Statistical Models with Variational Methods

February 3, 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 BASIC CONCEPTS

2.1 *Probability*

Definition 1. An *event* is a set of outcomes of an experiment to which a probability is assigned. This definition is made over the assumition that there is a *sample space* A, set of all possible outcomes of the experiment.

Definition 2. Let $\mathcal{P}(A)$ be the power set of A, then, $\mathcal{A} \subset \mathcal{P}(A)$ is called a *σ-algebra* if satisfies:

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

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

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

Definition 3. A *probability distribution* p over $(A, \mathcal{A})$ is a mapping $p : \mathcal{A} \to \mathbb{R}$, following:

- $p(\alpha) \geq 0 \ \forall \alpha \in \mathcal{A}$.
- $p(A) = 1$
- If α, $\beta \in S$ and $\alpha \cap \beta = \emptyset$, then $p(\alpha \cup \beta) = p(\alpha) + p(\beta)$.

The first condition implies non negativity. The second one states that the *trivial event* has the maximal possible probability of 1. The last condition states that given two mutually disjoint events, the probability of one of them is equal to the sum of the probabilities of each one.

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

Definition 4. A function $f : (A_1, \mathcal{A}_1) \to (A_2, \mathcal{A}_2)$ between two measurable spaces is said to be *measurable* if for every $\alpha \in \mathcal{A}_2$, it satisfies $f^{-1}(\alpha) \in \mathcal{A}_1$.

Definition 5. A *random variable* is a measurable function $X : A \to E$ from a set of possible outcomes A and a measurable space E.

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

$$p(X \in S) = p(\{\alpha \in \Omega \mid X(\alpha) \in S\})$$

We will adopt the following notation from now on, random variables will be denoted with lower case x and a set of variables with a calligraphic symbol like $\mathcal{V}$. The meaning of $p(state)$ will be clear without a reference to the variable. Otherwise $p(x = state)$ will be used. We will denote $p(x)$ the probability of x taking a specific value, this 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 from a given joint distribution $p(x, y)$, this is, the probability of both random variables.

Definition 6. 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 another variable.

Definition 7. 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.

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. Consider $p(d) = 10^{-5}$, $p(h) = 0.5$ and $p(h|d) = 0.9$. What is the probability that a person with habit h will have 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.

Definition 8. 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 follows that, if x and y are independent, then $p(x|y) = p(x)$.

2.2 *Graphical models*

Definition 9. A *graph* $G = (V, E)$ is a set of vertices or nodes V and edges $E \subset V \times V$ between the vertices. This 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*, if all of them are undirected, is called a *undirected graph*.

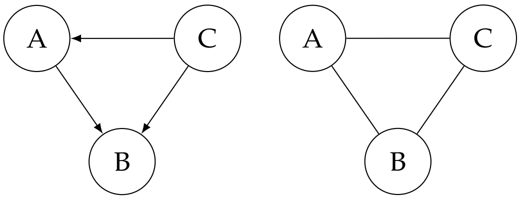


Figure 1: Directed and undirected graph respectively

Definition 10. A *path* $A \to B$ is a sequence of vertices $A_0 = A, A_1, \dots, A_{n-1}, A_n = B$ where $(A_n, A_{n=1})$ an edge of the graph. In a directed graph, if the edges follow the sequence, if is called a *directed path*.

Definition 11. 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 12. A *directed acyclic graph (DAG)* is a directed graph such that no directed path from any node to another revisits a vertex.

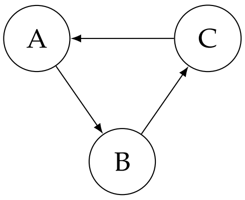


Figure 2: Example of graph which isn't a DAG

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 13. 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 follows for the *children* of a node.

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

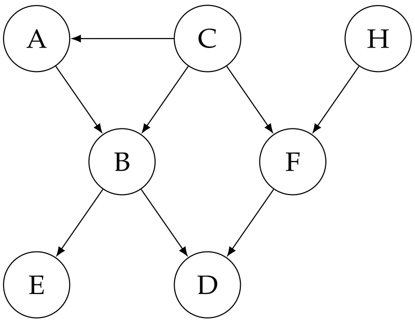


Figure 3: Directed acyclic graph

Definition 14. In a graph, the *neighbors* of a node are those directly connected to it.

We can use figure 3 to reflect this definitions. The parents of B are $\{A, C\}$ and it's children are $\{E, D\}$. With this, it's neighbors are $ne(B) = \{A, C, E, D\}$ and it's Markov blanket is $\{A, B, C, D, E, F\}$.

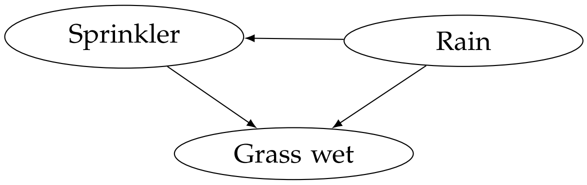
Definition 15. A *graphical model* is a probabilistic model for which a graph express the conditional dependence structure between random variables.

Commonly, they use 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* that differ in the set of independences they can encode and the factorization of the distribution that they include.

3 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.