Statistical Models with Variational Methods

February 14, 2020

LUIS ANTONIO ORTEGA ANDRÉS

End-of-degree Project
Granada, Spain

CONTENTS

1 INTRODUCTION

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

2 PARTE DE MATEMÁTICAS (ABSTRACT)IGNORE

3 BASIC CONCEPTS

3.1 *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 $P : \mathcal{F} \to \mathbb{R}$ which satisfies

- $P(\alpha) \geq 0 \ \ \forall \alpha \in \mathcal{F}$.
- $P(\Omega) = 1$.
- If $\alpha, \beta \in \mathcal{F}$ and $\alpha \cap \beta = \emptyset$, then $P(\alpha \cup \beta) = P(\alpha) + P(\beta)$.

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.

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

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

Note. Proof this or smth.

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 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

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

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 $\boldsymbol{X}$ 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, which means that

$$\int_x f(x) = \int_{dom(X)} f(X = s)\, ds$$

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.

introduce el concepto de variable continua y discreta.
usa p para la función de distribución de una discreta
y f para la densidad de una continua

Definition 5. 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 6. 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.

This formula is also known as *Bayes' rule*. With this definition the conditional probability is the probability of one event occurring in the presence of a second event.

Theorem 1. ***(Bayes' rule)***.

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 $p(\theta|\mathcal{D})$.

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.

Definition 7. We say that events 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 $p(x|y) = p(x)$.

3.2 *Graphical models*

Definition 8. A *graph* $G = (V, E)$ is a set of vertices or nodes V and edges $E \subset V \times V$ between them. These edges may be directed (have arrow in a single direction) or undirected. If all the edges of a graph are directed, it is called a *directed graph*, and if all of them are undirected, it is called an *undirected graph*.

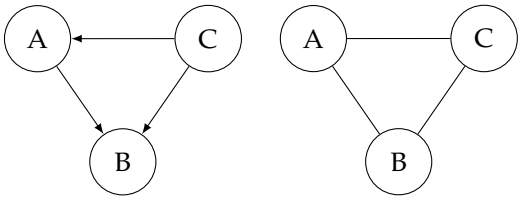


Figure 1: Example of directed and undirected graph, respectively.

Definition 9. A *path* $A \to B$ is a sequence of vertices $A = A_0, A_1, \ldots, A_{n-1}, A_n = B$ where (A_n, A_{n-1}) an edge of the graph. In a directed graph, if the edges follow the sequence, the resulting path is called a *directed path*.

Definition 10. Let A, B be two vertices. If $A \to B$ and $B \not\to A$, 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 11. A *directed acyclic graph (DAG)* is a directed graph such that no directed path between any two nodes revisits a vertex.

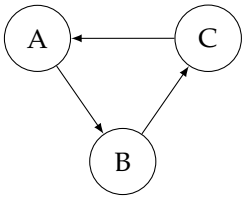


Figure 2: Example of graph which isn't a DAG.

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 12. 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.

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

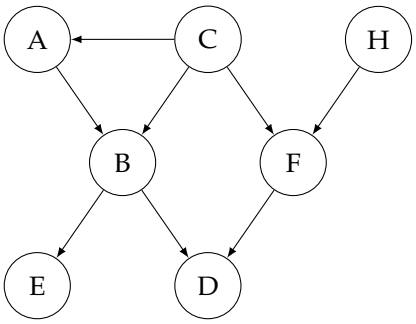


Figure 3: Directed acyclic graph

Definition 13. In a graph, the *neighbors* of a node are those directly connected to it.

We can use figure 3 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 $\{A, B, C, D, E, F\}$.

Definition 14. A *graphical model* is a probabilistic 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.

4 BELIEF 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 $\texttt{evidence} = \{x_e = \times_e \mid e \in \mathcal{E}\}$. Inference could be made by brute force:

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

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 15. A *belief network* is a distribution of the form

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

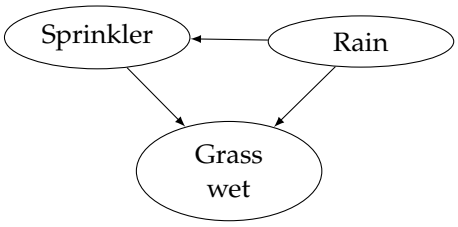
We can write it as a DAG where the i^{th} node corresponds to the factor $p(x_i|pa(x_i))$.

TODO Example

5 GRAPHICAL MODEL TEST WITH TIKZ

This is a test of making a graphical model in latex using Tikz package.

	Sprinkler	
Rain	T	F
F	0.4	0.6
T	0.01	0.99



Sprinkler	
T	F
0.2	0.8

		Grass wet	
Sprinkler	rain	T	F
F	F	0.4	0.6
F	T	0.01	0.99
T	F	0.01	0.99
T	T	0.01	0.99

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

REFERENCES

[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.