UNIVERSIDAD
DE GRANADA

STATISTICAL MODELS WITH VARIATIONAL METHODS

LUIS ANTONIO ORTEGA ANDRÉS

End-of-Degree Project

Computer Science and Mathematics

Tutor

Serafín Moral Callejón

FACULTY OF SCIENCE

H.T.S. OF COMPUTER ENGINEER AND TELECOMMUNICATIONS

Granada, Monday 30th March, 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 an *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

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

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 a set of pairwise 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

- $P(\emptyset) = 0$
- $P(\alpha \cup \beta) = P(\alpha) + P(\beta) - P(\alpha \cap \beta)$

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

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

We could make a question like "How likely is that the value of X equals a?". This is the same as asking for the probability of the set $\{\omega \in \Omega \mid X(w) = a\}$.

We define the *probability distribution* of an experiment as a function that provides the probability of occurrence of the different events in Ω. With this, if X is used to denote the outcome of the experiment, the probability distribution of X would be a function that gives the probability of every state of X. With this approach, we use the random variable X to "push-forward" the probability P on Ω to a probability P_X in the measurable space E. Typically, this measurable space is the set of real numbers $\mathbb{R}$ along with Borel's σ-algebra $\mathcal{B}$.

The probability distribution of a random variable is typically described by the *probability mass function* or the *probability density function*, depending on whether X is discrete or not. We will define this concepts after setting the notation that is going to be used, that is: 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. Using a lower case letter like $P(x)$ will denote the probability of the corresponding upper case variable X taking a specific value.

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 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 *multivariate random variable* or *random vector* is a column vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ whose components are real-valued random variables on the same probability space.

Note that we use the same symbol $\boldsymbol{X}$ for random vectors and sets of variables, but the meaning will be clear within the context.

Now we are going to define some concepts over events $\alpha, \beta \in \mathcal{F}$, then we will set those over random variables and their distributions.

- The *joint probability* $P(\alpha, \beta)$ the probability of both events occurring.
- The *marginal probability* $P(\alpha)$ is the probability of occurring α irrespective of the other event.
- The *conditional probability* $P(\alpha|\beta)$ is the probability of occurring α given β.

Let's illustrate this with a simple example.

Example 1. Suppose Ω is the set of outcomes of rolling a dice, that is $\Omega = \{1, 2, 3, 4, 5, 6\}$. We define $\alpha = \{2, 4, 6\}$ and $\beta = \{4, 5\}$. Then the joint probability is $P(\{4\}) = 1/6$, the marginal probability $P(\alpha) = 3/6$ is easy to calculate as we can do it straight-forward, but if we couldn't, we should use the joint probability. We said that the marginal $P(\alpha)$ is the probability of α irrespective of β. Then, we can calculate it as the probability of occurring both $P(\alpha, \beta)$ plus the probability of occurring α and not β $P(\alpha, \{1, 2, 3, 6\})$.

The conditional probability $P(\alpha|\beta)$ forces the outcome to be either 4 or 5, and the only option for it to be in α is that it is 4, then $P(\alpha|\beta) = 0.5$.

Definition 8. Let $\boldsymbol{X} = \{X_1, X_2, \ldots, X_n\}$ be a set of random variables, the *joint probability distribution* for $\boldsymbol{X}$ is function that gives the probability of each random variable X_i falling in a particular range or discrete set of values for that variable. It is called a *multivariate distribution*.

When using only two random variables, then is called a *bivariate distribution*.

This distribution can be expressed in terms of a joint cumulative distribution function

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)^1$$

or using a probability density function (all variables must be continuous) or a probability mass function (all variables must be discrete).

Definition 9. The *marginal distribution* of a subset of random variables is the probability distribution of the variables contained in that subset.

Let X, Y be two random variables, it follows that[2]

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

1 Where $\boldsymbol{x} = (x_1, \ldots, x_n)$

2 Using both continuous and discrete notation

Definition 10. The *conditional probability* of a subset of random variables is the probability distribution of them when the rest of the variables are known to be a particular value.

Theorem 1. ***(Bayes' theorem)***. *Let α, β be two events of an experiment, given that $P(\beta) \neq 0$. Then*

$$P(\alpha|\beta) = \frac{P(\beta|\alpha)P(\alpha)}{P(\beta)}$$

We can extend this theorem to a pair of random variables X, Y as

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

Example 2. 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 11. 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 12. 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 are not 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.

Definition 13. We say that a set of n random variables $\{X_1, \ldots, X_n\}$ defined to assume values in $I \subset \mathbb{R}$ are *independent and identically distributed (i.i.d)* if and only if they are independent

$$F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = F_{X_1}(x_1) \ldots F_{X_n}(x_n) \ \forall x_1, \ldots, x_n \in I$$

and are identically distributed

$$F_{X_1}(x_1) = F_{X_k}(x_k) \ \forall k \in \{2, \ldots, n\} \text{ and } \forall x \in I$$

2

DISTRIBUTIONS

In this section we will summarize some concepts concerning probability distributions among with some of the most used ones.

From now on, let X be a random variable and P its probability distribution.

Definition 14. The *mode* X_* of the probability distribution P is the state of X where the distribution takes it's highest value

$$X_* = \arg\max_x P(x)$$

A distribution could have more than one mode, in this case we say it is *multi-modal*.

Definition 15. The notation $\mathbb{E}[X]$ is used to denote the *average* or *expectation* of the values the variable takes respect to its distribution. If X is non-negative, it is defined as

$$\mathbb{E}[X] = \int_\Omega X(\omega) dP(\omega)^1$$

For a general variable X it is defined as $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$

Suppose now that X is a real-valued random variable, in case it is also continuous

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x f(x) dx$$

and if it is discrete, let x_i be the values X can take

$$\mathbb{E}[X] = \sum_{i=1}^{+\infty} x_i p(x_i) dx$$

Let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function, then $g \circ X$ is another random variable and we can talk about $\mathbb{E}[g(X)]$, so in case X is continuous, we have that

$$\mathbb{E}[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x) dx$$

1 P is a measure over Ω

Definition 16. We define the k^{th} moment of a distribution as the average of X^k over the distribution

$$\mu_k = \mathbb{E}[X^k]$$

For $k = 1$ it is typically denoted as μ.

Definition 17. The *variance* of a distribution is defined as

$$Var(X) = \sigma^2 = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X]^2 - \mathbb{E}[X^2]$$

The square root of the variance σ is called the *standard deviation*.

When using a multivariate distribution $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ we can talk about the *covariance matrix* $\boldsymbol{\Sigma}$ whose elements are

$$\begin{aligned}\boldsymbol{\Sigma}_{ij} &= \mathbb{E}[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])] = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)] \\ &= \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]\end{aligned}$$

The following result will be helpful later on

Proposition 1. *Let $\{X_1, \ldots, X_n\}$ be a set of random variables and P their joint probability distribution. It follows that the expectation of a function f on a subset of the variables $\mathcal{X}_0 \subset \mathcal{X} = \{X_1, \ldots, X_n\}$, verifies*

$$\mathbb{E}[f(\mathcal{X}_0)]_{P(\mathcal{X})} = \mathbb{E}[f(\mathcal{X}_0)]_{P(\mathcal{X}_0)}$$

that is, we only need to know the marginal distribution of the subset in order to carry out the average.

We are going to discuss now some examples of probability distributions that are going to be used from now on.

2.1 DISCRETE DISTRIBUTIONS

2.1.1 *Bernoulli distribution*

The Bernoulli distribution describes a discrete binary variable X that takes the value 1 with probability p and the value 0 with probability $1 - p$.

$$p(x) = \begin{cases} p & \text{if } x = 1 \\ 1 - p & \text{if } x = 0 \end{cases}$$

2.1.2 *Binomial distribution*

The binomial distribution describes the number of successes in a sequence of independent Bernoulli Trials. A discrete binary random variable X follows a *binomial distribution* of parameters $n \in \mathbb{N}$ and $p \in [0,1]$, denoted as $X \sim B(n,p)$ if and only if

$$P(x=k) = \binom{n}{k} p^k (1-p)^{n-k}$$

2.2 CONTINUOUS DISTRIBUTIONS

2.2.1 *Univariate Normal distribution*

The *normal* or *Gaussian distribution* is a type of continuous probability distribution for real-valued random variables.

Definition 18. We say the real valued random variable X follows a *normal distribution* of parameters $\mu, \sigma \in \mathbb{R}$, denoted as $X \sim N(\mu, \sigma)$ if and only if, its probability density function exists and is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The parameter μ is the mean or expectation of the distribution and σ is its standard deviation.

The simplest case of a normal distribution is known as *standard normal distribution*, denoted as Z. It is a special case where $\mu = 0$ and $\sigma = 1$, then its density function is

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

One of the properties of the normal distribution is that if $X \sim N(\mu, \sigma)$, $a, b \in \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ be defined as $f(x) = ax + b$, then $f(X) \sim N(\mu + b, a^2\sigma)$

2.2.2 *Multivariate Normal Distribution*

This distribution plays a fundamental role in this project so we will discuss its properties in more detail.

This distribution is an extension of the uni-variate one when having a multivariate random variable.

Definition 19. We say that a random vector $\boldsymbol{X} = (X_1, \ldots, X_p)$ follows a *multivariate normal distribution* of parameters $\boldsymbol{\mu} \in \mathbb{R}^p$ and $\boldsymbol{\Sigma} \in \mathbb{M}_n(\mathbb{R})$, denoted as $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ if and only if its probability density function is

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{det(2\pi\boldsymbol{\Sigma})}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Where $\boldsymbol{\mu}$ is the mean vector of the distribution, and $\boldsymbol{\Sigma}$ the covariance matrix. The inverse matrix $\boldsymbol{\sigma}^{-1}$ is called *precision*. It also satisfies that

$$\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}] \qquad \boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T]$$

As $\boldsymbol{\Sigma}$ is a real symmetric matrix, it can be eigendecomposed

$$\boldsymbol{\Sigma} = \boldsymbol{E}\boldsymbol{\Delta}\boldsymbol{E}^T$$

where $\boldsymbol{E}^T\boldsymbol{E} = \boldsymbol{I}$ and $\boldsymbol{\Delta} = diag(\lambda_1, \ldots, \lambda_n)$.

Using the transformation

$$\boldsymbol{y} = \boldsymbol{\Delta}^{\frac{1}{2}}\boldsymbol{E}^T(\boldsymbol{x}-\boldsymbol{\mu})$$

we get that

$$(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}(\boldsymbol{x}-\boldsymbol{\mu}) = \boldsymbol{y}^T\boldsymbol{y}$$

Using this, the multivariate Normal Distribution reduces to a product of n univariate standard normal distributions.

TODO: Add more properties as they are needed

2.2.3 *Beta Distrbution*

Another continuous distribution that we are going to use is the *Beta distribution*.

Definition 20. We say that a continuous random variable X defined on the interval $[0,1]$ follows a *Beta distribution* of parameters $\alpha, \beta > 0$, denoted as $X \sim Beta(\alpha, \beta)$ if and only if its density function is

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

where $B(\alpha,\beta)$ is the *beta function* defined as

$$B(\alpha,\beta) = \int_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$$

2.3 KULLBACK-LEIBLER DIVERGENCE

Definition 21. Let P and Q be two probability distributions over the same set of random variables $\boldsymbol{X}$, the *Kullback-Leibler divergence* $KL(Q|P)$ measures the 'difference' between both distributions as

$$KL(Q|P) = \mathbb{E}[\log Q(x) - \log P(x)]_Q$$

Proposition 2. *The Kullback-Leibler divergence is always non-negative.*

Proof. As the logarithm is bounded by $x - 1$, we can bound $\log \frac{P(x)}{Q(x)}$

$$\log x \leq x - 1 \implies \frac{P(x)}{Q(x)} - 1 \geq \log \frac{P(x)}{Q(x)}$$

Since probabilities are non-negative, we can multiply by $Q(x)$

$$P(x) - Q(x) \geq Q(x) \log P(x) - Q(x) \log Q(x)$$

Now we integrate (sum in case of discrete variables) both sides

$$0 \geq \mathbb{E}[\log P(x) - \log Q(x)]_Q \implies \mathbb{E}[\log Q(x) - \log P(x)]_Q \geq 0$$

□

As a result, the Kullback-Leibler divergence is 0 if and only if the two distributions are equal almost everywhere.

3

GRAPH THEORY

Definition 22. 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*.

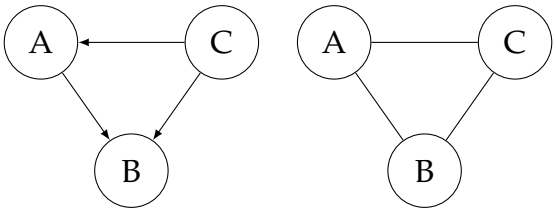


Figure 1: Example of directed and undirected graph, respectively.

Definition 23. 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 24. 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 25. 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 26. 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.

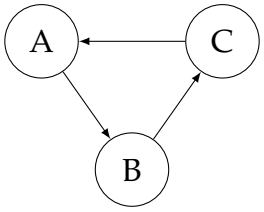


Figure 2: Example of graph which isn't a DAG.

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

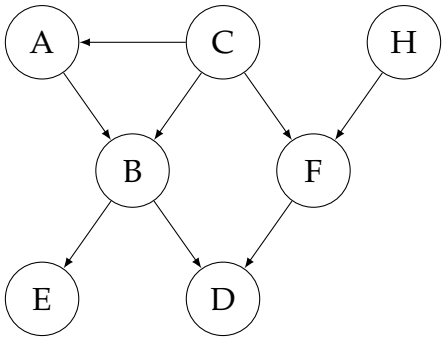


Figure 3: Directed acyclic graph

Definition 27. 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 28. Let G be a DAG, U be a path between two vertex and $A \in U$

- A is called a *collider* if $\forall B \in ne(A) \cap U, (B, A) \in E$.
- A is called a *fork* if $\forall B \in ne(A) \cap U, (A, B) \in E$.

Notice, a vertex can be a collider for a path but not for others.

For example in figure 3, D is a collider and C is a fork.

Definition 29. Let G be an undirected graph, a *clique* is a maximally connected subset of vertices. That is, all the members of the clique are connected to each others and there is no bigger clique that constains another.

Formally, $S \subset V$ is a *clique* if and only if $\forall A, B \in S,\ \{A, B\} \in E$ and $\nexists C \in V \backslash S$ such that $\forall A \in S,\ \{A, C\} \in E$.

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.

4

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 30. 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 31. A *belief network or Bayesian network* is a pair (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.

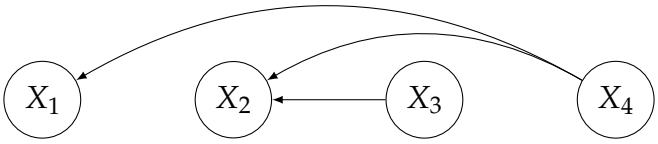


Figure 4: Bayesian Network factorizing $P(x_1, x_2, x_3, x_4) = P(x_1|x_4)P(x_2|x_3, x_4)P(x_3)P(x_4)$

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 3, $X_1 \perp\!\!\!\perp X_2 \mid X_4$ 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 32. 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 33. 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.

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.*

1 Continuous variable notation is used

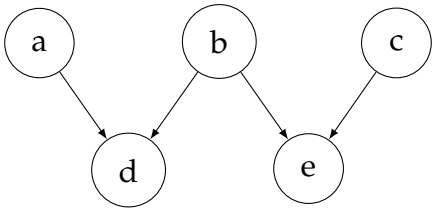


Figure 5: D-separation example

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

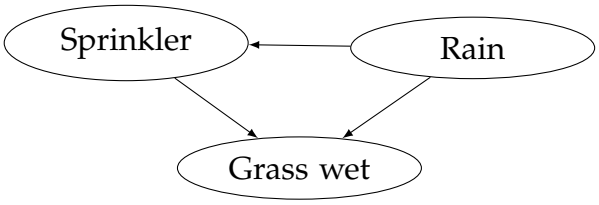
Example 3. In this example we are modeling three discrete random variables: Sprinkler (S), Rain (R) and Grass wet (G).

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:

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

In some situations our Belief Networks will contain a number of nodes that are essentially the same but repeated a number of times, for this, we are going to introduce the *plate notation*. Suppose we have the situation that figure 6 shows on the left. The we can collapse all B_i variables in a box, indicating there number of variables inside it.

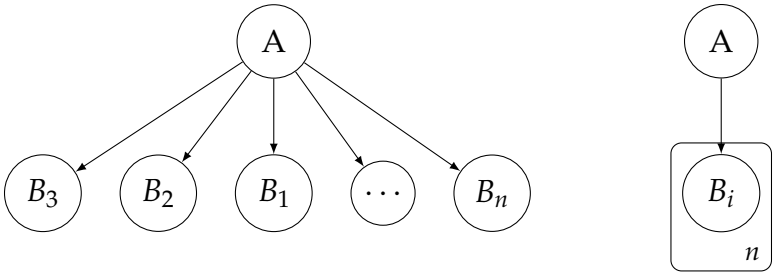


Figure 6: Plate notation example. Standard notation on the left and plate on the right

5

MARKOV RANDOM FIELDS

Definition 34. A *potential* ϕ is a non-negative function. It is worth to mention that a probability distribution is a special case of a potential.

Definition 35. Let $\boldsymbol{X}$ be a set of random variables, G an undirected graph, $\boldsymbol{X}_c, c \in \{1, \ldots, C\}$ be the maximal cliques of G and P a probability distribution over $\boldsymbol{X}$. The pair (G, P) is said to be a *Markov network or Markov random field* if, and only if

$$P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c=1}^{C} \phi_c(\boldsymbol{X}_c)$$

where Z is a constant that ensures normalization.

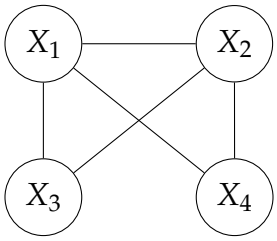


Figure 7: Markov Network $P(x_1, x_2, x_3, x_4) = \phi(x_1, x_2, x_3)\phi(x_2, x_3, x_4)/Z$

In figure 7 we can see an example of the factorization, without giving any reference of the potentials.

Let (G, P) be a Markov network, then it satisfies the following properties known as Markov properties:

- Pairwise Markov property. Any two non-adjacent variables are conditionally independent given all other variables.
- Local Markov property. A variable is conditionally independent over all other variables given it's neighbors. That is,

$$P(x_i | x_{\setminus i}) = p(x_i | ne(x_i))^1$$

1 $X_{\setminus i} = \{X_j \mid j \neq i\} \subset \boldsymbol{X}$

- Global Markov property. Any two subsets of variables are conditionally independent given a separating subset (any path from one set to the other passes trough this one).

Remark 1. A Markov network can also be defined as a pair (G, P) such as all Markov properties are satisfied. The clique factorization definition is a special case of these properties.

Definition 36. Let G be an undirected graph and P a probability distribution over a set of random variables $\boldsymbol{X}$. The pair (G, P) is called a Markov Random Field if and only if it follows the local Markov Property.

Part III

NAME THIS PART

6

LEARNING AS INFERENCE

In Machine Learning and related fields, the distributions are not fully specified and need to be learned from the data.

From now on, $\mathcal{V}$ will denote the known data and θ the set of parameters of the data distributions. The main task is to determine this set of parameters using the information given by the data.

Definition 37. *Priors* and *posteriors* typically refer to the parameter distribution before and after seeing the data, respectively. Using Bayes' rule

$$P(\theta \mid \mathcal{V}) = \frac{P(\mathcal{V} \mid \theta)P(\theta)}{P(\mathcal{V})}$$

The factor $P(\mathcal{V} \mid \theta)$ is called the *likelihood*.

Let us see an example of our goal, in it we will try to learn the bias of a coin, given a set of tossing results.

Example 4. Let $\{v_n\}_{n\in 0,\dots,N}$ be the results of tossing a coin $N \in \mathbb{N}$ times, let 1 symbolize *heads* and 0 *tails*.

Our objective is to estimate the probability θ that the coin will be head $P(v_n = 1 \mid \theta)$, for this we have the random variables $v_1, \dots, v_n$ and θ, and we require a model $P(v_1, \dots, v_n, \theta)$. We are considering the variables v_i to be independent to each others, we have a Belief Network depicted in figure 8

$$P(v_1, \dots, v_n, \theta) = P(\theta) \prod_{n=1}^{N} P(v_n \mid \theta)$$

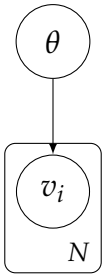


Figure 8: Belief network for coin tossing

We want to calculate

$$P(\theta \mid v_1, \ldots, v_n) = \frac{P(v_1, \ldots, v_n \mid \theta) P(\theta)}{P(v_1, \ldots, v_n)}$$

to do so, we need to specify the prior $P(\theta)$, we are using a discrete model where

$$P(\theta = 0.2) = 0.1 \qquad P(\theta = 0.5) = 0.7 \qquad P(\theta = 0.8) = 0.2$$

This means that we have a 70% belief that the coin is fair, a 10% belied that is biased to tails and 20% that is biased to heads. Notice that $P(v_n \mid \theta) = \theta$ if $v_n = 1$ and $P(v_n \mid \theta) = 1 - \theta$ if $v_n = 0$.

Let n_h be the number of heads in our observed data and n_t the number of tails

$$P(\theta \mid v_1, \ldots, v_n) = \frac{P(\theta)}{P(\mathcal{V})} \theta^{n_h} (1 - \theta)^{n_t}$$

Suppose now that $n_h = 2$ and $n_t = 0.8$, then

$$P(\theta = 0.2 \mid \mathcal{V}) = \frac{1}{P(\mathcal{V})} \times 0.1 \times 0.2^2 \times 0.8^8 = \frac{1}{P(\mathcal{V})} \times 6.71 \times 10^{-4}$$
$$P(\theta = 0.5 \mid \mathcal{V}) = \frac{1}{P(\mathcal{V})} \times 0.7 \times 0.5^2 \times 0.5^8 = \frac{1}{P(\mathcal{V})} \times 6.83 \times 10^{-4}$$
$$P(\theta = 0.8 \mid \mathcal{V}) = \frac{1}{P(\mathcal{V})} \times 0.2 \times 0.2^2 \times 0.8^8 = \frac{1}{P(\mathcal{V})} \times 3.27 \times 10^{-7}$$

Now, we can compute

$$\frac{1}{P(\mathcal{V})} = 6.71 \times 10^{-4} + 6.83 \times 10^{-4} + 3.27 \times 10^{-7} = 0.00135$$

So,

$$P(\theta = 0.2 \mid \mathcal{V}) = 0.4979$$
$$P(\theta = 0.5 \mid \mathcal{V}) = 0.5059$$
$$P(\theta = 0.8 \mid \mathcal{V}) = 0.00024$$

These are the posterior parameter beliefs of our experiment. Given this, it we were to choose a single value for the posterior it would be $\theta = 0.5$. This result is intuitive since, we had a strong belief of the coin being fair and even though the number of tails was quite bigger than heads, it was not enough to make the difference. Obviously the posterior of the coin being biased to tails is now bigger than the prior.

Let us use an uniform prior distribution so that $P(\theta) = k \implies \int_0^1 P(\theta)d\theta = k = 1$ due to normalization.

Using the previous calculations we have

$$P(\theta \mid \mathcal{V}) = \frac{1}{P(\mathcal{V})}\theta^{n_h}(1-\theta)^{n_t}$$

where

$$P(\mathcal{V}) = \int_0^1 \theta^{n_h}(1-\theta)^{n_t}d\theta$$

this implies that

$$P(\theta \mid \mathcal{V}) = \frac{\theta^{n_h}(1-\theta)^{n_t}}{\int_0^1 u^{n_h}(1-u)^{n_t}du} \equiv Beta(n_h+1, n_t+1)$$

Definition 38. If the posterior distribution is in the same probability distribution family as the prior distribution, they are then called *conjugate distributions*, and the prior is called a *conjugate prior* of the likelihood distribution.

Let's use a Beta distribution as the prior in the last example

$$P(\theta) = \frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \equiv Beta(\alpha,\beta)$$

then, repeating the same as before we get that

$$P(\theta,\mathcal{V}) = \frac{1}{B(\alpha+n_h,\beta+n_t)}\theta^{\alpha+n_h-1}(1-\theta)^{\beta+n_t-1} \equiv Beta(\alpha+n_h,\beta+n_t)$$

So both the prior and posterior are Beta distributions, then the Beta distribution is called "conjugate" of the Binomial distribution.

6.1 UTILITY

The Bayesian posterior says nothing about how to summarize the beliefs it represents, in order to do this we need to specify the utility of each decision.

With this idea we define an utility function over the parameters

$$U(\theta, \theta_{true}) = \alpha \mathbb{I}[\theta = \theta_{true}] - \beta \mathbb{I}[\theta \neq \theta_{true}]$$

where θ_{true} symbolizes the true value of the parameter, and $\alpha, \beta \in \mathbb{R}$.

Then the expected utility of a parameter θ_0 is calculated as

$$U(\theta = \theta_0) = \sum_{\theta_{true}} U(\theta = \theta_0, \theta_{true}) P(\theta = \theta_{true} \mid \mathcal{V})$$

Using the last example, we may define out utility function as

$$U(\theta, \theta_{true}) = 10 \mathbb{I}[\theta = \theta_{true}] - 20 \mathbb{I}[\theta \neq \theta_{true}]$$

so the expected utility of the decision that the parameter is $\theta = 0.2$ is

$$\begin{aligned} U(\theta = 0.2) = {} & U(\theta = 0.2, \theta_{true} = 0.2) P(\theta_{true} = 0.2 \mid \mathcal{V}) \\ & + U(\theta = 0.2, \theta_{true} = 0.5) P(\theta_{true} = 0.5 \mid \mathcal{V}) \\ & + U(\theta = 0.2, \theta_{true} = 0.8) P(\theta_{true} = 0.8 \mid \mathcal{V}) \end{aligned}$$

6.2 MAXIMUM A POSTERIORI AND MAXIMUM LIKELIHOOD

The posterior reflects our beliefs about the full range of probabilities, but we may want to summarize all this information, even though, we may lose loads of it.

Definition 39. Maximum Likelihood is calculated as

$$\theta^{ML} = \arg\max_{\theta} p(\mathcal{V} \mid \theta)$$

it refers to the value of the parameter θ for which the observed data better fits the model.

Definition 40. Maximum A Posteriori is calculated as

$$\theta^{MAP} = \arg\max_{\theta} p(\mathcal{V} \mid \theta) P(\theta)$$

The decision of taking the Maximum A Posteriori can be motivated using an utility that equals zero for all but the correct parameter

$$U(\theta, \theta_{true}) = \mathbb{I}[\theta = \theta_{true}]$$

using this, the expected utility of a parameter $\theta = \theta_0$ is

$$U(\theta = \theta_0) = \sum_{\theta_{true}} \mathbb{I}[\theta_{true} = \theta_0] P(\theta = \theta_{true} \mid \mathcal{V}) = P(\theta_0 \mid \mathcal{V})$$

This means that the maximum utility decision is to take the value θ_0 with the highest posterior value.

It is worth mentioning that when using a flat prior $\theta^{ML} = \theta^{MAP}$.

Let $\{x_1, \ldots, x_m\}$ be a set of discrete variables, we can define the empirical distribution as a distribution of the variables whose mass probability function is

$$Q(x) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}[x = x_i]$$

Now, we are going to show the relation between the Maximum Likelihood and the Kullback-Leibler divergence of the empirical distribution and our model. We may calculate this Kullback-Leibler divergence and study their functional independence.

$$KL(Q \mid P) = \mathbb{E}[\log(Q(x))]_Q - \mathbb{E}[\log(P(x))]_Q$$

Notice the term $\mathbb{E}[\log(Q(x))]_Q$ is a constant and assume the data is i.i.d, so that

$$\mathbb{E}[\log(P(x))]_Q = \frac{1}{m} \sum_{i=1}^{m} logP(x_i)$$

where the right side is the log likelihood under $P(x)$. As the logarithm is a strictly increasing function, maximizing the log likelihood equals to maximize the likelihood itself, and we can see here how it is equivalent to minimize the Kullback-Leibler divergence between the empirical distribution and our distribution.

In case $P(x)$ is unconstrained, the optimal choice is $P(x) = Q(x)$, that is, the maximum likelihood distribution corresponds to the empirical distribution.

For a Belief Network $P(x)$ presents the following constraint

$$P(x) = \prod_{i=1}^{K} P(x_i \mid pa(x_i))$$

We now want to minimize the Kullback-Leibler divergence between the empirical distribution $Q(x)$ and $P(x)$ in order to get the Maximum Likelihood.

$$\begin{aligned} KL(Q \mid P) &= -\mathbb{E}\big[\sum_{i=1}^{K} \log P(x_i \mid pa(x_i))\big]_Q + \mathbb{E}\big[\sum_{i=1}^{K} \log P(x_i \mid pa(x_i))\big]_P \\ &= -\sum_{i=1}^{K} \mathbb{E}\big[\log P(x_i \mid pa(x_i))\big]_Q + \sum_{i=1}^{K} \mathbb{E}\big[\log P(x_i \mid pa(x_i))\big]_P \end{aligned}$$

We can now use that $\log P(x_i \mid pa(x_i))$ only depends on $Q(x_i \mid pa(x_i))$ to rewrite it as

$$KL(Q \mid P) = \sum_{i=1}^{K} \mathbb{E}\Big[\log Q(x_i \mid pa(x_i))\Big]_{Q(x_i, pa(x_i))} - \mathbb{E}\Big[\log P(x_i \mid pa(x_i))\Big]_{Q(x_i, pa(x_i))}$$
$$= \sum_{i=1}^{K} \mathbb{E}\Big[KL\Big(Q(x_i \mid pa(x_i)) \mid P(x_i \mid pa(x_i))\Big)\Big]_{Q(x_i, pa(x_i))}$$

The minimal setting is then

$$P(x_i \mid pa(x_i)) = Q(x_i \mid pa(x_i))$$

in terms of the initial data it is to set $P(x_i \mid pa(x_i))$ to the number of times the state appears in it.

6.3 BAYESIAN BELIEF NETWORK TRAINING

A Bayesian approach where we set a distribution over the parameters is an alternative to ML training of a Bayesian Network.

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.