dots, x_N)$ 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. A *belief network* or *Bayesian network* is a probabilistic graphical model that represents conditional dependencies of a set of variables X_1, \dots, X_n via a directed acyclic graph following

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

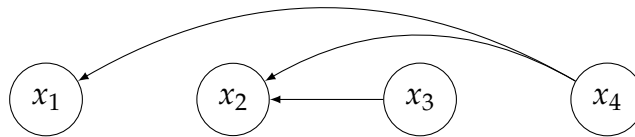


Figure 4: $P(x_1, x_2, x_3, x_4) = P(x_1|x_4)P(x_2|x_3, x_4)P(x_3)P(x_4)$

Cites so the references appear (testing) [2, 1, 3]

BIBLIOGRAPHY

- [1] David Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press.
- [2] Daphne Koller and Nir Friedman. *Probabilistic Graphical Models, Principles and Techniques*. The MIT Press.
- [3] Martin J. Wainwright and Michael I. Jordan. *Graphical Models, Exponential Families and Variational Inference*. Now Publishers Inc.