DM n° 1 - Bayesian Reasoning *(deadline October 24)*

Exercice 1. *Bayes' Formula (Bayes 1763, Laplace 1812)*

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probabilisty space, let $A \in \mathscr{F}$ and $\{B_n, n \geq 0\}$ be a finite or countable partition of Ω into events in $\mathscr{F}$ of non null measures. Show that for all $n \geq 0$, the next formula is true :

$$\boxed{\mathbb{P}(B_n|A) = \frac{\mathbb{P}(A|B_n)\mathbb{P}(B_n)}{\sum_{i\geq 0}\mathbb{P}(A|B_i)\mathbb{P}(B_i)}}$$

In particular, given a partition $\{B, \overline{B}\}$, the formula becomes :

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A|B)\mathbb{P}(B)}{\mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|\overline{B})\mathbb{P}(\overline{B})}$$

Bayes' formula is sometimes called the "formula of inverted probabilies" or the "formula of causes", because it can be interpreted as a way to calculate the probability of a cause starting an observed effect. Such a flip of data initially in the form $\mathbb{P}$(effect|cause) has long been viewed with caution by statisticians of the nineteenth century.

Exercice 2. *Serial conditioning*

Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and A_1,...,A_n some events in $\mathscr{F}$. In this exercise and all the other ones, "$A_n, \ldots, A_1$" will denote the event $A_n \cap \cdots \cap A_1$.

1. Assuming that $\mathbb{P}(A_n, A_{n-1}, \ldots, A_1) > 0$, prove the next formula :

$$\boxed{\mathbb{P}(A_n, A_{n-1}, \ldots, A_1) = \mathbb{P}(A_n|A_{n-1}, \ldots, A_1)\mathbb{P}(A_{n-1}|A_{n-2}, \ldots, A_1)\cdots\mathbb{P}(A_2|A_1)\mathbb{P}(A_1)}$$

2. What about the next formula : is it true or can you find a counterexampe?

$$\mathbb{P}(A_n, A_{n-1}, \ldots, A_1) = \mathbb{P}(A_n|A_{n-1})\mathbb{P}(A_{n-1}|A_{n-2})\cdots\mathbb{P}(A_2|A_1)\mathbb{P}(A_1)\ ?$$

Exercice 3. *Traceability*

A factory produces smart cards with three machines A, B, C. Machine A provides 20% of the whole production, but 5% of the cards produced by A are defective. Machine B provides 30% of the whole production, but 4% of the cards produced by B are defective. Machine C provides 50% of the whole production, but 1% of the cards produced by C are defective.

What is the probability that a defective smart card come from A? from B? from C?

Exercice 4. *Bayesian networks (Pearl 1985)*

A ***bayesian network*** is a probabilistic model which consists of :

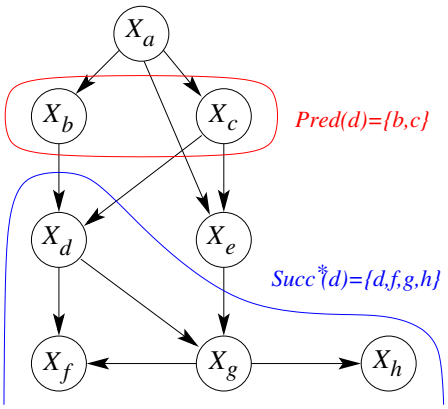
- a ***finite family of random variables*** $(X_v)_{v\in V}$ with discrete values in the respective spaces E_v, $v \in V$;
- an ***acyclic graph*** $G = (V, E)$ where each vertex v represents the random variable X_v, the direct predecessors of v are denoted $Pred(v) = \{w \in V | wv \in E\}$ and the vertices reachable from v are denoted $Succ^*(v) = \{w \in V | \exists$ chemin de v à $w\}$ (v is included) ;
- a ***dependence relation between random variables*** formalised by the graph in the following way : for all $v \in V$, X_v is *conditionally independent* from $(X_w)_{w\in V\setminus Succ^*(v)}$ *given* $(X_w)_{w\in Pred(v)}$, that is for any combination of values $x_v \in E_v$ and $x_w \in E_w$, $w \in V \setminus Succ^*(v)$,

  $$\mathbb{P}(X_v = x_v | \forall\, w \in V \setminus Succ^*(v),\ X_w = x_w) = \mathbb{P}(X_v = x_v | \forall\, w \in Pred(v),\ X_w = x_w),$$

  a lighter notation is $\boxed{\mathbb{P}(X_v|X_w,\ w \in V \setminus Succ^*(v)) = \mathbb{P}(X_v|X_w,\ w \in Pred(v))}$ which indicates that the equality is true for any set of values assigned to the random random variables (note that $Pred(v) \subseteq V \setminus Succ^*(v)$ since the graph is acyclic) ;

- a ***set of conditional probabilities*** $\mathbb{P}(X_v = x_v | \forall w \in Pred(v),\ X_w = x_w)$ for all $v \in V$ and any combination of values $x_v \in E_v$ and $x_w \in E_w$, $w \in Pred(v)$.

The next figure shows a bayesian network with $V = \{a,b,c,d,e,f,g,h\}$ and the random variables $(X_v)_{v\in V}$. As an example, the conditional independence relation associated with d is written $\mathbb{P}(X_d|X_a,X_b,X_c,X_e) = \mathbb{P}(X_d|X_b,X_c)$, this equality is true for any combination of values : suppose that the random variables have values in $\{0,1\}$, we have e.g. $\mathbb{P}(X_d = 1|X_a = 0,X_b = 1,X_c = 0,X_e = 1) = \mathbb{P}(X_d = 1|X_b = 1,X_c = 0)$.



1. Prove that, for any set of values x_v, $v \in V$, the probability that $(X_v)_{v\in V} = (x_v)_{v\in V}$ written $\mathbb{P}(X_v, v \in V)$ in a light way (with the convention $\mathbb{P}(X_v|\emptyset) = \mathbb{P}(X_v)$) satisfies :

$$\mathbb{P}(X_v, v \in V) = \prod_{v\in V} \mathbb{P}(X_v|X_w,\ w \in Pred(v))$$

A new virus is spreading through Internet. It can cause some memory overload of your servers, as well as a strong traffic increase. These two disruptions may lead to a shutdown of your whole system. Note that an increase of the traffic may also cause a major slowdown of response times.

2. Consider the following random variables with values in $\{\texttt{false},\texttt{true}\}$:
— I = "infected by the virus",
— S = "saturation of memory",
— T = "traffic increase",
— R = "response time slowdown",
— P = "system shutdown".
Draw the bayesian network matching the information you have.

3. From disclosed statistics about the virus and its effects, you dispose of the following numbers : $\mathbb{P}(I = \texttt{true}) = 0,2$, $\mathbb{P}(T = \texttt{true}|I = \texttt{true}) = 0,2$, $\mathbb{P}(T = \texttt{true}|I = \texttt{false}) = 0,05$, $\mathbb{P}(S = \texttt{true}|I = \texttt{true}) = 0,8$, $\mathbb{P}(S = \texttt{true}|I = \texttt{false}) = 0,2$, $\mathbb{P}(R = \texttt{true}|T = \texttt{true}) = 0,8$, $\mathbb{P}(R = \texttt{true}|T = \texttt{false}) = 0,6$, $\mathbb{P}(P = \texttt{true}|S = \texttt{true}, T = \texttt{true}) = \mathbb{P}(P = \texttt{true}|S = \texttt{true}, T = \texttt{false}) = \mathbb{P}(P = \texttt{true}|S = \texttt{false}, T = \texttt{true}) = 0,8$, $\mathbb{P}(P = \texttt{true}|S = \texttt{false}, T = \texttt{false}) = 0,05$.
Your system users are complaining about a slowdown of response times, nonetheless the global system has not crashed. Recent studies have estimated that approximatively 20% of systems are infected, i.e. $\mathbb{P}(I = \texttt{true}) = 20\%$. Compared to this a priori probability, do you think that your system has a higher or a lower chance to be really infected?

Exercice 5. *Computational complexity of bayesian inference (Cooper 1990)*

The *inference* problem for bayesian networks is stated as follows :

INPUT : a bayesian network described as Exercice 5, two subsets of random variables indexed by $S, T \subseteq V$, a set of values $x_v \in E_v$ for $v \in S \cup T$. OUTPUT : compute $\mathbb{P}(X_t = x_t,\ t \in T \mid X_s = x_s,\ s \in S)$.

1. Suppose that all the spaces E_v of values are finite, describe an algorithm which solves this inference problem. Is it a polynomial time algorithm? If YES, contact me as soon as possible. If NO, try the next question.

2. [Tiebreaker] To show that this inference problem is rapidly complex, one can narrow the study to restricted inputs where $\forall v \in V$, $E_v = \{\texttt{false},\texttt{true}\}$, and $|T| = 1$, and the output question is $\mathbb{P}(X_t = x_t \mid X_s = x_s,\ s \in S) > 0$? Prove that even with these simplifications, this decision problem is NP-complete.
Suggestion : design a reduction from 3-SAT by building a kind of boolean circuit equivalent to the input logical formula.

Exercice 6. *Alternative facts*

After an anonymous reporting, a person has been accused of murder, shaking up his quiet town of 1000000 inhabitants. The culprit has left DNA traces and a test has shown that these traces correspond to the DNA of the accused. During the trial, the prosecutor calls an expert who explains that " if the accused is innocent, there is 1 chance out of 100000 that his DNA matches the one found at the crime scene". Addressing the jury, the prosecutor concludes : "as you have just heard, there is only 1 chance out of 100000 that the accused, whose DNA matches the one found at the crime scene, is innocent ... that is to say he is guilty!".

What do you think of the prosecutor's argument and of the guilt of the accused?