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, Sunday 3rd May, 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. Given two events $\alpha, \beta \in \mathcal{F}$, with $P(\beta) \neq 0$, the conditional probability of α given β is defined as the quotient of the probability of the joint events and the probability of β:

$$P(\alpha \mid \beta) = \frac{P(\alpha \cap \beta)}{P(\beta)}$$

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

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

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 $P(h \mid d) = 0.9$. What is the probability that a person with habit h will have the disease d?

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

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 4. We say that two events $\alpha, \beta \in \mathcal{F}$ are *independent* if knowing one of them does not give any extra information about the other. Mathematically,

$$P(\alpha \cap \beta) = P(\alpha)P(\beta) \qquad\qquad P(\alpha \mid \beta) = P(\alpha)$$

Let $\gamma \in \mathcal{F}$, we say that α and β are *conditionally independent* on γ, $\alpha \perp\!\!\!\perp \beta \mid \gamma$ if and only if

$$P(\alpha \cup \beta \mid \gamma) = P(\alpha \mid \gamma)P(\beta \mid \gamma)$$

Otherwise, they are said to be *conditionally dependent* on γ, $\alpha \not\perp\!\!\!\perp \beta \mid \gamma$.

Now we are going to introduce the concept of *random variable* and some properties as we have done with events.

Definition 5. 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 6. A *random variable* is a measurable function $X : \Omega \to E$ from a probability space $(\Omega, \mathcal{F}, P)$ to a measurable space $(E, \mathcal{F}')$ verifying $X(\omega) \in \mathcal{F}' \ \forall \omega \in \Omega$.

The probability of X taking a value on a measurable set $S \in E$ 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 will set the following 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 7. The *cumulative distribution function* of a real-valued 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 8. 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 and real, then X is called a *continuous random variable* if there exists is a non-negative Lebesgue-integrable f, called its *probability density function* such that

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

A *mixed random variable* is a random variable who is neither discrete nor continuous, it can be realized as the sum of a discrete and continuous random variables. An example of a random variable of mixed type would be based on an experiment where a coin is flipped and a random positive number is chose only if the result of the coin toss is heads, -1 otherwise.

From now on, $P(x)$ will denote $f_X(x)$ when X is a continuous random variable. We will define the *probability distribution* P_X of a random variable X over the probability space $(\Omega, \mathcal{F}, P)$ as the pushforward measure of it, that is, $P_X = PX^{-1}$.

As summation is integration with respect to the *counting measure* defined as

$$\#(dx) = \sum_{n \in \mathcal{I}} \delta(x - n)dx$$

Where $\mathcal{I}$ is the set of values X can take, and δ is the Dirac distribution. Then

$$\int_x P(x)\#(dx) = \sum_{n \in \mathcal{I}} \int_x P(x)\delta(x - n)dx = \sum_{n \in I} P(n)$$

Where we used that $\int f(x)\delta(x - x_0) = f(x_0)$. Given this, from now on, we will use the integration notation for both discrete and continuous variables given that the integrals will be respect to the counting measure when needed.

Definition 9. As we did for events, we can define the *conditional probability* over random variables, let X, Y be random variables,

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

It is required that $P(y) \neq 0$ for the conditional probability to be defined.

We can also enunciate the *Bayes' theorem*.

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

Definition 10. 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

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

Definition 11. 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 12. 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 $P(x \mid y) = P(x)$.

Definition 13. Let X, Y and Z be three random variables, then X and Y are *conditionally independent* given Z if and only if

$$P(x, y \mid z) = P(x \mid z)P(y \mid z)$$

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

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

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

Definition 14. 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

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

Definition 15. A *multivariate random variable* or *random vector* is a column vector $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ whose components are random variables that can be defined over different probability spaces.

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.

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 16. 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 17. The notation $\mathbb{E}[X]$ is used to denote the *average* or *expectation* of the values a real-valued variable takes respect to its distribution. It is worth mentioning that it might not exists. 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 $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$. Where

$$X^+(\omega) = \max(X(\omega), 0) \qquad X^-(\omega) = \min(X(\omega), 0)$$

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 18. 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. Note μ_k can also denote the k^{th} element in the mean vector of a multivariate variable.

Definition 19. 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 $\boldsymbol{X} = (X_1, \dots, X_n)^T$ 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 $\mathcal{X} = \{X_1, \dots, X_n\}$ be a set of random variables, $\mathcal{X}_0 \subset \mathcal{X}$ and $P(\mathcal{X}), P(\mathcal{X}_0)$ their probability distributions. It follows that the expectation of a function g over $\mathcal{X}_0$, verifies*

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

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

Proof. Let $\mathcal{I} = (i_1, \dots, i_k)$ be the indexes corresponding to $\mathcal{X}_0$, then

$$\begin{aligned}\mathbb{E}_{P(\mathcal{X})}\Big[g(\mathcal{X}_0)\Big] &= \int_{x_1} \cdots \int_{x_n} g(x_{i_1}, \dots, x_{i_k}) f(x_1, \dots, x_n) \\ &= \int_{x_{i_1}} \cdots \int_{x_{i_k}} g(x_{i_1}, \dots, x_{i_k}) \int \cdots \int f(x_1, \dots, x_n) d_{x_1}, \dots, d_{x_n} \\ &= \int_{x_{i_1}} \cdots \int_{x_{i_k}} g(x_{i_1}, \dots, x_{i_k}) f(x_{i_1}, \dots, x_{i_k}) = \mathbb{E}_{P(\mathcal{X}_0)}\Big[g(\mathcal{X}_0)\Big]\end{aligned}$$

Where in the second-last equality we used marginalization. □

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 *Categorical Distribution*

A generalization of the Bernoulli Distribution when the variable can take more than two states is the *Categorical Distribution*. Let $Dom(X) = \{1, \ldots, N\}$, then X follows a categorical distribution of parameters $\theta = (\theta_1, \ldots, \theta_N)$ if and only if

$$P(x \mid \theta) = \prod_{i=1}^{N} \theta_i^{\mathbb{I}[x=i]} \text{ and } \sum_{i=1}^{N} \theta_i = 1$$

2.1.3 *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 20. 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 univariate one when having a multivariate random variable.

Definition 21. 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

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

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.

2.2.3 *Beta Distribution*

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

Definition 22. 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

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

The mean is given by $\mathbb{E}\Big[X\Big] = \frac{\alpha}{\alpha+\beta}$

2.2.4 *Dirichlet Distribution*

The Dirichlet distribution is a family of continuous multivariate probability distributions parameterized by a vector $\boldsymbol{\alpha}$ of positive reals. It is a multivariate generalization of the Beta Distribution.

Definition 23. We say that a continuous random multivariate variable $\boldsymbol{X}$ with order $K \geq 2$, follows a *Dirichlet Distribution* with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, if and only if its density function is defined as

$$f(\boldsymbol{x}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} x_k^{\alpha_k - 1}$$

and it satisfies that

$$\sum_{k=1}^{K} x_k = 1 \text{ and } x_k > 0 \; \forall k = 1, \ldots, K$$

Where the normalization constant is the multivariate beta function

$$B(\boldsymbol{\alpha}) = \frac{\prod_k \Gamma(\alpha_k)}{\Gamma\left(\sum_k \alpha_k\right)}$$

Proposition 2. *Let* $(X_0, \ldots, X_n) \sim Dirichlet(\alpha_0, \ldots, \alpha_n)$, *then* $X_0 \sim Beta(\alpha_0, \alpha_1 + \cdots + \alpha_n)$

Proof. Following [Farrow (2008)], we can write the joint probability as

$$f(x_1, \ldots, x_n) = f_1(x_1) f_2(x_2 \mid x_1) \ldots f_{n-1}(x_{n-1} \mid x_1, \ldots, x_{n-2})$$

We do not need the last term because it is fixed given the others. In fact, let $A = \sum_i \alpha_i$, we can write it as

$$\left(\frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(A-\alpha_1)} x_1^{\alpha_1-1}(1-x_1)^{A-\alpha_1-1}\right)\left(\frac{\Gamma(A-\alpha_1)}{\Gamma(\alpha_2)\Gamma(A-\alpha_1-\alpha_2)} \frac{x_2^{\alpha_2-1}(1-x_1-x_2)^{A-\alpha_2-\alpha_1-1}}{(1-\alpha_1)^{A-\alpha_1-1}}\right)$$

$$\dots \Big(\frac{\Gamma(A-\alpha_1-\dots-\alpha_{n-2})}{\Gamma(\alpha_{n-1})\Gamma(A-\alpha_1-\dots-\alpha_{n-1})} \frac{x_{n-1}^{\alpha_n-1} x_n^{\alpha_n-1}}{(1-x_1-\dots-x_{n-2})^{\alpha_{n-1}+\alpha_n-1}} \Big)$$

From this, we get that

$$f_1(x_1) = \frac{\Gamma(A)}{\Gamma(\alpha_1)\Gamma(A-\alpha_1)} x_1^{\alpha_1-1}(1-x_1)^{A-\alpha_1-1} \implies X_1 \sim \text{Beta}(\alpha_1, A-\alpha_1)$$

Making the decomposition over any other X_j, results on $X_j \sim \text{Beta}(\alpha_j, A-\alpha_j)$. □

2.3 KULLBACK-LEIBLER DIVERGENCE

Definition 24. 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 3. *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 25. 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*.

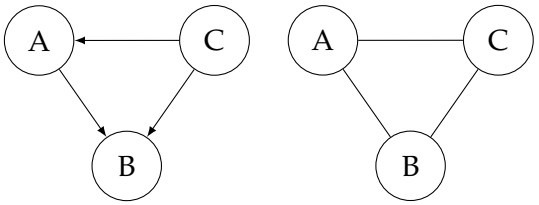


Figure 1: Example of directed and undirected graph, respectively.

Definition 26. 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 27. 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 28. A *directed acyclic graph (DAG)* is a directed graph such that no directed path between any two nodes revisits a vertex.

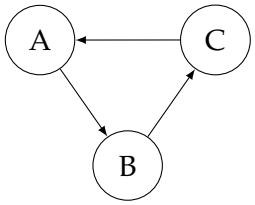


Figure 2: Example of graph which isn't a DAG.

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 29. 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.

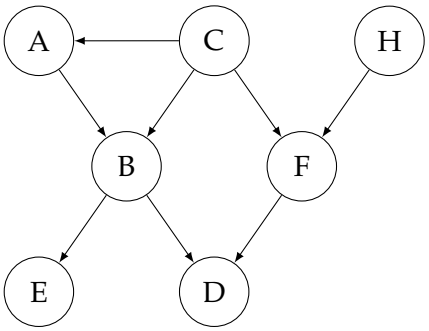


Figure 3: Directed acyclic graph

Definition 30. 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 31. 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. A vertex is said to be a collider or a fork without any reference to the path when it is for any path that goes through it. This happens when the edge direction condition is satisfied for every neighbor.

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

Definition 32. 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 contains 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:

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

Let us 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 33. 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 34. 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

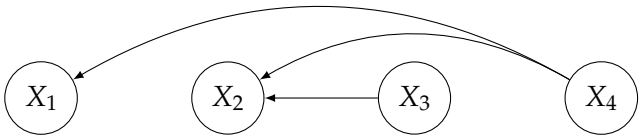


Figure 4: Bayesian Network factorizing $P(x_1, x_2, x_3, x_4) = P(x_1|x_4)P(x_2|x_3, x_4)P(x_3)P(x_4)$

large, this is still intractable so the tables are generally parameterized in a low dimensional manner.

Bayesian Networks are good for encoding conditional independence over the variables, but aren't for encoding dependence. For example, with the following network

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

represented as $x \to y$ in a DAG (a Bayesian network can be defined giving the graph or the joint probability equally). 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 35. 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 36. 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

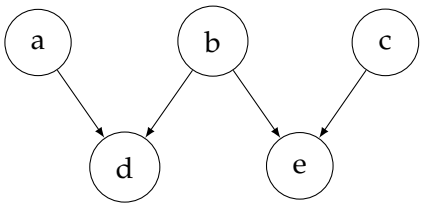


Figure 5: D-separation example

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

Theorem 2 (Pearl & Dechter (2013)). *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 [Shachter (2013)] provides a linear time complexity algorithm that computes conditional independent using this theorem.

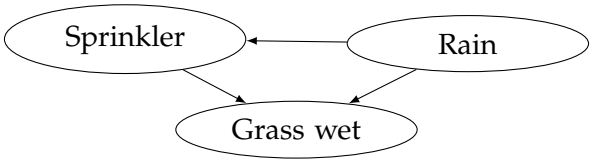
Example 2. In this example we are modeling three discrete random variables: Sprinkler (S), Rain (R) and Grass wet (G).

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:

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

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.

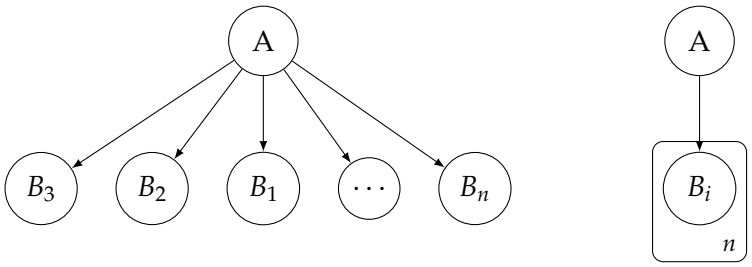


Figure 6: Plate notation example. Standard notation on the left and plate on the right

5

MARKOV RANDOM FIELDS

Definition 37. 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 38. 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.

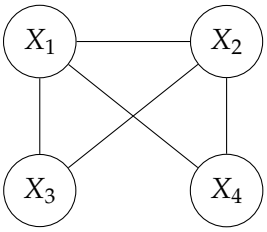


Figure 7: Markov Network $P(x_1, x_2, x_3, x_4) = \phi(x_1, x_2, x_3)\phi(x_2, x_3, x_4)/Z$

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,

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

- 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).

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

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 39. 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

BAYESIAN NETWORKS LEARNING

6

INTRODUCTION

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 40. *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 3. Let $\mathcal{V} = \{v_n\}_{n\in 0,\dots,N}$ 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 i.i.d random variables $v_1, \dots, v_n$ and θ, and we require a model $P(v_1, \dots, v_n, \theta)$. We have a Belief Network shown in figure 8

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

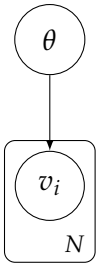


Figure 8: Belief network for coin tossing

We want to calculate

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

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 $P(v_n = 1 \mid \theta) = \theta$ and $P(v_n = 0 \mid \theta) = 1 - \theta$.

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

$$n_h = \#\{v = 1\} \qquad n_t = \#\{v = 0\}$$

then the posterior has the form

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

Suppose now that $n_h = 2$ and $n_t = 8$, then

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

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,

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

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, 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.

Suppose 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

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

Definition 41. 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

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

then, repeating the same as before we get that

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

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 benefit from 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 $\alpha, \beta \in \mathbb{R}$. This symbolizes the gains or looses of choosing the parameter θ, when the true value of the parameter is supposed to be θ_{true}. 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

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

Where we interpret that the loss of choosing the wrong parameter is twice as important as the gains from doing it right.

The expected utility of the decision that the parameter is $\theta = 0.2$ in our discrete example would be

$$\begin{aligned} U(\theta = 0.2) &= U(\theta = 0.2, \theta_{true} = 0.2)P(\theta_{true} = 0.2 \mid \mathcal{V}) \\ &\quad + U(\theta = 0.2, \theta_{true} = 0.5)P(\theta_{true} = 0.5 \mid \mathcal{V}) \\ &\quad + U(\theta = 0.2, \theta_{true} = 0.8)P(\theta_{true} = 0.8 \mid \mathcal{V}) \\ &= 10 \times 0.4979 - 20 \times 0.5059 - 20 \times 0.00024 \\ &= -5.1438 \\ U(\theta = 0.5) &= -4.9038 \\ U(\theta = 0.8) &= -20.0736 \end{aligned}$$

This illustrate how an utility function can affect the results of the inference. The most probable value for θ was 0.2, but, using this utility function, 0.5 is the one with which we expect minor losses.

7

MAXIMUM LIKELIHOOD TRAINING

In this section we will introduce two concepts, Maximum Likelihood and Maximum a Posteriori, showing that *training* a model's parameter to maximize the Maximum Likelihood equals to take the empirical distribution as it.

Let $\{X_1, \ldots, X_N\}$ be a set of real i.i.d random variables and $\mathcal{V} = \{x_1, \ldots, x_N\}$ be the set of observations and θ the considered parameters of the model.

Definition 42. 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 43. Maximum A Posteriori refers to

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

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.

Remark 2. When using a flat prior, i.e, $P(\theta)$ is constant, $\theta^{ML} = \theta^{MAP}$.

7.1 ML AND KL DIVERGENCE

Now, we are going to show the relation between the Maximum Likelihood and the Kullback-Leibler divergence of the empirical distribution and our model. Firstly, we define the empirical distribution as a distribution whose probability mass function Q is

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

We may calculate the Kullback-Leibler divergence between the empirical and our considered model $P(x \mid \theta)$ and study their functional independence.

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

Notice the term $\mathbb{E}_Q\Big[\log(Q(x))\Big]$ is a constant and the log likelihood under Q takes the form

$$\mathbb{E}_Q\Big[\log(P(x \mid \theta))\Big] = \frac{1}{N}\int_x \sum_{i=1}^{N} \mathbb{I}[x = x_i] \log P(x \mid \theta) = \frac{1}{N}\sum_{i=1}^{N} \log P(x_i \mid \theta)$$

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.

$$\arg\min_\theta KL\Big(Q \mid P\Big) = \arg\min_\theta -\mathbb{E}_Q\Big[\log(P(x \mid \theta))\Big] = \arg\max_\theta \mathbb{E}_Q\Big[\log(P(x \mid \theta))\Big]$$

$$\arg\max_\theta \frac{1}{N}\sum_{i=1}^{N} \log P(x_i \mid \theta) = \arg\max_\theta \frac{1}{N}\sum_{i=1}^{N} P(x_i \mid \theta) = \arg\max_\theta \sum_{i=1}^{N} P(x_i \mid \theta) = \theta^{ML}$$

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

For a Belief Network we know there is the following constraint

$$P(x_1, \ldots, x_m \mid \theta) = \prod_{i=1}^{K} P(x_i \mid pa(x_i), \theta)$$

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

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

Notice the expectation is over the full distribution $Q(x_1, \ldots, x_k)$, we can use proposition 1 on $\log P(x_i \mid pa(x_i), \theta)$ and $Q(x_i, pa(x_i))$.

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

The minimal setting is then

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

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.

8

BAYESIAN TRAINING

A Bayesian approach where we set a distribution over the parameters is an alternative to Maximum Likelihood training of a Bayesian Network, as we did in the coin tossing example.

We go deep into it using the following scenario, consider a disease D and two habits A and B. Let $N = 7$ be the number of observations of the variables as shown in the table below, that means we are considering the following i.i.d variables $\{A_1, \ldots, A_N\}$, $\{B_1, \ldots, b_N\}$ and $\{D_1, \ldots, D_N\}$ governed by the parameters θ_A, θ_B and θ_D as shown in the following Bayesian Network.

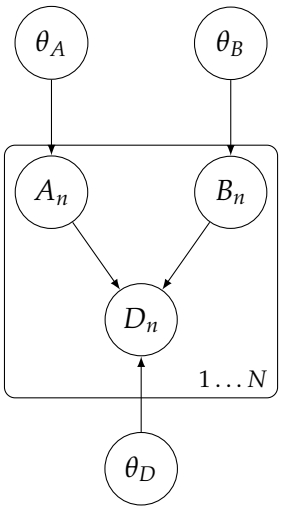


Figure 9: Bayesian parameter model for the relation between A, B, D

A	B	D
1	1	1
1	0	0
0	1	1
0	1	0
1	1	1
0	0	0
1	0	1

Table 1: Observations

The graph gives the following joint probability distribution over A, B and D

$$P(a, b, d) = P(d \mid a, b)P(a)P(b)$$

Let $\mathcal{V} = \{(a_n, b_n, d_n), n = 1, \ldots, N\}$ be the set of observations.

We need a notation for the parameters, as all the variables are binary we are going to use

$$P(A = 1 \mid \theta_A) = \theta_A, \qquad P(B = 1 \mid \theta_B = \theta_B), \qquad P(D = 1 \mid A = 0, B = 1, \theta_D) = \theta_1$$

$$\theta_D = (\theta_0, \theta_1, \theta_2, \theta_3)$$

Where we are using a binary to decimal transformation between the states of A and B and the subindex of θ.

We need to specify a prior and since dealing with multi-dimensional continuous distributions is computationally problematic it is normal to use uni-variate distributions.

A convenient assumption is that the prior factorizes, this is usually called *global parameter independence*. We assume then

$$P(\theta_A, \theta_B, \theta_D) = P(\theta_A)P(\theta_B)P(\theta_D)$$

Assuming our data is i.i.d, we have

$$P(\theta_A, \theta_B, \theta_D, \mathcal{V}) = P(\theta_A)P(\theta_B)P(\theta_D)\prod_n P(a_n \mid \theta_A)P(b_n \mid \theta_B)P(d_n \mid a_n, b_n, \theta_D)$$

Learning then corresponds to inference

$$\begin{aligned} P(\theta_A, \theta_B, \theta_D \mid \mathcal{V}) &= \frac{P(\mathcal{V} \mid \theta_a, \theta_B, \theta_D)P(\theta_A, \theta_B, \theta_D)}{P(\mathcal{V})} = \frac{P(\mathcal{V} \mid \theta_A, \theta_B, \theta_D)P(\theta_A)P(\theta_B)P(\theta_D)}{P(\mathcal{V})} \\ &= \frac{1}{P(\mathcal{V})}P(\theta_A)\prod_n P(a_n \mid \theta_A)P(\theta_B)\prod_n P(b_n \mid \theta_B)P(\theta_D)\prod_n P(d_n \mid a_n, b_n, \theta_D) \\ &= P(\theta_A \mid \mathcal{V}_A)P(\theta_B \mid \mathcal{V}_B)P(\theta_D \mid \mathcal{V}) \end{aligned}$$

Where V_i is the subset of the data restricted to the variable i. If we further assume that $P(\theta_D)$ factorizes as $P(\theta_D) = P(\theta_0)P(\theta_1)P(\theta_2)P(\theta_3)$, this is called *local parameter independence*, then it follows that

$$P(\theta_D \mid \mathcal{V}) = P(\theta_0 \mid \mathcal{V})P(\theta_1 \mid \mathcal{V})P(\theta_2 \mid \mathcal{V})P(\theta_3 \mid \mathcal{V})$$

8.1 LEARNING BINARY VARIABLES

The simplest cases to continue are $P(a \mid \theta_A)$ and $P(b \mid \theta_b)$ since they require only a univariate prior distribution $P(\theta_A)$ or $P(\theta_b)$. We use $P(\theta_A)$ as the other case is analogous.

The posterior is

$$P(\theta_A \mid \mathcal{V}_A) = \frac{1}{P(\mathcal{V}_A)}P(\theta_A)\theta_A^{\#(a=1)}(1-\theta_A)^{\#(a=0)}$$

The most convenient choice for the prior is a Beta distribution as conjugacy will hold.

$$\theta_A \sim \text{Beta}(\alpha_A, \beta_A) \implies P(\theta_A) = \frac{1}{B(\alpha_A, \beta_A)}\theta_A^{\alpha_A - 1}(1-\theta_A)^{\beta_A - 1}$$

So it follows that

$$(\theta_A \mid \mathcal{V}_A) \sim \text{Beta}(\theta_A \mid \alpha_A + \#(A=1), \beta_A + \#(A=0))$$

The marginal is then

$$\begin{aligned} P(A=1 \mid \mathcal{V}_A) &= \frac{P(A=1,\mathcal{V}_A)}{P(\mathcal{V}_A)} = \int_{\theta_A} \frac{P(A=1,\mathcal{V}_A,\theta_A)}{P(\mathcal{V}_A)} = \int_{\theta_A} \frac{P(A=1 \mid \mathcal{V}_A,\theta_A)P(\mathcal{V}_A,\theta_A)}{P(\mathcal{V}_A)} \\ &= \int_{\theta_A} \frac{P(A=1 \mid \mathcal{V}_A,\theta_A)P(\theta_A \mid \mathcal{V}_A)P(\mathcal{V}_A)}{P(\mathcal{V}_A)} = \int_{\theta_A} P(\theta_A \mid \mathcal{V}_A)\theta_A = \mathbb{E}\left[\theta_A \mid \mathcal{V}_A\right] \\ &= \frac{\alpha_A + \#(A=1)}{\alpha_A + \#(A=1) + \beta_A + \#(A=0)} \end{aligned}$$

For $P(d \mid a,b)$ the situation is more complex, the most convenient way is to specify a Beta prior for each one of the four components of θ_D. Lets focus on $P(D=1 \mid A=1,B=0)$, notice the parameters α and β we used before now depend on a and b, for this reason we are using $\alpha_D(a,b)$ and $\beta_D(a,b)$ as prior parameters, these are called *hyperparameters*.

$$P(\theta_2) = B(\theta_2 \mid \alpha_D(1,0) + \#(D=1,A=1,B=0), \beta_D(1,0) + \#(D=0,A=1,B=0))$$

As before we got that

$$P(D=1 \mid A=1,B=0,\mathcal{V}) = \frac{\alpha_D(1,0) + \#(D=1,A=1,B=0)}{\alpha_D(1,0) + \beta_D(1,0) + \#(A=1,B=0)}$$

In case we had no preference, we could set all hyperparameters to the same value, and, a complete ignorance prior would correspond to set them to 1.

Let now consider two limit possibilities, the one where we have no data at all, and the one where we have infinite data.

In case we have no data, the marginal probability corresponds to the prior which in the last case is

$$P(D=1 \mid A=1,B=0,\mathcal{V}) = \frac{\alpha_D(1,0)}{\alpha_D(1,0) + \beta_D(1,0)}$$

Note that equal hyperparameters would give a result of 0.5.

When infinite data is available, the marginal is generally dominated by it, this corresponds to the Maximum Likelihood solution.

$$P(D=1 \mid A=1,B=0,\mathcal{V}) = \frac{\#(D=1,A=1,B=0)}{\#(A=1,B=0)}$$

This happens unless the prior has a pathologically strong effect.

Consider the data given in the table in figure 9, and equal parameters and hyperparameters 1. Then we can compute the differences between this and using the Maximum Likelihood technique.

$$P(A=1 \mid \mathcal{V}) = \frac{1 + \#(A=1)}{2+N} = \frac{5}{9} \approx 0.556$$

By comparison, the Maximum Likelihood result is $4/7 = 0.571$, the Bayesian result is more prudent than this one, which fits in with our belief that any setting is equally probable i.e 0.5.

8.2 LEARNING DISCRETE VARIABLES

The natural generalization to more than two states is using a Dirichlet distribution as prior, assuming i.i.d data and local and global prior independence. We are considering two different scenarios, firstly one where the variable has no parents, as the case for A and B in the previous example. Secondly, we will consider a variable with a non void set of parents, as in the case with the disease D.

8.2.1 *No parents*

Consider a variable X with $Dom(X) = \{1, \dots, I\}$, $\theta = (\theta_1, \dots, \theta_I)$, then

$$P(x \mid \theta) = \prod_{i=1}^{I} \theta_i^{\mathbb{I}[x=i]} \text{ with } \sum_{i=1}^{I} \theta_i = 1$$

So that the posterior (considering N observations of the variable $(x_1, \dots, x_N) = \mathcal{V}$) is

$$P(\theta \mid x_1, \dots, x_N) = \frac{1}{P(\mathcal{V})} P(\theta) \prod_{n=1}^{N} \prod_{i=1}^{I} \theta_i^{\mathbb{I}[x_n=i]} = \frac{1}{P(\mathcal{V})} P(\theta) \prod_{i=1}^{I} \theta_i^{\sum_n \mathbb{I}[x_n=i]}$$

Then assuming a Dirichlet prior with hyperparameters $\boldsymbol{u} = (u_1, \dots, u_I)$

$$P(\theta) = \frac{1}{B(\boldsymbol{u})} \prod_{i=1}^{I} \theta_i^{u_i - 1} \implies P(\theta \mid \mathcal{V}) = \frac{1}{B(\boldsymbol{u}) P(\mathcal{V})} \prod_{i=1}^{I} \theta_i^{u_i - 1 + \sum_n \mathbb{I}[x_n=i]}$$

Which means that, defining $\boldsymbol{c} = (\sum_{n=1}^{N} \mathbb{I}[x_n = i])_{i=1,\dots,I}$

$$P(\theta \mid \mathcal{V}) \sim \text{Dirichlet}(\boldsymbol{u} + \boldsymbol{c})$$

Remark 3. Summarizing the above information, we just proved that the Dirichlet distribution is the conjugate prior of the Categorical Distribution.

The marginal is then given by

$$\begin{aligned} P(X = i \mid \mathcal{V}) &= \int_{\theta} P(X = i \mid \theta) P(\theta \mid \mathcal{V}) = \int_{\theta} \theta_i P(\theta \mid \mathcal{V}) \\ &= \int_{\theta_i} \theta_i P(\theta_i \mid \mathcal{V}) = \mathbb{E}\left[\theta_i \mid \mathcal{V}\right] \end{aligned}$$

Where we used that

$$\int_{\theta_{j \neq i}} \theta_i P(\theta \mid \mathcal{V}) = \theta_i \prod_{k \neq j} P(\theta_k \mid V) \int_{\theta_j} P(\theta_j \mid \mathcal{V}) = \theta_i \prod_{k \neq j} P(\theta_k \mid \mathcal{V})$$

As we already know from Proposition 2, the univariate marginal of a Dirichlet distribution is a Beta Distribution, then

$$(\theta_i \mid \mathcal{V}) \sim \text{Beta}(u_i + c_i, \sum_{j \neq i} u_j + c_j)$$

So the marginal is

$$P(X = i \mid \mathcal{V}) = \frac{u_i + c_i}{\sum_j u_j + c_j}$$

8.2.2 *Parents*

Consider now that X has a set of parent variables $pa(X)$, in this case, we want to compute the marginal given a state of its parents and the data

$$P(X = i \mid pa(X) = \boldsymbol{j}, \mathcal{V})$$

Let set the following notation for the parameters

$$P(X = i \mid pa(X) = \boldsymbol{j}, \theta) = \theta_{i,\boldsymbol{j}} \qquad \boldsymbol{\theta_j} = (\theta_{1,\boldsymbol{j}}, \dots, \theta_{I,\boldsymbol{j}})$$

Local independence means that

$$P(\boldsymbol{\theta}) = \prod_j P(\boldsymbol{\theta}_j)$$

As we did before, we consider a Dirichlet prior

$$\boldsymbol{\theta_j} \sim Dirichlet(\boldsymbol{u_j})$$

the posterior is then

$$\begin{aligned}
P(\boldsymbol{\theta} \mid \mathcal{V}) &= \frac{P(\boldsymbol{\theta})P(\mathcal{V} \mid \boldsymbol{\theta})}{P(\mathcal{V})} = \frac{1}{P(\mathcal{V})}\Big(\prod_{\boldsymbol{j}} P(\boldsymbol{\theta}_j)\Big)P(\mathcal{V} \mid \boldsymbol{\theta}) \\
&= \frac{1}{P(\mathcal{V})}\Big(\prod_{\boldsymbol{j}} \frac{1}{B(\boldsymbol{u_j})}\prod_i \theta_{i,\boldsymbol{j}}^{u_{i,\boldsymbol{j}}-1}\Big)P(\mathcal{V} \mid \boldsymbol{\theta}) \\
&= \frac{1}{P(\mathcal{V})}\Big(\prod_{\boldsymbol{j}} \frac{1}{B(\boldsymbol{u_j})}\prod_i \theta_{i,\boldsymbol{j}}^{u_{i,\boldsymbol{j}}-1}\Big)\Big(\prod_n\prod_{\boldsymbol{j}}\prod_i \theta_{i,\boldsymbol{j}}^{\mathbb{I}[x_n=i,pa(x_n)=\boldsymbol{j}]}\Big) \\
&= \frac{1}{P(\mathcal{V})}\prod_{\boldsymbol{j}} \frac{1}{B(\boldsymbol{u_j})}\prod_i \theta_{i,\boldsymbol{j}}^{u_{i,\boldsymbol{j}}-1+\#(X=i,pa(X)=\boldsymbol{j})}
\end{aligned}$$

Naming $\boldsymbol{v_j} = \boldsymbol{u_j} + \#(X = i, pa(X) = \boldsymbol{j})$, the posterior is

$$(\boldsymbol{\theta} \mid \mathcal{V}) \sim \prod_j \text{Dirichlet}(\boldsymbol{v_j})$$

Noting $v_{i,j}$ the components of $\boldsymbol{v_j}$, the marginal is then

$$P(X = i, pa(X) = \boldsymbol{j}, \mathcal{V}) = \frac{v_{i,j}}{\sum_i v_{i,j}}$$

Notice all the above has been done using a fixed variable X, so that all the parameters depend on that variable.

Using the above calculations, we can define the data likelihood under a model, usually called the *model likelihood*

$$\begin{aligned}
P(\mathcal{V} \mid \mathcal{M}) &= \prod_x\prod_n P(x_n \mid pa(x_n), \mathcal{M}) = \prod_x \int_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) \prod_n P(x_n \mid pa(x_n), \boldsymbol{\theta}, \mathcal{M}) \\
&= \prod_x\prod_{\boldsymbol{j}} \frac{1}{B(\boldsymbol{u_j})} \int_{\boldsymbol{\theta}} \prod_i \theta_{i,\boldsymbol{j}}^{u_{i,\boldsymbol{j}}-1+\#(x=i,pa(x)=\boldsymbol{j})} \\
&= \prod_x\prod_j \frac{B(\boldsymbol{v_j})}{B(\boldsymbol{u_j})}
\end{aligned}$$

9

STRUCTURE LEARNING

So far, both data and the BN have given to us. However, the BN structure is not always given and may be learned also form the data. Even considering complete data (no missing observations), there are some problems that need to be taken into account.

- The number of Belief networks is exponential over the number of variables so a brute force algorithm would not be viable.
- Testing dependencies requires a large amount of data. So a threshold must be set to measure when a dependence is significant.
- A Belief Network or a Markov Network may not be enough to represent the data due to the existence of unobserved variables.

Algorithm 1: PC Algorithm

Data: Complete undirected graph G, with vertices $\mathcal{V}$
Result: G with removed links
$i = 0$;
while *all nodes have* $\leq i$ *neighbors* **do**
  for $X \in \mathcal{V}$ **do**
    for $Y \in ne(x)$ **do**
      if $\exists S \subset ne(X) \backslash Y$ *such that* $\#S = i$ *and* $X \perp\!\!\!\perp Y \mid S$ **then**
        Remove $X - Y$ from G;
        $S_{XY} = S$;
      end
    end
  end
  $i = i + 1$;
end

9.1 PC ALGORITHM

An approach to learn the structure is the PC algorithm, it begins with a complete graph G and tries to remove as many links as possible studying the independence of the variables.

The algorithm 2 iterates over a natural counter, ending when it gets bigger than all existent neighborhoods. It chooses a linked pair of variables $X - Y$ and a subset $S_{XY} \subset ne(X)$,

following it has the desired size and $Y \notin S_{XY}$. If $X \perp\!\!\!\perp Y \mid S$, then the link is removed and S_{XY} is stored. The main idea behind this is the set of independencies is faithful to a graph then there is no link between two nodes X and Y if and only if there exists a subset of $ne(X)$ such that they are independent given this subset.

When this process ends, the undirected graph may be constructed following one rule, for any undirected link $X - Y - Z$, if $Y \notin S_{XZ}$ then set $X \to Y \leftarrow Z$. The rest of links may oriented arbitrarily not creating cycles or colliders. The reasoning behind this is using the d-separation theorem 2, if $Y \in S_{XZ}$ and $X \perp\!\!\!\perp Z \mid S_{XZ}$ then we want S_{XZ} to d-separate them, that is, using any configuration that doesn't create a collider in S_{XZ}. On the other hand if $Y \notin S_{XZ}$ then $X \not\perp\!\!\!\perp Z \mid Y$ so Y d-connect them, to get this we set it as a collider.

9.2 INDEPENDENCE LEARNING

Our main concern now is given three variables X, Y, Z to measure $X \perp\!\!\!\perp Y \mid Z$. One approach is to measure the empirical *conditional mutual information* of the variables.

Definition 44. Given two random variables X, Y, we define their *mutual information* as the Kullback-Leibler divergence of their joint distribution and the product of their marginals and

$$MI(X;Y) = KL(P_{X,Y} \mid P_X P_Y)$$

Definition 45. Given three random variables X, Y, Z we define the *conditional mutual information* of X and Y over Z as

$$MI(X;Y \mid Z) = \mathbb{E}_Z\Big[KL(P_{X,Y|Z} \mid P_{X|Z} P_{Y|Z})\Big]$$

Where $MI(X;Y \mid Z) \geq 0$ and $MI(X;Y \mid Z) = 0 \iff P_{X,Y|Z} = P_{X|Z} P_{Y|Z} \iff X \perp\!\!\!\perp Y \mid Z$. We can estimate this using the empirical distributions, however, this *empirical* mutual information will be typically greater than 0 even when $X \perp\!\!\!\perp Y \mid Z$, therefore a threshold must be established.

A Bayesian approach would be comparing the model likelihood under independence and dependence hypothesis. That is computing the model likelihood for the below joint distributions assuming local and global parameter independence

$$P_{indep}(x,y,z) = P(x \mid z, \theta_1) P(y \mid z, \theta_2) P(z \mid \theta_3) P(\theta_1) P(\theta_2) P(\theta_3)$$

$$P_{dep}(x,y,z) = P(x,y,z \mid \theta) P(\theta)$$

10

LEARNING WITH MISSING VARIABLES AND DATA

Until this moment we have assume that the data we are given is completed but in practice this data is not in two different ways. There may be unobserved or *hidden* variables that affect the visible ones, and there may be *missing* information, that is, states of visible variables that are missing.

Think about the example with the disease and the two habits we used in the last section, missing data would be a row in the table where some entry is missing, for example $x_3 = \{D = 1, A = 1\}$, where we know that this person got the disease and had habit A but we have no information about habit B, this is an example of *missing* data.

One approach to handle this situation would be marginalizing over that variable

$$P(x_3 \mid \theta) = \int_b P(d_3, a_3, b \mid \theta) = P(a_3 \mid \theta_A) \int_b P(b \mid \theta_B) P(d_3 \mid b, a_3, \theta_D)$$

This leads to a non-factorized form of the posterior which is computationally difficult to handle, notice the problem is not conceptual but computational. Using the marginal to handle missing information does not always lead to this situation, in fact, marginalizing over a collider (D in our example) would mean loosing that variable as the integral simply equals 1.

$$P(x_3 \mid \theta) = \int_d P(d, a_3, b_3 \mid \theta) = P(a_3 \mid \theta_A) P(b_3 \mid \theta_B) \int_d P(d \mid b_3, a_3, \theta_D) = P(a_3 \mid \theta_A) P(b_3 \mid \theta_B)$$

There are three main types of missing data:

- **Missing completely at random (MCAR)**. If the events that lead to any particular data to be missing is independent from both the observed and the unobserved variables, and occur at random.
- **Missing at random (MAR)**. When the absence is not random but can be explained with the observed variables.
- **Missing not at random (MNAR)**. The missing data is related with the reason why it is missing. For example, skipping a question in a survey for being ashamed of the answer.

To express this mathematically, split the variables $\mathcal{X}$ into visible $\mathcal{X}_{vis}$ and hidden $\mathcal{X}_{hid}$, let M be a variable denoting that the state of the hidden variables is known (0) or unknown (1). So the difference between the three types resides on how $P(M = 1 \mid x_{vis}, x_{hid}, \theta)$ simplifies.

When data is *missing at random*, we assume that we can explain the missing information with the visible one, so the probability of being missing only depends on the visible data, that is

$$P(M = 1 \mid x_{vis}, x_{hid}, \theta) = P(M = 1 \mid x_{vis})$$

so that,

$$P(x_{vis}, M = 1 \mid \theta) = P(M = 1 \mid x_{vis})P(x_{vis} \mid \theta)$$

Assuming the data is *missing completely at random* is stronger, as we are supposing that there is no reason behind the missing data, so that it being missing is independent from the visible and hidden data.

$$P(M = 1 \mid x_{vis}, x_{hid}, \theta) = P(M = 1)$$

so now

$$P(x_{vis}, M = 1 \mid \theta) = P(M = 1)P(x_{vis} \mid \theta)$$

In both cases we may simply use the marginal $P(x_{vis} \mid \theta)$ to assess parameters as $P(x_{vis}, M = 1 \mid \theta)$ does not depend on the missing variables.

In case data is *missing not at random* then no independence assumption is made over the probability of the data being unknown, meaning it depends on both the visible and the hidden information. From now on, we will assume missing information is either MAR or MCAR, even though this could lead to a misunderstanding of the problem as in the following example.

Example 4. Consider a situation where data is obtained from a survey where people are asked to choose between 3 options A, B and C. Assume that no one chose option C because they are ashamed of the answer, and the answers are uniform between A, B and not answering.

Normalizing the missing information would lead to setting $P(A \mid \mathcal{V}) = 0.5 = P(B \mid \mathcal{V})$ and $P(C \mid \mathcal{V}) = 0$ when the reasonable result is that not answering equals to choosing C so that $P(A \mid V) = P(B \mid \mathcal{V}) = P(C \mid \mathcal{V}) = \frac{1}{3}$

10.1 EXPECTATION MAXIMIZATION

The *expectation maximization* algorithm is an iterative method to find maximum likelihood or maximum a posteriori estimates of parameters in statistical models, where the model depends on hidden variables. The main idea is to set a lower bound to the marginal likelihood, then using an iterative method to increase this lower bound.

10.1.1 *General case*

Consider we have only two variables, one visible V and one hidden H. Consider also the Kullback-Leibler divergence between a 'variational' distribution $Q(h \mid v)$ (where variational

means that this distribution is the object of an optimization problem) and a parametric one $P(h \mid v, \theta)$.

$$\begin{aligned} KL(Q(h \mid v) \mid P(h \mid v, \theta)) &= \mathbb{E}_Q\Big[\log Q(h \mid v) - \log P(h \mid v, \theta)\Big] \\ &= \mathbb{E}_Q\Big[\log Q(h \mid v)\Big] - \mathbb{E}_Q\Big[\log P(h, v \mid \theta)\Big] + \mathbb{E}_Q\Big[\log P(v \mid \theta)\Big] \\ &= \mathbb{E}_Q\Big[\log Q(h \mid v)\Big] - \mathbb{E}_Q\Big[\log P(h, v \mid \theta)\Big] + \log P(v \mid \theta) \geq 0 \end{aligned}$$

We got then a lower bound to $\log P(v \mid \theta)$

$$\log P(v \mid \theta) \geq \underbrace{-\mathbb{E}_Q\Big[\log Q(h \mid v)\Big]}_{\text{Entropy}} + \underbrace{\mathbb{E}_Q\Big[\log P(h, v \mid \theta)\Big]}_{\text{Energy}} \text{ [1]}$$

Assume a set of observations of the visible data, $\mathcal{V} = \{v_1, \ldots, v_N\}$ and the states of the hidden variables in that observations $\{h_1, \ldots, h_N\}$, notice that the values on this last set are unknown but they exists. Then has the variables of the observations are i.i.d we got that

$$\log P(\mathcal{V} \mid \theta) = \sum_{n=1}^{N} \log P(v_n \mid \theta) \geq \sum_{n=1}^{N} -\mathbb{E}_Q\Big[\log Q(h_n \mid v_n)\Big] + \mathbb{E}_Q\Big[\log P(h_n, v_n \mid \theta)\Big]$$

And we know equality holds if and only if $Q(h_n \mid v_n) = P(h_n \mid v_n, \theta)\ \forall n = 1, \ldots, N$.

This suggest the following iterative procedure to optimize the parameter θ consisted in two steps.

- **E-step**. For a fixed θ, find the distributions that maximize the above bound, i.e, choose $Q(h_n \mid v_n) = P(h_n \mid v_n, \theta)$.
- **M-step**. For a fixed distribution Q, find the parameter θ that maximizes the bound. Since Q does not depend on the parameter, this is equivalent to maximize the energy term.

Example 5. Consider a single variable V with $Dom(V) = \mathbb{R}$ and a single hidden variable H with $dom(H) = \{1, 2\}$. Consider the model

$$P(v \mid h, \theta) = \frac{1}{\sqrt{\pi}} e^{-(v - \theta h)^2}$$

and $P(H = 1) = P(H = 2) = 0.5$. Suppose an observation $v = 2.75$ we want to optimize the parameter θ in

$$P(V = 2.75 \mid \theta) = \int_h P(V = 2.75 \mid h, \theta) P(h) = \frac{1}{2\sqrt{\pi}} \left(e^{-(2.75 - \theta)^2} + e^{-(2.75 - 2\theta)^2}\right)$$

The lower bound given to the log likelihood is

$$\log P(v \mid \theta) \geq -Q(1) \log Q(1) - Q(2) \log Q(2) - \mathbb{E}_Q\Big[(v - \theta h)^2\Big] + \text{const.}$$

1 This terms come from a statistical physics terminology

The M-step can be done analytically, noticing that due to the negative sign we want to minimize $\mathbb{E}_Q\Big[(v-\theta h)^2\Big]$

$$\frac{d}{d\theta}\mathbb{E}_Q\Big[(v-\theta h)^2\Big] = \mathbb{E}_Q\Big[2vh+2\theta h^2\Big] = 2v\mathbb{E}_Q\Big[h\Big]+2\theta\mathbb{E}_Q\Big[h^2\Big] = 0 \iff \theta = \frac{v\mathbb{E}_Q\Big[h\Big]}{\mathbb{E}_Q\Big[h^2\Big]}$$

$$\frac{d^2}{d^2\theta}\mathbb{E}_Q\Big[(v-\theta h)^2\Big] = 2\mathbb{E}_Q\Big[h^2\Big] \geq 0$$

so the new parameter optimal parameter is

$$\theta_{new} = v\frac{\mathbb{E}_Q\Big[H\Big]}{\mathbb{E}_Q\Big[H^2\Big]}$$

The E-step would set $Q_{new}(h) = P(h \mid v, \theta)$, in this case

$$Q_{new}(h) = \frac{P(V=2.75 \mid h,\theta)P(H=2)}{P(V=2.75)} = \frac{e^{-(2.75-h\theta)}}{e^{-(2.75-\theta)}+e^{-(2.75-2\theta)}}$$

Algorithm 2: Expectation Maximization Algorithm

Data: A distribution $P(x \mid \theta)$ and a dataset $\mathcal{V}$. Where X splits in visible variables V and hidden variables H

Result: Parameter θ that maximizes the likelihood

while *Convergence stop criteria* **do**

  for $n \in 1,\ldots,N$ **do**

    $Q(h \mid v_n) = P(h \mid v_n, \theta)$;

  end

  $\theta = \arg\max_\theta \sum_{n=1}^{N} \mathbb{E}_{Q(h|v_n)}\Big[\log P(h \mid v_n, \theta)\Big]$;

end

return θ;

It is clear that the EM algorithm does increase the lower bound in each iteration but we would like it to increase not only the bound but also the marginal likelihood. We will be using a single data point as it easy holds by summation when using the full dataset.

Let θ^{new} be the value of the parameter after one iteration, we now that Q is set to

$$Q(h \mid v_n) = P(h \mid v_n, \theta)$$

So the lower bound in terms of θ and θ^{new} is

$$LB(\theta^{new} \mid \theta) = \underbrace{-\mathbb{E}_{P(h|v,\theta)}\Big[\log P(h \mid v, \theta)\Big]}_{\text{Entropy}} + \underbrace{\mathbb{E}_{P(h|v,\theta)}\Big[\log P(h, v \mid \theta^{new})\Big]}_{\text{Energy}}$$

From the definition of the Kullback-Leibler we get that

$$\log P(v \mid \theta^{new}) = LB(\theta^{new} \mid \theta) + KL\big(P(h \mid v, \theta) \mid P(h \mid v, \theta^{new})\big)$$

We could use θ in the above formula getting

$$\log P(v \mid \theta) = LB(\theta \mid \theta) + KL\big(P(h \mid v, \theta) \mid P(h \mid v, \theta)\big) = LB(\theta \mid \theta)$$

So we can compute the difference between the log likelihood between two consecutive iterations as

$$\log P(v \mid \theta^{new}) - \log P(v \mid \theta) = LB(\theta^{new} \mid \theta) - LB(\theta \mid \theta) + KL\big(P(h \mid v, \theta) \mid P(h \mid v, \theta^{new})\big)$$

Where we know the last term is always positive, about the difference of bounds, the M-step ensures the new parameter makes the lower bound higher or equal to the current one, so that difference is also positive.

10.1.2 *Belief Networks case*

As we did before let $\mathcal{X} = (\mathcal{V}, \mathcal{H}) = \{X_1, \ldots, X_M\} = \{(V_1, H_1), \ldots, (V_M, H_M)\}$ be the set of variables partitioned in visible and hidden. Let $\mathcal{D} = \{v^i, \ldots, v^N\}$ be the set of observations and $\{h^1, \ldots, h^N\}$ the corresponding values of the hidden variables.

The *energy term* in a Bayesian networks has the form

$$\sum_{n=1}^{N} \mathbb{E}_{Q(h^n \mid v^n)}\Big[\log P(x^n \mid \theta)\Big] = \sum_{n=1}^{N} \sum_{i=0}^{M} \mathbb{E}_{Q(h^n \mid v^n)}\Big[\log P(x_i^n \mid pa(x_i^n, \theta))\Big]$$

It is useful to use the following notation that defines a conditional distribution of the hidden variable when the visible one equals v^n.

$$Q^n(x) = Q^n(v, h) = Q(h \mid v^n)\mathbb{I}(v = v^n)$$

We can define the mixture distribution

$$Q(x) = \frac{1}{N} \sum_{n=1}^{N} Q^n(x)$$

Then we have that

$$\begin{aligned}
\mathbb{E}_{Q(x)}\Big[\log P(x \mid \theta)\Big] &= \int_x Q(x) \log P(x \mid \theta) = \int_x \frac{1}{N} \sum_{n=1}^{N} Q^n(x) \log P(x \mid \theta) \\
&= \frac{1}{N} \int_x \sum_{n=1}^{N} Q(h \mid v^n)\mathbb{I}[v = v^n] \log P(x \mid \theta) \\
&= \frac{1}{N} \sum_{n=1}^{N} \mathbb{E}_{Q(h \mid v^n)}\Big[\log P(x^n \mid \theta)\Big]
\end{aligned}$$

Using the Belief Network structure

$$\begin{aligned}\mathbb{E}_{Q(x)}\Big[\log P(x \mid \theta)\Big] &= \sum_{i=1}^{M} \mathbb{E}_{Q(x)}\Big[\log P(x_i \mid pa(x_i, \theta))\Big] \\ &= \sum_{i=1}^{M} \int_x Q(x) \log P(x_i \mid pa(x_i, \theta)) \\ &= \sum_{i=1}^{M} \mathbb{E}_{Q(pa(x_i), theta)}\Big[\mathbb{E}_{Q(x_i|pa(x_i),\theta)}\Big[\log P(x_i \mid pa(x_i), \theta)\Big]\Big]\end{aligned}$$

We add a constant to the last term so it comes with the structure of a Kullback-Leibler Divergence (notice it has its sign has changed)

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

$$= \sum_{i=1}^{M} E_{Q(pa(x_i))}\Big[KL\Big(Q(x_i \mid pa(x_i), \theta) \mid P(x_i \mid pa(x_i), \theta)\Big)\Big]$$

So maximizing the energy term is equivalent to minimize the above formula, that is, setting

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

So the first observation is that θ is not needed in order to maximize the energy term due to the Belief Network structure. The second one is that storing the full $Q(x)$ on each iteration is not needed as only the distribution on the family of each variable X_i is required by the M-step.

The M-step is then equivalent to set

$$P(x_i \mid pa(x_i)) = Q(x_i \mid pa(x_i)) = \frac{\sum_{n=1}^{N} Q^n(x_i, pa(x_i))}{\sum_{n=1}^{N} Q^n(pa(x_i))}$$

10.2 EM EXTENSIONS

10.2.1 *Partial steps*

Making a partial M-step consist on not using the optimal parameter for the energy term, but using one with just higher energy. Finding this values can be easier than finding the optimal one and convergence still follows as the only requirement to make the likelihood increase was to increase the lower bound.

On the other hand, when studying the increase on the likelihood, we supposed that the optimal E-step was being used. In general, we cannot guarantee that a partial step would increase the likelihood. In fact, it is guaranteed to increase the lower bound, but nothing can be said about the log likelihood.

Another important factor is, that the EM algorithm assumes that the energy term is possible to calculate, which may not be. As an approach to solve this situation, we can set a clsas of

distributions $\mathcal{Q}$, and minimize the KullBack-Leibler divergence between $P(h \mid v, \theta_1)$ and a distribution $Q \in \mathcal{Q}$, so we pick a distribution such that

$$Q = \arg\min_{Q \in \mathcal{Q}} KL\big(Q(h \mid v, \theta_2) \mid P(h \mid v, \theta_1)\big)$$

An extreme case is to choose $\mathcal{Q}$ as delta functions, where the energy term is now a constant. And the optimal chose setting is

$$Q(h^n \mid v^n) = \delta(h^n, h^n_{opt}) \qquad\qquad h^n_{opt} = \arg\max_h P(h, v^n \mid \theta)$$

This is called *Viterbi training* and does not guarantee that the log likelihood is being increased in each iteration.

10.3 VARIATIONAL BAYES

Another method to deal with hidden variables is using *Variational Bayes (VB)*, in contrast with the EM algorithm, this one uses a distribution that better represents the posterior than the one using a Maximum Likelihood approach.

Consider a simple datapoint $x = (v, h)$, in this situation we focus our interest on the posterior distribution.

$$P(\theta, v) = \frac{P(v \mid \theta)P(\theta)}{P(v)} = \frac{1}{P(v)} \int_h P(v, h \mid \theta)P(\theta)$$

Variational Bayes tries to factorize the joint hidden a parameter posterior, i.e, computes two distributions, one over the hidden variable and one over the parameter such that

$$P(h, \theta \mid v) \approx Q(h)Q(\theta)^{2}$$

To achieve that, we minimize the Kullback-Leibler divergence between them.

$$KL\Big(Q(h)Q(\theta) \mid P(h, \theta \mid v)\Big) = \mathbb{E}_{Q(h)}\Big[\log Q(h)\Big] + \mathbb{E}_{Q(\theta)}\Big[\log Q(\theta)\Big] - \mathbb{E}_{Q(h)Q(\theta)}\Big[\log P(h, \theta \mid v)\Big] \geq 0$$

We use that $\log P(v)$ is independent from $Q(h)Q(\theta)$ to get the desired inequality.

$$\begin{aligned}
0 &\leq \mathbb{E}_{Q(h)}\Big[\log Q(h)\Big] + \mathbb{E}_{Q(\theta)}\Big[\log Q(\theta)\Big] - \mathbb{E}_{Q(h)Q(\theta)}\Big[\log P(h, \theta \mid v)\Big] \\
&= \mathbb{E}_{Q(h)}\Big[\log Q(h)\Big] + \mathbb{E}_{Q(\theta)}\Big[\log Q(\theta)\Big] - \mathbb{E}_{Q(h)Q(\theta)}\Big[\log P(h, \theta, v)P(v)\Big] \\
&= \mathbb{E}_{Q(h)}\Big[\log Q(h)\Big] + \mathbb{E}_{Q(\theta)}\Big[\log Q(\theta)\Big] - \mathbb{E}_{Q(h)Q(\theta)}\Big[\log P(h, \theta, v)\Big] - \mathbb{E}_{Q(h)Q(\theta)}\Big[\log P(v)\Big] \\
&= \mathbb{E}_{Q(h)}\Big[\log Q(h)\Big] + \mathbb{E}_{Q(\theta)}\Big[\log Q(\theta)\Big] - \mathbb{E}_{Q(h)Q(\theta)}\Big[\log P(h, \theta, v)\Big] - \log P(v)
\end{aligned}$$

We got the following lower bound

$$\log P(v) \geq -\mathbb{E}_{Q(h)}\Big[\log Q(h)\Big] - \mathbb{E}_{Q(\theta)}\Big[\log Q(\theta)\Big] + \mathbb{E}_{Q(h)Q(\theta)}\Big[\log P(h, \theta, v)\Big]$$

2 We use the same letter for both distributions, as they can be differenced from the context.

So that minimizing the Kullback-Leibler divergence is equivalent to find the tightest lower bound.

The procedure is then split in two steps to keep the structure of the EM algorithm.

- **E-step**. Given a fixed $Q(\theta)$, minimize the Kullback-Leibler divergence.

$$Q^{new}(h) = \arg\min_{Q(h)} KL\Big(Q(h)Q(\theta) \mid P(h, \theta \mid v)\Big)$$

- **M-step**. Given a fixed $Q(h)$, minimize the Kullback-Leibler divergence.

$$Q^{new}(\theta) = \arg\min_{Q(\theta)} KL\Big(Q(h)Q(\theta) \mid P(h, \theta \mid v)\Big)$$

As in the case of the EM algorithm, each iterations guarantees an increase in the lower bound of the marginal likelihood, but increasing the marginal likelihood itself is not guaranteed.

When using an i.i.d dataset $(\mathcal{V}, \mathcal{H})$, we may assume that $Q(\mathcal{H})$ can be factorized:

$$Q(h_1, \ldots, h_N) = \prod_{n=1}^{N} Q(h_n)$$

The lower bound to the marginal likelihood is then written as a summation of the bounds on each datapoint.

$$\log P(\mathcal{V}, \theta) = \sum_n - \mathbb{E}_{Q(h_n)}\Big[\log Q(h_n)\Big] - \mathbb{E}_{Q(\theta)}\Big[\log Q(\theta)\Big] + \mathbb{E}_{Q(h_n)Q(\theta)}\Big[\log P(v_n, h_n, \theta)\Big]$$

10.3.1 *VB is a generalization of the EM algorithm*

We start considering a distribution over the parameter that sumarize the information in the optimal point, let θ_{opt} be the optimal value of θ.

$$Q(\theta) = \delta(\theta - \theta_{opt})$$

The lower bound takes the form

$$\log P(v \mid \theta_{opt}) \geq -\mathbb{E}_{Q(h)}\Big[\log Q(h)\Big] + \mathbb{E}_{Q(h)}\Big[\log P(h, v, \theta_{opt})\Big] + \text{ const.}$$

The M-step is then picking the optimal parameter θ_{opt} given a fixed $Q(h)$:

$$\begin{aligned}
\theta_{opt} &= \arg\max_{\theta} \Big(\mathbb{E}_{Q(h)}\Big[\log P(v, h, \theta)\Big]\Big) \\
&= \arg\max_{\theta} \Big(\mathbb{E}_{Q(h)}\Big[\log P(v \mid h, \theta) P(h \mid \theta) P(\theta)\Big]\Big) \\
&= \arg\max_{\theta} \Big(\mathbb{E}_{Q(h)}\Big[\log P(v \mid h, \theta) P(h \mid \theta)\Big] + \log P(\theta)\Big)
\end{aligned}$$

If we take a flat prior ($P(\theta)$ constant), this term is equivalent to the energy one in the EM bound $\mathbb{E}_{Q(h|v)}\Big[\log P(h, v \mid \theta)\Big]$.

The VB E-step consists on minimizing $KL\Big(Q(h) \mid P(h, \theta_{opt} \mid v)\Big)$, as

$$P(h, \theta_{opt} \mid v) \propto P(h, \theta_{opt}, v) \propto P(h \mid \theta_{opt}, h)$$

this is equivalent to the E-step of the EM algorithm which consisted on minimizing the Kullback-Leibler divergence $KL\Big(Q(h \mid v) \mid P(h \mid v, \theta)\Big)$.

BIBLIOGRAPHY

Barber, David. 2007. *Bayesian Reasoning and Machine Learning*. Cambridge University Press.

Farrow, Malcolm. 2008. MAS3301 Bayesian Statistics.

Koller, Daphne, & Friedman, Nir. 2009. *Probabilistic Graphical Models, Principles and Techniques*. The MIT Press.

Pearl, Judea, & Dechter, Rina. 2013. Identifying Independences in Casual Graphs with Feedback.

Shachter, Ross D. 2013. Bayes-Ball: The Rational Pastime.

Wainwright, Martin J., & Jordan, Michael I. 2008. *Graphical Models, Exponential Families and Variational Inference*. Now Publishers Inc.