Introduction

Consider a set of variables $X_1, \ldots, X_n$, we will define a model that will consider the joint probability density function of the variables

$$P(X_1, \ldots, X_n)$$

There are a few forms of representing this

- Bayesian Networks, which are represented by directed graphs in which each node represents conditional probabilities, and the edges represent dependencies. (Drawing from the blackboard)

- Markov Networks. In this case, we have undirected graphs, where the edges are *factors* (tables of probability)

Bayesian networks have the *factors* in the nodes, while Markov networks have the factors in the edges.

Our goal will be to make inference about variables using the available information. Sometimes, making inference is understood as **marginalizing the joint distribution**.

$$P(A) = \sum_{B,C} P(A, B, C) = \sum_{B,C} P(A)P(B|A)P(C|B, A)$$

There are different algorithms to compute this probabilities, such as *variable elimination* or *message-passing*.

Also, there are different ways of **reasoning**, such as

- Causal reasonament, which studies causalities
- Evidencial reasonament

With all this types of reasoning, we are assuming **two main points**:

- That we have all the information about both the structure of the network and the *factors* or probabilities.

Having both the network structure and the probabilities of each of the factors, we can marginalize to obtain the joint probability distributions. However, we can consider a case where we do **not** know the structure of the network but we know the probabilities of the factors. This is also a *branch* of study, which we will not go deep in. A last case is the one where we **know** the structure of the network, but we are **lacking** parts of the table which we would like to infere. This is the case that we will focus in this course.

Example

Consider the following random variable, modeling the probability of obtaining heads or tails in a coinflip

$$v^n = \begin{cases} 1 & \text{heads} \\ 0 & \text{tails} \end{cases}$$

However, our coin might have different weights for each of the outcomes (biased coin). For instance, consider that $\theta = P(\text{heads}) = P(v^n = 1|\theta)$. Hence, $P(\text{tails}) = 1 - \theta$. In this case, our goal would be to **determine** θ.

If we tossed the coin n times, and **we knew the probability** θ, the coin tosses would be **independent**. However, if we **do not know** the probability θ, the coin tosses **would be dependent** since the **outcome** of the experiment affects θ. (Diagrams from the blackboard). In this case, the joint probability would be

$$P(\theta, v^1, \ldots, v^n) = P(\theta) \prod_{i=1}^{n} P(v^i|\theta)$$

Estimations of the parameters are sometimes done using the empirical distribution function. However, there are other methods of estimating the joint pdf, such as *maximum likelihood estimation.*

Maximum likelihood estimation

We define the **likelihood** of the data as

$$L(\theta; D) = P(D|\theta) = \prod_{j=1}^{N} P(v^j|\theta)$$

Maximum likelihood estimation determines the likelihood function and tries to find (or approximate) a maximum of it.

We can generalize our previous coin toss example. Consider that we obtained M_h heads and M_t tails in our coin toss problem. In this case, our likelihood function would be

$$L(\theta; M_h, M_t) = \theta^{M_h}(1 - \theta)^{M_t}$$

The most common approach is to apply the logarithm to the likelihood function, which is a monotonous increasing function, to convert the product into sums, and then maximize the **log-likelihood**. In the previous example, our log-likelihood would be

$$\ell = \log L(\theta; M_h, M_t) = M_h \log \theta + M_t \log(1 - \theta)$$