ELU 501
Data science, graph theory and social network studies

Yannis Haralambous (IMT Atlantique)

April 27, 2018

Part III

Lecture 4
Bayesian networks

Random variables

- A *(discrete) random variable* ⟨🇫🇷 *variable aléatoire (discrète)*⟩ is a variable over a space of *events* ⟨🇫🇷 *évènement*⟩. Its values are the *domain* of the variable.
- Giving the probability of each value is giving the *distribution* ⟨🇫🇷 *distribution*⟩ of the variable. We call it a random variable if

$$\sum_{x \in \mathrm{dom}(X)} P(X = x) = 1$$

and P must be additive on mutually exclusive events:

$$P(X = x \text{ or } X = y) = P(X = x) + P(X = y).$$

Joint distribution

- A *joint distribution* ⟨🇫🇷 *distribution jointe*⟩ $P(X, Y)$ of two variables X and Y is the probability of the Cartesian product of the event: $P(X = x \text{ and } Y = y)$.
- We can return from a joint distribution to a single-variable distribution by *marginalization* ⟨🇫🇷⟩*marginalisation*:

$$P(X = x) = \sum_{y \in \mathrm{dom}(Y)} P(X = x, Y = y).$$

- Attention: notation $P(X, Y)$ must not be confused with the one of a function of two variables $f(x, y)$. In $P(X, Y)$ the order of variables makes no difference, this is why some authors write $P(X \cap Y)$ instead.

Conditional probability

- The *probability of $X = x$ knowing $Y = y$* is defined by
$$P(X = x \mid Y = y) := \frac{P(X = x, Y = y)}{P(Y = y)}.$$
This makes sense only when $Y = y$ is not an impossible event.
- By commutativity of joint distribution we get *Bayes' formula*:
$$\begin{aligned} P(X = x \mid Y = y) &= \frac{P(Y = y \mid X = x)\, P(X = x)}{P(Y = y)} \\ &= \frac{P(Y = y \mid X = x)\, P(X = x)}{\sum_x P(X = x, Y = y)}, \end{aligned}$$
where we call $P(X = x)$ the *priori* ⟨🇫🇷 *probabilité antérieure*⟩, $P(X = x \mid Y = y)$ the *posteriori* ⟨🇫🇷 *probabilité postérieure*⟩, and $P(Y = y \mid X = x)$ the *likelihood* ⟨🇫🇷 *fonction de vraisemblance*⟩ of X.

Conditional probability

- Attention: $x \mapsto P(X = x \mid Y = y)$ is a distribution, but *not so* $x \mapsto P(X = x, Y = y)$, since $\sum_* P(X = *, Y = y) = P(Y = y) \neq 1$.
- To get a distribution out of
$$x \mapsto P(X = x, Y = y)$$
, we just need to divide it by
$$\sum_* P(X = *, Y = y),$$
and we get...
$$x \mapsto P(X = x \mid Y = y).$$
- The two functions are equal up to multiplication with a normalization constant.

Independence

- Two variables are *independent* if knowledge of the state of one of them does not provide any information about the state of the other.
- Mathematically this means that
$$P(X, Y) = P(X)P(Y)$$
or that
$$P(X \mid Y) = P(X) \text{ and } P(Y \mid X) = P(Y).$$
- More generally, if $P(X = x, Y = y) = kf(x)g(y)$ where k is constant, and f and g positive functions, then X and Y are independent. We denote this by $X \perp\!\!\!\perp Y$. When two variables are dependent, we will write $X \not\!\perp\!\!\!\perp Y$.
- Independence and dependance are commutative but not transitive (give an example).

Independence

A PSEUDO-PARADOX.
Take distribution $P(X = a, Y = 1) = 1$ and $P(X, Y) = 0$ otherwise. Obviously $P(X = a) = 1$, $P(Y = 1) = 1$ and otherwise $P(X) = P(Y) = 0$ so that we do have $P(X, Y) = P(X)P(Y)$, and hence $X \perp\!\!\!\perp Y$.

But still they are always in the same joint state...

Explanation: knowing the state of X does not provide any additional information on the state of Y since we know it already.

Independence

Another pseudo-paradox.
Take the following events: $X = \{\text{a dice gives an even number}\}$, $Y = \{\text{the same dice gives 1 or 2}\}$. Here also, $X \perp\!\!\!\perp Y$!!

Indeed, $P(X, Y) = P(\{2\}) = \frac{1}{6} = \frac{1}{2} \cdot \frac{1}{3} = P(X)P(Y)$.

Also: $P(X \mid Y) = P(\text{even} \mid \{1, 2\}) = \frac{1}{2} = P(\{2, 4, 6\}) = P(X)$ (numerically probabilities are equal even though for different reasons).
Idem: $P(Y \mid X) = P(\{1, 2\} \mid \text{even}) = P(\{1\} \mid \text{even}) + P(\{2\} \mid \text{even}) = 0 + \frac{1}{3} = P(\{1, 2\}) = P(Y)$. Confusing, no?

Confusion comes from the fact that X (resp. Y) provides information about dice rolling in general, but not on Y (resp. X) per se.

Conditional independence

- *Conditional independence* ⟨*indépendance conditionnelle*⟩ will be the fundamental concept of this lecture.
- X and Y are *independent knowing* Z, if
$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$
for all values of X, Y and Z. We write $X \perp\!\!\!\perp Y \mid Z$ (although it would be smarter to write $X \perp\!\!\!\perp_Z Y$ but it is too late to change notation). When X and Y are conditionally dependent, we will write $X \top\!\!\!\top Y \mid Z$.
- Attention! We have neither $X \perp\!\!\!\perp Y \mid Z \Rightarrow X \perp\!\!\!\perp Y$, nor $X \perp\!\!\!\perp Y \Rightarrow X \perp\!\!\!\perp Y \mid Z$.

Unobserved variables represent *opinions*. Observed variables represent *facts*. Opinions can be influenced from facts but also from other opinions.

Conditional independence

- In the real world we rarely encounter independent events (and when so, then they are probably of no interest...).
- Therefore, when modeling the world we rather use *conditional independence* ⟨*indépendance conditionnelle*⟩, a much more subtle notion.
- Example: Let there be events A = "Where I live it rains often" and B = "Where I live people eat a lot of crêpes". They are not independent since Darwinian evolution and history of mankind have resulted in having a place where it rains often and people eat a lot of crêpes. Therefore A and B are not independent and knowing that we are in a place where people eat a lot of crêpes allows us to assume that it probably also rains a lot, and vice-versa.
- Consider now the event C = "I live in Brest". *Conditionally to* C, *A and B are independent* because once I know the place is Brest, A does not provide any extra information about B nor B about A.

An exercise to rack one's brains

Let there be two factories, A and B, producing watches. Factory A produces one broken watch in a 100, and factory B one in 200. Let us imagine a reseller gets a shipment of watches not knowing from which factory they come. She tests the first watch: it works.

She puts back the first watch and draws a second. What is the probability that the second watch works as well?

The experiment we did is a draw with replacement ⟨tirage avec remise⟩. Are the random variable describing the states of the two watches independent?

An exercise to rack one's brains

Let X_n be the state of the nth watch, and Y the factory from which originates the shipment. Let us suppose that $P(Y=A) = P(Y=B) = \frac{1}{2}$. We know that $P(X_n = 1 \mid Y = A) = 0.99$ and $P(X_n = 1 \mid Y = B) = 0.995$. We want to calculate $P(X_2 = 1 \mid X_1 = 1)$.

By Bayes,

$$P(X_2 = 1 \mid X_1 = 1) = P(X_1 = 1, X_2 = 1)/P(X_1 = 1).$$

We have

$$\begin{aligned} P(X_1 = 1, X_2 = 1) = {} & P(X_1 = 1, X_2 = 1 \mid Y = A)P(Y = A) \\ & + P(X_1 = 1, X_2 = 1 \mid Y = B)P(Y = B) \end{aligned}$$

and

$$P(X_1 = 1) = P(X_1 = 1 \mid Y = A) + P(X_1 = 1 \mid Y = B).$$

An exercise to rack one's brains

As the draw is with replacement, knowing the factory, X_1 and X_2 are conditionally independent. Therefore $P(X_1 = 1, X_2 = 1) = 0.98505625$, while $P(X_1 = 1)P(X_2 = 1) = 0.98505$, therefore X_1 are X_2 *are not* independent(!).

Hence, $P(X_2 = 1 \mid X_1 = 1) = 0.9925063$ is not equal to $P(X_2 = 1) = 0.9925$.

Therefore if the first watch works, there is a small additional chance that the second will work too.

Properties of conditional independence

Conditional independence has some very nice properties:

- *Symmetry* $X \perp\!\!\!\perp Y \mid Z \Rightarrow Y \perp\!\!\!\perp X \mid Z$.
- *Decomposition* $X \perp\!\!\!\perp \{Y, W\} \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z$.
- *Weak union* $X \perp\!\!\!\perp \{Y, W\} \mid Z \Rightarrow X \perp\!\!\!\perp Y \mid Z, W$.
- *Contraction* $(X \perp\!\!\!\perp W \mid Z, Y) \wedge (X \perp\!\!\!\perp Y \mid Z) \Rightarrow X \perp\!\!\!\perp \{Y, W\} \mid Z$.
- *Intersection* If the distributions are positive $(\alpha \neq \emptyset \Rightarrow P(\alpha) > 0)$,
  $(X \perp\!\!\!\perp Y \mid Z, W) \wedge (X \perp\!\!\!\perp W \mid Z, Y) \Rightarrow X \perp\!\!\!\perp \{Y, W\} \mid Z$.

In these properties you can replace X, Y, Z, etc. by sets of random variables. (In the case of intersection, the sets of variables must be mutually disjoint.)

Queries

Let $\mathcal{X}$ be the total set of random variables of our model, E the subset of evidence variables and Y the subset of query variables. Let $W := \mathcal{X} \setminus E$.

- A *probability query* is the computation of $P(Y \mid E = e)$, called the *posterior probability distribution conditioned by the fact that* $E = e$.
- A *MAP* (= maximum a posteriori probability estimate) ⟨*estimateur du maximum a posteriori*⟩ is the most likely joint assignment of all (non-evidence) variables, given the evidence $E = e$: $\mathrm{MAP}(W \mid E = e) := \mathrm{argmax}_w P(w, E = e)$ (which is also $\mathrm{argmax}_w P(w \mid E = e)$).
- If $Z = \mathcal{X} \setminus (Y \cup E)$, a *marginal MAP* is the most likely assignment to the variables in Y given $E = e$: $\mathrm{MAP}(Y \mid e) := \mathrm{argmax}_y P(y \mid e) = \mathrm{argmax}_y \sum_z P(Y, Z \mid E = e)$.

Conditional probability tables

The probabilities that a British subject is of English, Scottish or Welsh origin (O) are resp. 0.88, 0.08 and 0.04. The conditional probabilities (knowing her origin) that the mother tongue (L) of the given subject if English, Scottish or Welsh are:

$$P(L = \text{line} \mid O = \text{column}) = \begin{pmatrix} 0.95 & 0.7 & 0.6 \\ 0.04 & 0.3 & 0.0 \\ 0.01 & 0.0 & 0.4 \end{pmatrix}.$$

To obtain the joint distribution we just write $P(L, O) = P(L \mid O)P(O)$ (where $P(O)$ is the diagonal matrix of the distribution of O):

$$P(L = \text{line}, O = \text{column}) = \begin{pmatrix} 0.836 & 0.056 & 0.024 \\ 0.0352 & 0.024 & 0 \\ 0.0088 & 0 & 0.016 \end{pmatrix}.$$

Probabilistic inference

We will identify all relevant random variables $X_1, \ldots, X_N$ in a given context and will try to establish a probabilistic model $P(X_1, \ldots, X_N)$ of their interactions. Inference will consists in inserting information on the state of some of the variables and see how this affects the model, conditionally to this information.

Example: *Scientists found that 90% of Kreuzfeld-Jacob patients were eating hamburgers. In the US population, there is one patient in 100 000 individuals. If we consider that every second person eats hamburgers, what is the probability of catching K-J when you eat hamburgers?*
And if the ratio of hamburger eaters was lesser (like in India) what would happen to this probability?

Kreuzfeld-Jacob and hamburgers

Let KJ be the fact of having the illness, H of eating hamburgers. The statement says that $P(H \mid KJ) = 0.9$, $P(KJ) = 10^{-5}$ et $P(H) = 0.5$. We seek for (happiness and) $P(KJ \mid H)$:

$$P(KJ \mid H) = \frac{P(KJ, H)}{P(H)} = \frac{P(H \mid KJ)P(KJ)}{P(H)} = 1.8 \cdot 10^{-5}.$$

It is 80% more frequent than when there is no constraint but the probability is still quite weak.

And what of those not eating hamburgers?

$$P(KJ \mid \neg H) = \frac{P(KJ, H)}{P(\neg H)} = \frac{P(\neg H \mid KJ)P(KJ)}{P(\neg H)} = 0.2 \cdot 10^{-5}.$$

And if it was only one person in 1000 eating hamburgers? We find $P(KJ \mid H) = 1/100$.

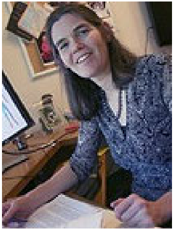
Modus ponens

- *Modus ponens* ⟨🇫🇷 *modus ponens*⟩ is a logical inference rule saying that "every apple is a fruit" and "every fruit grows on a tree" entails that "every apple grows on a tree".
- Let us use a probabilistic inference to show it.
- "Every apple is a fruit": $P(F \mid A) = 1$. "Every fruit grows on a tree": $P(T \mid F) = 1$. Let us show that $P(T \mid A) = 1$:
- Showing that $P(T \mid A) = 1$ is equivalent to showing that $P(\neg T \mid A) = 0$, which again is equivalent to $P(\neg T, A) = 0$. The latter can be written:
$$P(\neg T, F, A) + P(\neg T, \neg F, A),$$
let us show that both terms are zero:
(a) $P(\neg T, F, A) \leq P(\neg T, F) = P(\neg T \mid F)P(F) = 0$,
(b) $P(\neg T, \neg F, A) \leq P(\neg F, A) = P(\neg F \mid A)P(A) = 0$. QED

Daphne Koller's quote

As Daphne Koller [**?**, p. 45] puts it:

> *The explicit representation of the joint distribution is unmanageable from every perspective. Computationally, it is very expensive to manipulate and generally too large to store in memory. Cognitively, it is impossible to acquire so many numbers from a human expert; moreover, the numbers are very small and do not correspond to events that people can reasonably contemplate.*



Naive Bayes model

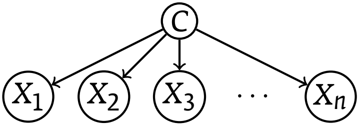
- Imagine you have a situation with $n+1$ random variables $C, X_1, X_2, \ldots, X_n$, where the X_* are observed and we wish to infer on C.
- The *Naive Bayes assumption* is to assume that, for all i,

$$X_i \perp\!\!\!\perp \{X_1, \ldots, \hat{X}_i, \ldots, X_n\} \mid C.$$

- Then the joint distribution factorizes as

$$P(C, X_1, \ldots, X_n) = P(C) \prod_{i=1}^{n} P(X_i \mid C),$$

which is a small set of factors. (Think of X_* as symptoms and C as disease, can we have $X_i \perp\!\!\!\perp X_j \mid C$?)



Belief networks, an example

When we have many variables, calculations for obtaining joint distributions or marginalizations rapidly become heavy. A solution is to find (in)dependencies between variables to simplify the model. We then build a *belief network* ⟨🇫🇷 *réseau de croyances*⟩.

The most paradigmatic belief network example is the one of the *sprinkler* ⟨🇫🇷 *arroseur*⟩: *One morning, by leaving her house, Tracey realizes that the grass is wet* (T)*. She asks herself whether it has rained* (R) *or whether she left the sprinkler on* (S)*. Then she notices that her neighbor's Jack's grass is also wet* (J)*, so now she believes that* R *is the most probable event.*

Normally we need $2^4 - 1$ values to describe entirely the distribution $P(S, R, T, J)$.
We always have the chain rule:

$$P(S, R, T, J) = P(T \mid J, R, S)P(J \mid R, S)P(R \mid S)P(S).$$

The sprinkler under the rain

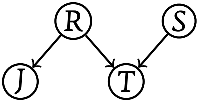
Now we can do some simplifications:

1. if we know whether it has rained or not *and* whether the sprinkler was left on or not, the information on Jack's grass tells us nothing new about Tracey's grass: $P(T \mid J, R, S) = P(T \mid R, S)$;
2. if we know whether it has rained or not, the fact whether Tracey has left her sprinkler on tells us nothing bout Jack's grass: $P(J \mid R, S) = P(J \mid R)$;
3. and finally, sprinklers do not cause or prevent rain: $P(R \mid S) = P(R)$.

The formula becomes (8 parameters!):

$$P(S, R, T, J) = P(T \mid R, S)P(J \mid R)P(R)P(S).$$

Here is the graphical model of the factorization:



The sprinkler under the rain

Let us give values to the 8 parameters: $P(R=1)=0.2$, $P(S=1)=0.1$, $P(J=1\mid R=1)=1$, $P(J=1\mid R=0)=0.2$, $P(T=1\mid R=1,S=0)=1$, $P(T=1\mid R=1,S=1)=1$, $P(T=1\mid R=0,S=1)=0.9$, $P(T=1\mid R=0,S=0)=0$. And now, *calculemus*:

Our priori: $P(S=1)=0.1$.

First posteriori: $P(S=1\mid T=1)=\frac{\sum_{J,R}P(T=1,J,R,S=1)}{\sum_{J,R,S}P(T=1,J,R,S)}=0.3382$.

Second posteriori:

$P(S=1\mid T=1,J=1)=\frac{\sum_{R}P(T=1,J=1,R,S=1)}{\sum_{R,S}P(T=1,J=1,R,S)}=0.1604$.

Note that the second is significantly smaller than the first.

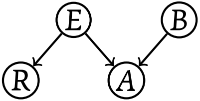
By concluding that rain has wet her grass and not the sprinkler, Tracey's brain has somehow done these calculations and has compared the results.

The shaken burglar

Returning home, Tracey discovers that the alarm is on ($A = 1$). Was it caused by burglary ($B = 1$) or by an earthquake ($E = 1$)? She turns the radio on and finds out ($R = 1$) that there has indeed been an earthquake. What shall she conclude?

The distribution is

$$\begin{aligned} P(B, E, A, R) &= P(A \mid B, E, R)P(R \mid B, E)P(E \mid B)P(B) \\ &= P(A \mid B, E)P(R \mid E)P(E)P(B) \quad \text{(explain why).} \end{aligned}$$



$P(A = 1 \mid B, E)$	B	E
0.9999	1	1
0.99	1	0
0.99	0	1
0.0001	0	0

$P(R = 1 \mid E)$	E
1	1
0	0

$$P(B = 1) = 0.01$$
$$P(E = 1) = 0.000001.$$

Bayesian networks

A *Bayesian network stucture* ⟨*structure de réseau bayésien*⟩ is a dag $\mathcal{G}$ whose vertices represent random variables $X_1, \dots, X_n$. Let $\mathrm{pa}(X_i)$ denote the parents of X_i and $\mathrm{notdesc}(X_i)$ the vertices which are not descendants of X_i (this set does not include X_i itself). Then $\mathcal{G}$ encodes the following set of conditional independence assumptions, called *local independencies* and denoted by $\mathcal{I}_\ell(\mathcal{G})$:

$$\text{for every } X_i,\ X_i \perp\!\!\!\perp \mathrm{notdesc}(X_i) \mid \mathrm{pa}(X_i).$$

We will show that this is equivalent to the fact that the joint probability distribution is factorized in a very specific way.

I-maps

- Let P be a distribution over $\mathcal{X}$. We define $\mathcal{I}(P)$ to be the set of independence assertions of the kind $X \perp\!\!\!\perp Y \mid Z$ (where $X, Y, Z \subset \mathcal{X}$) that hold in P.

> Consider influence in the network as a *fluid* propagating through edges. Independencies will block that fluid, since independence means that having an opinion on one variable does not change our opinion on the other.

- Let $\mathcal{K}$ be a graph object encoding a set of independencies $\mathcal{I}(\mathcal{K})$. We say that $\mathcal{K}$ is an *I-map* ⟨*I-map*⟩ for a set of independencies I if $\mathcal{I}(\mathcal{K}) \subseteq I$.
- Finally, we say that a dag $\mathcal{K}$ is an *I-map of a probability distribution* P, if $\mathcal{I}(\mathcal{K}) \subseteq \mathcal{I}(P)$.

Bayesian networks

Let $\mathcal{G}$ be a BN structure over variables $X_1, \ldots, X_n$. We say that a distribution P *factorizes according to $\mathcal{G}$* if P can be expressed as a product:

$$P(X_1, \ldots, X_n) := \prod_{i=1}^{n} P(X_i \mid \mathrm{pa}(X_i)).$$

A *Bayesian network* ⟨🇫🇷 *réseau bayésien*⟩ is a pair $\mathcal{B} = (\mathcal{G}, P)$ where P factorizes over $\mathcal{G}$.

Bayesian networks

Theorem

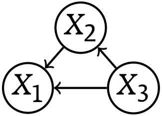
Let $\mathcal{G}$ be a BN structure over $\mathcal{X}$ and let P be a joint distribution over $\mathcal{X}$. If $\mathcal{G}$ is an I-map for P then P factorizes according to $\mathcal{G}$.

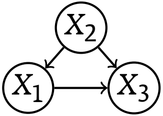
The proof uses the fact that there is a topological ordering $X_1, \ldots, X_n$ of $\mathcal{G}$ and writes $P(X_1, \ldots, X_n) = \prod P(X_i \mid X_1, \ldots, X_{i-1})$. By the I-map assumption we get that the latter is equal to $P(X_i \mid \mathrm{pa}(X_i))$.

Theorem

Let $\mathcal{G}$ be a BN structure over $\mathcal{X}$ and let P be a joint distribution over $\mathcal{X}$. If P factorizes according to $\mathcal{G}$ then $\mathcal{G}$ is an I-map for P.

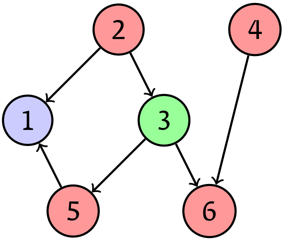
Like $P(X_1, X_2, X_3)$ is equal to $P(X_1 \mid X_2, X_3)P(X_2 \mid X_3)P(X_3)$ as well as $P(X_3 \mid X_1, X_2)P(X_1 \mid X_2)P(X_2)$, we realize that the same distribution can be described by more than one dags. We have to find the most interesting one.





Markov blanket

The *Markov blanket* ⟨*couverture de Markov*⟩ of a vertex is the set of its parents, children and parents of children.



E.g., here, the Markov blanket of 3 is $\{2, 4, 5, 6\}$.

```
from libpgm.graphskeleton import GraphSkeleton
from libpgm.nodedata import NodeData
from libpgm.discretebayesiannetwork import DiscreteBayesianNetwor
from libpgm.tablecpdfactorization import TableCPDFactorization
def getTableCPD():
   nd = NodeData()
   skel = GraphSkeleton()
   jsonpath="job_interview.txt"
   nd.load(jsonpath)
   skel.load(jsonpath)
   bn = DiscreteBayesianNetwork(skel, nd)
   tablecpd=TableCPDFactorization(bn)
   return tablecpd
tcpd=getTableCPD()
print tcpd.specificquery(dict(Offer='1'),dict())
```

Draw the Bayesian network. What does the code? Why `dict()`? Calculate $P(\text{Offer} = 1 \mid \text{Grades} = 0, \text{Experience} = 0)$ (causal reasoning), $P(\text{Experience} = 1)$ (evidential reasoning) and $P(\text{Experience} = 1 \mid \text{Interview} = 2)$ (intercausal reasoning).

Show that the probability of Offer knowing Interview doesn't depend on the observed values of Experience and Grades. Why? What happens if Interview is not observed?

Markov blanket

We may ask the question: if a graph encodes some local independencies, are there other ones? Do we have an independency whenever there is no edge between two vertices?

Theorem

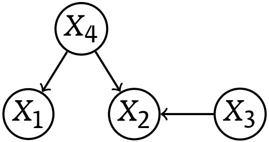
Let X be a BN vertex and $\mathrm{MB}(X)$ is Markov blanket. then for every $Y \notin \mathrm{MB}(X)$, we have

$$X \perp\!\!\!\perp Y \mid \mathrm{MB}(X).$$

I.e., the Markov blanket contains all the information available in the BN concerning a given vertex.

Examples of conditional independencies

Yannis Haralambous (IMT Atlantique)



In this BN we can ask the question: are X_1 and X_2 cond. indep. knowing X_4?
It happens that yes, because

$$P(X_1, X_2 \mid X_4) = \frac{1}{P(X_4)} \sum_{X_3} P(X_1, X_2, X_3, X_4)$$
$$= P(X_1 \mid X_4) \sum_{X_3} P(X_2 \mid X_3, X_4) P(X_3),$$
$$P(X_2 \mid X_4) = \frac{1}{P(X_4)} \sum_{X_1, X_3} P(X_1, X_2, X_3, X_4)$$
$$= \sum_{X_3} P(X_2 \mid X_3, X_4) P(X_3)$$

and therefore $P(X_1, X_2 \mid X_4) = P(X_1 \mid X_4) P(X_2 \mid X_4)$.

Examples of conditional independencies

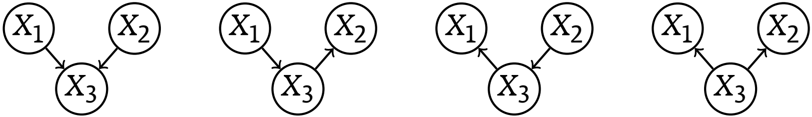
It would be nice to have a general algorithm for finding out if two vertex sets are independent knowing a third set (the set of observed variables).

Let us take the case of three random variables X_1, X_2, X_3.

We always have

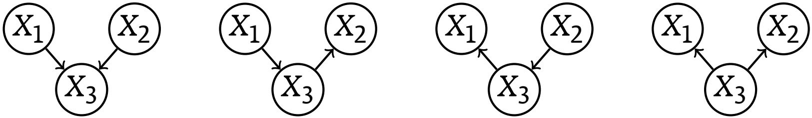
$$P(X_1, X_2, X_3) = P(X_{i_1} \mid X_{i_2}, X_{i_3})P(X_{i_2} \mid X_{i_3})P(X_{i_3}).$$

Encode an independence would be removing an edge of the triangle $X_{i_1}, X_{i_2}, X_{i_3}$. There are four cases:



common effect causal trail evidential trail common cause

Examples of conditional independencies



Calculation gives

$$
\begin{aligned}
P(X_2 \mid X_3)P(X_3 \mid X_1)P(X_1) = \frac{P(X_2, X_3)P(X_3, X_1)}{P(X_3)} &= P(X_1 \mid X_3)P(X_2, X_3) \\
&= P(X_1 \mid X_3)P(X_3 \mid X_2)P(X_2) \\
&= P(X_1 \mid X_3)P(X_2 \mid X_3)P(X_3).
\end{aligned}
$$

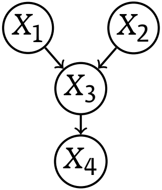
i.e., the three last cases represent the same distribution, with $X_1 \perp\!\!\!\perp X_2 \mid X_3$.

In the first case we do not have $X_1 \perp\!\!\!\perp X_2 \mid X_3$ but

$$
\begin{aligned}
P(X_1, X_2) = \sum_{X_3} P(X_1, X_2, X_3) &= \sum_{X_3} P(X_3 \mid X_1, X_2)P(X_1)P(X_2) \\
&= P(X_1)P(X_2) \quad \text{and therefore } X_1 \perp\!\!\!\perp X_2.
\end{aligned}
$$

Examples of conditional independencies

And what about the graph:



$$\begin{aligned} P(X_1, X_2 \mid X_4) &= \frac{P(X_1, X_2, X_4)}{P(X_4)} \\ &= \frac{1}{P(X_4)} \sum_{X_3} P(X_4 \mid X_3) P(X_3 \mid X_1, X_2) P(X_1) P(X_2) \\ &\neq P(X_1 \mid X_4) P(X_2 \mid X_4) \quad \text{in general.} \end{aligned}$$

On the opposite

$$\begin{aligned} P(X_1, X_2) &= \sum_{X_3, X_4} P(X_1, X_2, X_3, x_4) = \sum_{X_3, X_4} P(X_4 \mid X_3) P(X_3 \mid X_1, X_2) P(X_1) P(X_2) \\ &= P(X_1) P(X_2) \quad \text{and therefore we have } X_1 \perp\!\!\!\perp X_2 \end{aligned}$$

like in the case $X_1 \rightarrow X_3 \leftarrow X_2$.

Conditional independencies

- We call $X \to Y \leftarrow Z$ a *v-structure* ⟨*v-structure*⟩.
- If we have $X \leftarrow Z \to Y$, $X \to Z \to Y$ or $X \leftarrow Z \leftarrow Y$ and we condition with respect to Z then there can be no conditional dependency between X and Y.
- If we have a v-structure $X \to Z \leftarrow Y$ and neither Z nor any of its descendants are in the conditioning set, then we can have no dependency between X and Y.

I.e.,

- Conditioning with respect to Z a network $X \leftarrow Z \to Y$, $X \to Z \to Y$ or $X \leftarrow Z \leftarrow Y$, makes X and Y (cond.) independent. Marginalizing with respect to Z makes them dependent.
- Conditioning with respect to Z (or a descendant of Z) a v-structure $X \to Z \leftarrow Y$, makes X and Y (cond.) dependent. Marginalizing with respect to Z makes them independent.

D-separation

Let $\mathcal{G}$ be a BN structure, and $X_1 \rightleftharpoons \cdots \rightleftharpoons X_n$ a trail in $\mathcal{G}$. We say that the trail is *active given the set of observed variables* Z if

- whenever we have a v-structure $X_{i-1} \rightarrow X_i \leftarrow X_{i+1}$, then X_i or one of its descendants are in Z;
- no other vertex along the trail is in Z.

Let X, Y, Z be three sets of nodes in $\mathcal{G}$. We say that X and Y are *d-separated* ⟨*d-séparés*⟩ given Z, and we denote $\text{d-sep}(X; Y \mid Z)$ if there is no active trail between any vertex of X and any vertex of Y, given Z.

D-separation

We call

$$\mathcal{I}(\mathcal{G}) = \{(X \perp\!\!\!\perp Y \mid Z) \mid \text{d-sep}(X; Y \mid Z)\}$$

the set of *global Markov independencies*.

We will see that $\mathcal{I}(\mathcal{G})$ are precisely those that are guaranteed to hold for every distribution over $\mathcal{G}$.

Concerning d-separation *as a method for determining conditional independence* we need to show two aspects:

1. *soundness* ⟨*correction*⟩: if X and Y are d-separated given Z are they guaranteed to be cond. indep. given Z in *any* distribution?
2. and *completeness* ⟨*complétude*⟩: does d-separation detect *all* possible independencies in all possible distributions?

D-separation

Theorem

If a distribution P factorizes according to $\mathcal{G}$ then $\mathcal{I}(\mathcal{G}) \subseteq \mathcal{I}(P)$.

This establishes soundness.

Completeness in the sense that "if X and Y are not d-separated given Z then X and Y are dependent in *all* distributions P" ($\Leftrightarrow$ "if there is *some* P such that $X \perp\!\!\!\perp Y$ then X and Y are d-separated") doesn't work:

Take the graph $a \to c \leftarrow b$ and the distribution $P(c = 1 \mid a, b) = 0.5$, $P(a = 1) = 0.3$ et $P(b = 1) = 0.4$. Let $X = a$, $Y = b$, $Z = \emptyset$. d-separation gives us that $a \perp\!\!\!\perp b$. But numerically we have also independencies $a \perp\!\!\!\perp c$ and $b \perp\!\!\!\perp c$.

D-separation

Theorem

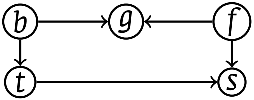
If X and Y are not d-separated given Z then X and Y are dependent given Z in *some* distribution P that factorizes over $\mathcal{G}$.

SKETCH OF THE PROOF: Since X and Y are not d-separated there is an active trail $U_1, \ldots, U_k$. We define CPDs for the variables on the trail so that each pair is correlated; in case of a v-structure $U_i \rightarrow U_{i+1} \leftarrow U_{i+2}$ we define the CPD of U_{i+1} as well as down-stream evidence so that correlation between U_i and U_{i+2} is activated. All other CPDs of the graph are chosen to be uniform, so influence flows only along the trail.

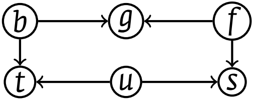
Theorem

For all distributions P that factorize over $\mathcal{G}$ except for a set of measure zero in the space of CPD parametrization, we have $\mathcal{I}(P) = \mathcal{I}(\mathcal{G})$.

D-separation examples



Do we have $t \perp\!\!\!\perp f$ (i.e., $t \perp\!\!\!\perp f \mid \emptyset$)?
Is $t \perp\!\!\!\perp f \mid g$?



Is $\{b, f\} \perp\!\!\!\perp u$ (i.e., $\{b, f\} \perp\!\!\!\perp u \mid \emptyset$)?
Is $\{b, f\} \perp\!\!\!\perp u \mid g$? $\{b, f\} \perp\!\!\!\perp u \mid \{t, s\}$?

I-equivalence

- There is *I-equivalence* ⟨ *équivalence markovienne*⟩ between graph structures $\mathcal{K}_1$ and $\mathcal{K}_2$, if $\mathcal{I}(\mathcal{K}_1) = \mathcal{I}(\mathcal{K}_2)$.
- Examples: $a \to c \to b$, $a \leftarrow c \leftarrow b$ and $a \leftarrow c \to b$ are I-equivalent. This is not the case for $a \to c \leftarrow b$.
- Vertices A, B and C form an *immorality* ⟨ *immoralité*⟩ if C is a child of A and B and there are no edges $A \to B$ or $A \leftarrow B$.

Theorem

Let $\mathcal{G}_1'$, $\mathcal{G}_2'$ the undirected graphs underlying $\mathcal{G}_1$, $\mathcal{G}_2$. Then $\mathcal{G}_1$ and $\mathcal{G}_2$ are I-equivalent iff

1. $G_1' = G_2'$ and
2. G_1 and G_2 have the same immoralities.

From distributions to graphs

- QUESTION: *If we start with a distribution* P*, to what extent can we construct* $\mathcal{G}$ *such that* $\mathcal{I}(\mathcal{G})$ *are a good surrogate for* $\mathcal{I}(\mathrm{P})$*?*
- The question may seem pointless because we never have an entire joint distribution, but it will help for constructing Bayesian networks.
- Notice that the complete graph is an I-map for any distribution, yet it does not reveal any of the independency structures of the distribution.
- A graph $\mathcal{K}$ is a *minimal I-map* for a set of independencies $\mathcal{I}$ is an I-map such that the removal of a single edge invalidates the I-map property.

Algorithm for building minimal I-maps

Data: $X_1, \ldots, X_n$ an ordering of random variables in $\mathcal{X}$,
$\mathcal{I}$ set of independencies
Result: A minimal I-map

```
Set 𝒢 to an empty graph over 𝒳;
for i = 1, ..., n do
    U ← {X_1, ..., X_{i-1}};    // U candidate for parents of X_i
    for U′ ⊆ {X_1, ..., X_{i-1}} do
        if U′ ⊂ U ∧ (X_i ⊥⊥ {X_1, ..., X_{i-1}} \ U′ | U′) ∈ 𝓘 then
            U ← U′
        end
    end
    // U minimal set satisfying (X_i ⊥⊥ {X_1, ..., X_{i-1}} \ U | U)
    // Now set U to be the parents of X_i
    for X_j ∈ U do
        Add X_j → X_i to 𝒢
    end
end
return 𝒢
```

P-maps

- A minimal I-map is still not able to provide all independencies in P out of $\mathcal{G}$.
- We say that a graph $\mathcal{K}$ is a *perfect map (P-map)* for a set $\mathcal{I}$ if we have $\mathcal{I}(\mathcal{K}) = \mathcal{I}$. It is a perfect map for P if $\mathcal{I}(\mathcal{K}) = \mathcal{I}(P)$.
- P-maps do not always exist, and when they exist, they are not unique. But all P-maps are I-equivalent.
- To find a P-map $\mathcal{G}^*$ for a distribution P we will
  1. identify the undirected skeleton;
  2. identify immoralities;
  3. represent equivalence classes of graphs.

Algorithm for finding P-maps

- The algorithms returns a *compact representation of a DAG I-equivalence class*.
- FIRST STEP: IDENTIFY THE UNDIRECTED SKELETON.
- We will create an undirected graph S that contains an edge $X - Y$ if X and Y are adjacent in $\mathcal{G}^*$.
- Lemma: in P we have $X \not\perp\!\!\!\perp Y \mid U$ for any set U that does not include X and Y.
- Lemma: Let X and Y be non-adjacent, then either $X \perp\!\!\!\perp Y \mid \mathrm{pa}(X)$ or $X \perp\!\!\!\perp Y \mid \mathrm{pa}(Y)$.
- I.e., if X and Y are not adjacent then we can always find a set U so that $X \perp\!\!\!\perp Y \mid U$. We call this set U a *witness* ⟨*témoin*⟩ of their independence.
- Furthermore $\#U \leq \mathrm{indegree}(\mathcal{G}^*)$.

Algorithm for finding P-maps

Data: $X_1, \ldots, X_n$ set of random variables, P, d bound on witness set
Result: An undirected skeleton of $\mathcal{G}^*$
Let $\mathcal{H} := K_{\#\mathcal{X}}$ (complete graph);
for X_i, X_j *in* $\mathcal{X}$ **do**
    $U_{X_i,X_j} \leftarrow \emptyset$;
    for $U \in \text{Witnesses}(X_i, X_j, \mathcal{H}, d)$ **do**
        if $(X_i \perp\!\!\!\perp X_j \mid U)$ **then**
            $U_{X_i,X_j} \leftarrow U$;
            Remove $X_i - X_j$ from $\mathcal{H}$;
            break;
        end
    end
end
return $(\mathcal{H}, \{U_{X_i,X_j}$ for all $1 \leq i, j \leq n\})$

Algorithm for finding P-maps

- SECOND STEP: IDENTIFY IMMORALITIES.
- We will use *potential* immoralities and determine whether they are real immoralities. A *potential immorality* is a triplet X, Y, Z such that the skeleton contains $X - Y - Z$ but not $X - Y$.
- If it is an immorality then $X \not\perp\!\!\!\perp Y \mid U$ for any U that contains Z.
- If it is not an immorality, then X and Y are d-separated only if Z is observed, i.e., if $X \perp\!\!\!\perp Y \mid U$, then $Z \in U$.
- All we need to do is to take all potential immoralities X, Y, Z and the witness sets $U_{*,*}$ obtained from the previous algorithm and check whether $Y \in U_{X,Z}$.

Algorithm for finding P-maps

Data: $X_1, \ldots, X_n$ set of random variables, S skeleton, $\{U_{X_i,X_j}\}$ from previous algorithm

Result: An undirected skeleton $\mathcal{K}$ of $\mathcal{G}^*$ with immoralities

$\mathcal{K} \leftarrow S$;

for *X_i, X_j, X_k in S such that $X_i - X_j - X_k$ and $X_i \not- X_k$ in S* **do**

    `// ` $X_i - X_j - X_k$ ` is a potential immorality`

    if $X_j \notin U_{X_i,X_k}$ **then**

        Add the orientations $X_i \to X_j$ and $X_j \leftarrow X_k$ to $\mathcal{K}$;

    end

end

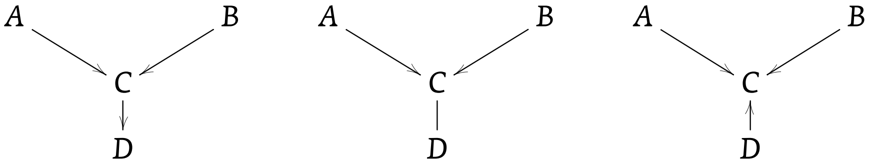
return $\mathcal{K}$

Algorithm for finding P-maps

- THIRD STEP: FIND THE EQUIVALENCE CLASS OF $\mathcal{G}^*$.
- The previous step provided an undirected skeleton and immoralities, but no orientation for the remaining edges.
- We define a *trail* as a path of directed and/or undirected edges where orientation does not need to be respected.
- We define a *partially directed acyclic graph (PDAG)* or *chain graph* as a mixed directed/undirected graph without trails which are cycles.
- For every PDAG there is a partition $K_1, \ldots, K_\ell$ such that each K_i contains only undirected edges and for any $X \in K_i$ and $Y \in K_j$ there can be only directed edges between X and Y. A K_i is called a *chain component*.

Algorithm for finding P-maps

- Let $\mathcal{G}$ be a dag. A chain graph $\mathcal{K}$ is a *class PDAG of the I-equivalence class of $\mathcal{G}$* if it shares the same skeleton and contains a directed edge $X \to Y$ iff all $\mathcal{G}'$ I-equivalent to $\mathcal{G}$ also contain the edge $X \to Y$.
- Question: can we find edges which do not belong to immoralities but which will be directed in the class PDAG?



(a) (b) (c)

- (a) is $\mathcal{G}^*$, (b) is its skeleton with immoralities, (c) is a dag which is not I-equivalent to $\mathcal{G}^*$.

Algorithm for finding P-maps

ELU 501 Data science, graph theory and social network studies

Yannis Haralambous (IMT Atlantique)

Conditional probability
Independence
Conditional independence
Probabilistic inference
Belief networks, an example
Bayesian networks
I-maps
Bayesian networks
Bayesian networks

- Based on the previous example there are three rules that can be applied to produce more directed edges:
- R1: $\begin{tikzcd} X \arrow[d] & \\ Y \arrow[r, dash] & Z \end{tikzcd}$ gives $\begin{tikzcd} X \arrow[d] & \\ Y \arrow[r] & Z \end{tikzcd}$
- R2: $\begin{tikzcd} X \arrow[d] \arrow[dr, dash] & \\ Y \arrow[r] & Z \end{tikzcd}$ gives $\begin{tikzcd} X \arrow[d] \arrow[dr] & \\ Y \arrow[r] & Z \end{tikzcd}$
- R3: $\begin{tikzcd} & X \arrow[dl, dash] \arrow[dr, dash] \arrow[dd, dash] & \\ Y_1 \arrow[dr] & & Y_2 \arrow[dl] \\ & Z & \end{tikzcd}$ gives $\begin{tikzcd} & X \arrow[dl, dash] \arrow[dr, dash] \arrow[dd] & \\ Y_1 \arrow[dr] & & Y_2 \arrow[dl] \\ & Z & \end{tikzcd}$
- by applying these rules until convergence we obtain the class PDAG.

Algorithm for finding P-maps

Data: $X_1, \ldots, X_n$ set of random variables, P distibution, S, $\mathcal{K}$ and U_{X_i,X_j} from previous algorithms
Result: An class PDAG of $\mathcal{G}^*$
while *not converged* **do**
    Find a subgraph in $\mathcal{K}$ matching the left-hand side of a rule R1-R3;
    Replace the subgraph with the right-hand side of the rule;
end
return the new $\mathcal{K}$

Algorithm for finding P-maps

Theorem

Let P a distribution that has a P-map $\mathcal{G}^*$ and let $\mathcal{K}$ be the PDAG returned by the algorithm we described. Then $\mathcal{K}$ is a class PDAG of $\mathcal{G}^*$.

- We check the following:
  1. acyclicity: the graph returned by the algorithm is acyclic;
  2. soundness: if $X \rightarrow Y \in \mathcal{K}$ then $X \rightarrow Y$ appears in all DAGs in $\mathcal{G}^*$'s I-equivalence class;
  3. completeness: if $X - Y \in \mathcal{K}$ then we can find a dag $\mathcal{G}$ that is I-equivalent to $\mathcal{G}^*$ such that $X \rightarrow Y \in \mathcal{G}$ (and another one such that $X \leftarrow Y \in \mathcal{G}$).
- Attention! if P has no P-map in the first place, then the algorithm will still return a result, but the result will be wrong. So *checking the result* should be the last step of the algorithm.

Bibliography

Barber, D. (2012).
Bayesian Reasoning and Machine Learning.
Cambridge University Press.

House, T. (2011).
Modelling epidemics on networks.
preprint http://arxiv.org/abs/1111.4875.

Koller, D. and Friedman, N. (2009).
Probabilistic Graphical Models, Principles and Techniques.
The MIT Press.

Newman, M. (2010).
Networks, an introduction.
Oxford University Press.

Salganik, M. J., Dodds, P., and Watts, D. (2006).
Experimental study of inequality and unpredictability in an artificial cultural market.
Science, 311:854–856.