50.021 -AI

Alex

Week 12: Probabilistic Graphical Models

Probabilistic Graphical models –a way to efficiently capture probability distributions between large numbers of variables

The goals of this part in AI are:

- a recap on necessary probability knowledge
- see how joint probabilities can be modeled using Bayesian networks
- see a simple way of how to fit its parameters
- see how Bayesian networks imply independence
- see how to sample efficiently from Bayesian networks

In this class you should see – only for discrete variables

- a recap on joint probabilities
- a recap on marginal probabilities
- a recap on various types of independence
- a recap on conditional probabilities
- a recap on the interplay of independence and conditional probabilities
- Bayesian networks to model joint probabilities with less parameters by using conditional independence assumptions
- we need conditional independence for Bayesian networks

Humans want realiability and safety but the world is probabilistic in nature.

Recap: Joint Distribution: N variables

NOTE: In this AI class we will not use densities. For exam you can skip the density stuff, however in practice Bayesian models are often used with densities.

Also: check the uploaded pdf please - you can start at page 9: joint cumulative distribution on n variables. Its about modeling probability distributions of n variables.

Lets consider n variables: $X_1, X_2, \ldots, X_n$. Then the joint distribution is defined in **the Discrete Case:**

We assume that the Variables X_i have a finite (e.g. $X_i \in \{0, 1, 2\}$) or countably infinite domain of values

(e.g. $X_i \in$ integer numbers $\{\ldots, -5, -4, -3, -2, -1, 0, 1, 2, 3, \ldots\}$, or X_i has values from the set of all rational numbers $\frac{p}{q}$ such that both $q \neq 0, p \geq 0$ are integers)

$(x_1, x_2, \ldots, x_n)$ is an n-tuple from the respective domains $x_i \in dom(X_i)$

Let p be a function (**probability mass function, pmf**) such that

$$P(x_1, x_2, \ldots, x_n) \geq 0$$
$$\sum_{(x_1, x_2, \ldots, x_n)} p(x_1, x_2, \ldots, x_n) = 1$$

Let A be any set of values $(x_1, \ldots, x_n)$ for the variables $X_1, X_2, \ldots, X_n$, then the probability of this set A is defined as:

$$P((X_1, X_2, \ldots, X_n) \in A) = \sum_{(x_1, x_2, \ldots, x_n) \in A} p(x_1, x_2, \ldots, x_n)$$

If $X_i \in \{0, 1, 2\}$, then the values $(x_1, x_2, \ldots, x_n)$ are simply all the sequences of length n having all zeros,ones,twos like

$$\underbrace{01022102012010212010201 11}_{n}$$

Case with a density function f: - out of class

Densities are used for continuos spaces like intervals such as $[0, 1]$. Though not evrey distribution on such a space can be described by a density, or a density alone. Point masses cannot.

Suppose we have a density, and variables $X_1, X_2, \ldots, X_n$ that can take values in an space that has an integral. You can imagine that X_i takes values in an interval $[a_i, b_i]$ such like $[0, 2]$. Then the set of X_i's $X_1, X_2, \ldots, X_n$ takes values in the product space

$$\begin{aligned}\prod_i [a_i, b_i] &= [a_1, b_1] \times [a_2, b_2] \times \ldots \times [a_n, b_n] \\ &= \{(x_1, x_2, \ldots, x_n) \mid \forall i \; x_i \in [a_i, b_i]\}\end{aligned}$$

The last is just a way to show what this space contains.

$$\begin{aligned} f(x_1, x_2, \ldots, x_n) &\geq 0 \\ \int_{(x_1, x_2, \ldots, x_n)} f(x_1, x_2, \ldots, x_n) &= 1\end{aligned}$$

Let A be a measurable[1] subset of $\prod_i [a_i, b_i]$

$$P((X_1, X_2, \ldots, X_n) \in A) = \int_A f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n$$

[1] not all sets are measurable, see `https://en.wikipedia.org/wiki/Vitali_set`. You cannot compute an integral over a measurable subset

In case that the set A is a product of intervals,

$$A = [c_1, d_1] \times [c_2, d_2] \times \ldots [c_n, d_n]$$

then the last integral can be written in a more student-liked form:

$$\begin{aligned} &P((X_1, X_2, \ldots, X_n) \in A) = P((X_1, X_2, \ldots, X_n) \in [a_1, b_1] \times [a_2, b_2] \times \ldots [a_n, b_n]) \\ &= \int_{[a_1, b_1] \times [c_2, d_2] \times \ldots [a_n, b_n]} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n \\ &= \int_{x_1 = a_1}^{b_1} \int_{x_2 = a_2}^{b_2} \cdots \int_{x_n = a_n}^{b_n} f(x_1, x_2, \ldots, x_n) dx_1 dx_2 \cdots dx_n \end{aligned}$$

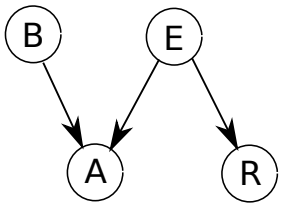
For sets A, that do not have such shape, the integral can be tried to be computed using **Integral substitution theorem** for n dimensions[2]

[2] `https://en.wikipedia.org/wiki/Integration_by_substitution`

What is a Bayesian Network

A Bayesian network is a directed acyclic graph used to represent a joint distribution between several variables.

- directed acyclic graph
- each node A is a conditional probability distribution representing $P(A|parents(A))$ where the parents of A are the incoming edges to node A.



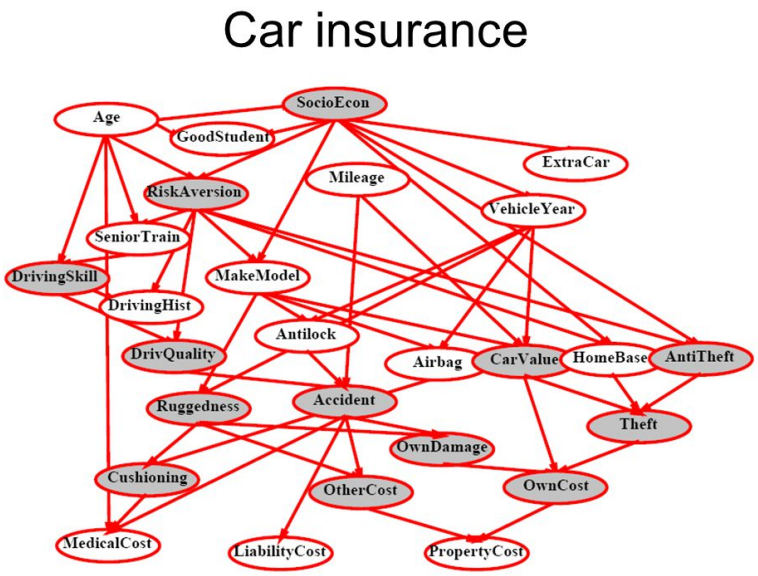
We have an alarm system, this can be activated either by a real burglary, or cause a false alarm due to earthquakes. We have 4 binary variables, A - alarm sounds or not, B - burglary occured or not, E - Earthquake, R - radio report of earthquakes.

This model encodes a joint probability as:

$$P(A, B, E, R) = P(A|B, E)P(R|E)P(B)P(E)$$

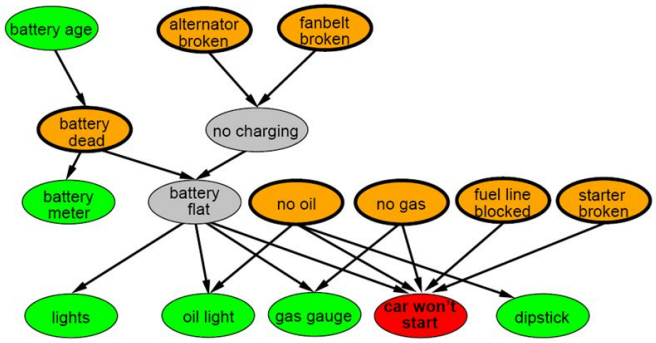
Usage:

- fit a model involving many variables
- Can use to compute marginal and conditional probabilities over a subset of variable.



A more realistic Bayes Network: Car diagnosis

- Initial observation: car won't start
- Orange: "broken, so fix it" nodes
- Green: testable evidence
- Gray: "hidden variables" to ensure sparse structure, reduce parameteres



Recap: Marginal Distribution: N variables

The marginal distribution is the distribution of a subset of the set of all variables. The idea to obtain a marginal is summing away variables that you do not want to see.

For a pmf $P(X_1, X_2, \ldots, X_n)$ we can compute marginals with $n-1, n-2, n-3, \ldots, 3, 2, 1$ variables.

Example 1: Marginal eliminating X_1:

For example, in the discrete case:

$$P(X_2, \ldots, X_n) = \sum_{x_1} P(X_1 = x_1, X_2, \ldots, X_n)$$

Whether a probability is a marginal or a joint probability – that depends on the set of variables which you consider. $P(x_2, \ldots, x_n)$ is a marginal for the set of variables $(X_1, X_2, \ldots, X_n)$, and a joint probability for the set of variables $(X_2, \ldots, X_n)$.

Example 2: Marginal eliminating X_2:

Another example, in the discrete case, now a marginal without X_2

$$P(X_1, X_3 \ldots, X_n) = \sum_{x_2} P(X_1, X_2 = x_2, \ldots, X_n)$$

We can define an marginal probability for $n-2$ variables, eliminating X_2, X_3:

Example2: Marginal eliminating X_2 and X_3:

$$P(X_1, X_4, \ldots, X_n) = \sum_{x_2} \sum_{x_3} P(X_1, X_2 = x_2, X_3 = x_3, X_4, X_5 \ldots, x_n)$$

Rule to compute marginal pmfs and pdfs:
The general rule is:

- If one wants to compute the marginal probability P for k variables in the discrete case, then one must sum over the other/remaining $n - k$ variables.
- If one wants to compute the marginal density f for k variables in the continuous case, then one must integrate over the other/remaining $n - k$ variables.

Question 1
Suppose you have a joint distribution of four variables (X_1, X_2, X_3, X_4). How many marginals consisting of $3, 2, 1$ variables? ${}^4C_3$

Question 2

- How many $n-1$-dimensional marginals we have if we had originally n variables?
- How many 1-dimensional marginals we have if we had originally n variables?
- How many 2-dimensional marginals we have if we had originally n variables?
- How many k-dimensional marginals we have if we had originally n variables? ${}^nC_k$
- How many marginals in total? $\sum_{k=0}^{n} \binom{n}{k} = 2^n$

Joint probabilities and the curse of dimensionality

Why do we need modeling at all? Why is that part of an AI lecture?

Example: n variables, each variable take value $\{0, 1\}$. Problem:

Joint distribution $P(X_1, \ldots, X_n)$ has how many parameters?

Need to encode probability $P(X_1 = x_1, \ldots X_n = x_n)$ for every possible tuple $(x_1, \ldots, x_n)$

Every state like $\underbrace{010010101110 \ldots 0101}_{n} \in X_1, \ldots, X_n$ has its own probability $P(X_1, \ldots, X_n)$. So 2^n possible states for $x_1, \ldots, x_n$.

Answer: Probability must sum up to one, so $2^n - 1$ possible values of P.

$2^{10} \approx 10^3$, $2^{50} \approx 10^{15}$, so a million times a billion. Now imagine how much data would you need to estimate 10^{15} probabilities?!

Same problem when computing a marginal distribution

$$P(X_1, X_2) = \sum_{X_3 \in \{0,1\}} \sum_{X_4 \in \{0,1\}} \sum_{X_5 \in \{0,1\}} \cdots \sum_{X_n \in \{0,1\}} P(X_1, \ldots, X_n)$$

is a sum over 2^{n-2} terms. Quickly too much to compute. Need a way to model with less parameters - we will make structural assumptions and assume conditional independence.

Recap: Independence: Two variables

Definition:

Two random variables X and Y are said to be **independent** if for every pair of x and y values

$$p(x, y) = p_X(x) \cdot p_Y(y)$$

when X and Y are discrete, or

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

when X and Y are continuous and there is a density for them.

If the above relation is **not** satisfied for all (x, y), then X and Y are said to be **dependent**.

The above definition implies that two variables are independent if their joint pmf or pdf is the product of the two marginal pmf's or pdf's.

Recap: Independence - an overview for N variables

For 2 variables it is clear: one variable can be independent the other. No more design choices left.

For n variables are multiple types of independences possible! One can ask if all N variables are independent from each other. For i.i.d. random variables, all N variables are independent from each other. However there are other types of independence: two sets of variables can be independent, even if all variables are not independent.

Recap: Independence: N variables (from each other) – case of discrete variables

Definition:

N random variables $X_1, \ldots, X_n$ are said to be **independent** (from each other) in the case of discrete variables if for every n-tuple of $x_1, \ldots, x_n$ values

$$P(X_1 = x_1, \ldots, X_1 = x_n) = P(X_1 = x_1) \cdot P(X_2 = x_2) \ldots \cdot P(X_n = x_n) = \prod_{i=1}^{n} P(X_i = x_i)$$

when X_i are discrete

If the above relation is **not** satisfied for all $(x_1, \ldots, x_n)$, then the X_i are said to be **dependent**.

If N variables are independent (from each other), then all subsets of K variables ($K < N$) are also independent from each other.

WARNING!

independence of all pairs $<$ independence of n variables In general: For n variables, independence of all pairs as in the case of 2 variables does not imply independence of n variables.

If all pairs of two variables are independent, it does NOT imply that n variables would be independent. Consider this counterexample:

$$(X, Y, Z) = \begin{cases} (0,0,0) & \text{with probability } 0.25 \\ (0,1,1) & \text{with probability } 0.25 \\ (1,0,1) & \text{with probability } 0.25 \\ (1,1,0) & \text{with probability } 0.25 \end{cases}$$

Compute $P(X = 0), P(Y = 0), P(Z = 0)$ and $P(X = 0)P(Y = 0)P(Z = 0)$. Verify that $P(Y = 0, Z = 0)$ is a product of the one-dimensional marginals.

*Recap: Independence: **between two subsets of** N variables – case of discrete variables*

Suppose we want to model 5 variables $(X_1, X_2, X_3, X_4, X_5)$. Suppose we have a probability for the first three $P(X_1, X_2, X_3)$ and a probability for the last two $P(X_4, X_5)$. You can design a probability for 5 variables from it by defining:

$$P(X_1, X_2, X_3, X_4, X_5) = P(X_1, X_2, X_3) \cdot P(X_4, X_5)$$

This property, if it holds for all possible values $x_i \in X_I$, defines independence betwee ntwo sets of variables

Above is a probability: products of nonnegative factors are nonnegative and it holds:

$$\sum_{X_1,X_2,X_3,X_4,X_5} P(X_1,X_2,X_3)\cdot P(X_4,X_5)$$
$$= \sum_{X_1,X_2,X_3}\sum_{X_4,X_5} P(X_1,X_2,X_3)\cdot P(X_4,X_5)$$

here take the $P(X_1,X_2,X_3)$ out of the second sum as it is a constant regarding the second sum

$$= \sum_{X_1,X_2,X_3} P(X_1,X_2,X_3)\cdot \sum_{X_4,X_5} P(X_4,X_5) = 1\cdot 1$$

It does **not mean that** these 5 variables are *independent from each other*. If $P(X_4,X_5) \neq P(X_4)\cdot P(X_5)$, then independence of each other will never be true.

The important thing to see is: in this case the set of variables (X_1,X_2,X_3) is independent from the set (X_4,X_5), even if the 5 variables are not independent of each other.

Lets formalize this for two sets:

Lets denote X_{Set1}, X_{Set2} to be an **ordered** subset of $(X_1,\ldots,X_n)$.

For example:

$n = 10$, $X_{Set1} = (X_2,X_6,X_9)$, $X_{Set2} = (X_1,X_3,X_4,X_5,X_7,X_8,X_{10})$ (in this case we have covered all 10 variables in Set1 and Set2)

or $X_{Set1} = (X_3,X_4,X_5)$, $X_{Set2} = (X_1,X_2,X_7)$ (in this case both sets cover only 6 out of 10 variables)

We define $X_{Set1\cup Set2}$ to be the ordered set union of these two sets.

Example:

$$X_{Set1} = (X_3,X_4,X_5,X_9), X_{Set2} = (X_1,X_2,X_7)$$
$$\rightarrow X_{Set1\cup Set2} = (X_1,X_2,X_3,X_4,X_5,X_7,X_9)$$

Definition: Independence: between two subsets of N variables

The two disjoint sets X_{Set1}, X_{Set2} of variables are independent, if

$$P(X_{Set1\cup Set2}) = P(X_{Set1})\cdot P(X_{Set2})$$

Relationship between Independence Categories

Independence of N variables is stronger than independence between two sets: if N variables are independent, then all subsets are independent, too.

*Recap: Independence: between **more than two** subsets of N variables – case of discrete variables*

Of course a probability $P(x_1, \ldots, x_n)$ can be a product of more than two factors. If it factors for example in three factors, then one would have three sets of variables $X_{set1}, X_{set2}, X_{set3}$ and

$$P(X_{Set1 \cup Set2 \cup Set3}) = P(X_{set1})P(X_{set2})P(X_{set3})$$

The important thing is to understand that independence in the case of N variables can have many forms. It is not as simple as independence between 2 variables. The simplest form of independence for N variables is that the N variables are independent from each other. This simplest form is also the strongest: It implies independence between all possible sets of variables (no matter 2,3 or any other number of sets).

Conditional Probability: Two Variables

The most important thing to understand that a conditional probability for variable X_1 conditioned on variable X_2 must sum/integrate to 1 over the variable X_1.

$$P(X_1|X_2) = \frac{P(X_1, X_2)}{P(X_2)} \text{ whenever } P(X_2) \neq 0$$

These definitions are not a surprise and can be remembered in the following way.

1. The conditional probability is the product of the joint probability and a normalizing constant (normalizer).
2. For a conditional probability $P(X_1|X_2)$ it must hold that: summing it over X_1 (the variable that is not used for conditioning) must yield 1.

   If we assume that, then we can derive the formula of conditional probability by solving for c:

$$\begin{aligned} P(X_1|X_2) &= cP(X_1, X_2) \\ 1 &= \sum_{x_1} P(X_1 = x_1|X_2) \\ \text{use: conditional is ``joint times normalizer''} \quad &\Rightarrow 1 = \sum_{x_1} cP(X_1 = x_1, X_2) \\ &\Rightarrow 1 = cP(X_2) \\ &\Rightarrow c = \frac{1}{P(X_2)} \end{aligned}$$

The important point is to remember: the conditional probability is constant times the joint probability but must sum up to 1 over all the variables that were not fixed by conditioning. Then one can immediately recover the formulas above.

Final remark: conditional probabilities are only defined where the marginal used to divide (e.g. $P(X_2)$) is non-zero. You cannot condition on an event of zero probability.

Conditional Probability: n Variables

We have the same effect as for marginals. We can have many different conditional probabilities and densities.

For example, consider 3 variables X_1, X_2, X_3. Then we can have
$P(X_1, X_3|X_2)$
$P(X_1, X_2|X_3)$
$P(X_2, X_3|X_1)$
$P(X_1|X_2, X_3)$
$P(X_2|X_1, X_3)$
$P(X_3|X_1, X_2)$
but we can also have
$P(X_3|X_1)$

Question: How many conditional probabilities do exist?

We can write down intuitively

$$P(X_1, X_3|X_2) = \frac{P(X_1, X_2, X_3)}{P(X_2)}$$
$$P(X_1|X_2) = \frac{P(X_1, X_2)}{P(X_2)}$$

Note that in the last term, the density in the enumerator (above the /) is a marginal density, too, as $X_1|X_2$ is only a subset of (X_1, X_2, X_3), so $P(X_1, X_2)$ is a marginal density

How to define all these conditional densities for n variables?

Lets denote X_{Set1}, X_{Set2} to be an **ordered** subset of $(X_1, \ldots, X_n)$. Both sets are **disjoint**.

We define $X_{Set1 \cup Set2}$ to be the ordered set union of these two sets.

Definition of conditional probability:

$$P(X_{Set1} \mid X_{Set2}) = \frac{P(X_{Set1 \cup Set2})}{P(X_{Set2})}$$

If $X_{Set1 \cup Set2}$ contains all variables in $(X_1, \ldots, X_n)$, then we have

$$P(X_{Set1} \mid X_{Set2}) = \frac{P(X_1, X_2, \ldots, X_n)}{P(X_{Set2})}$$

Again, this formula can be easily derived and remembered by understanding that

- **conditional = joint × a normalizer:** $P(X_{Set1}|X_{Set2}) = cP(X_{Set1 \cup Set2})$
- **summing this over all values for** X_{Set1} (the non-conditioning variables) must yield ONE.
  Obviously summing $P(X_{Set1 \cup Set2})$ over all values for X_{Set1} is marginalization which yields $P(X_{Set2})$

WARNING: A conditional probability sums up to one if you sum it over all its non-conditioning variables

$$\sum_{\text{all possible values for } X_{Set1}} P(X_{Set1}|X_{Set2}) = 1$$

If you compute a sum that includes at least one of the variables in the set used for conditioning, then it does **NOT** sum up to one anymore. It will be a funny number without any meaning about probabilities or densities.

$$\sum_{X_{Set1}} \sum_{X_s} P(X_{Set1}|X_{Set2}) = |dom(X_s)|,\ X_s \in X_{Set2}$$
$$\sum_{X_s} P(X_{Set1}|X_{Set2}) = \text{?? Whatever! },\ X_s \in X_{Set2}$$

Warning 2: There is a caveat to this

$$X_s \in X_{Set2} \Rightarrow \sum_{X_s} P(X_{Set1}|X_{Set2})P(X_s) = P(X_{Set1}|X_{Set2 \setminus \{s\}})$$

because $P(A|B)P(B) = P(A,B)$

an example

A	B	C	$P(A,B,C)$
0	0	0	.1
1	0	0	.2
0	1	0	.05
1	1	0	.05
0	0	1	.3
1	0	1	.05
0	1	1	.15
1	1	1	.1

$P(A=0) = 0.1 + 0.05 + 0.3 + 0.15 = 0.6$

$P(B=0) = 0.1 + 0.2 + 0.3 + 0.05 = 0.65$

- what is $P(A=0)$, $P(B=0)$?
- does knowing $P(A=0), P(B=0)$ help for computing $P(A=0, B=0)$?

No, they are not independent

$$\frac{P(A=1,\ C=0)}{P(C=0)} = \frac{0.25}{0.4} = \frac{5}{8}$$

- What is $P(A = 1|C = 0)$?
- What is $P(C = 1|B = 0)$?
- What is $P(C = 1|C = 0)$? $\longrightarrow$ 0
- What is $P(A = 0, C = 1|B = 0)$? $\longrightarrow$ $\frac{P(A=0,\ C=0,\ B=0)}{P(B=0)}$
- What is $P(A = 1, C = 1)$?

Do compute this in class please! If you fail on this, then exam might be a problem, too!

Conditional Probability and its relation to independence

If two sets of variables X_{Set1}, X_{Set2} are independent, then we have for all values $x_{Set1 \cup Set2}$:

$$P(X_{Set1}|X_{Set2}) = P(X_{Set1})$$
$$P(X_{Set2}|X_{Set1}) = P(X_{Set2})$$

Why does that hold?

by independence we have

$$P(X_{Set1 \cup Set2}) = P(X_{Set1}) \cdot P(X_{Set2})$$
$$\Leftrightarrow \frac{P(X_{Set1 \cup Set2})}{P(X_{Set1})} = P(X_{Set2})$$

definition of conditional: $\Leftrightarrow P(X_{Set2} \mid X_{Set1}) = P(X_{Set2})$

What does that mean? Consider the first equation.

$$P(X_{Set1}|X_{Set2}) = P(X_{Set1})$$

That means:

1. the conditional probability function with respect to X_{Set1} does not change when knowing the value x_{Set2} of the variables in the set X_{Set2} used for conditioning on.
2. That conditional probability with respect to X_{Set1} is same as the marginal probability over the variables X_{Set1}.

In short, knowing the value of the variable set used for conditioning, gives you no extra information about the conditional probability, if we have independence

Recap: Conditional Independence

It can happen that

$$P(A,B) \neq P(A)P(B)$$
$$P(A,B|C) = P(A|C)P(B|C) \text{ for all values of } C$$

The latter is **conditional independence**. Its just independence where you have to condition on a subset of variables. Can be more than just one variable C. Conditional independence is weaker than independence: it does not imply (unconditional) independence

Example:

P(A | C)	C=0	C=1
A=0	0.4	0.2
A=1	0.6	0.8

P(B | C)	C=0	C=1
B=0	0.5	0.2
B=1	0.5	0.8

Lets define

$$P(A,B|C) = P(A|C)P(B|C)$$

Then, conditional independence holds by design.

How about $P(A,B)$ versus $P(A)P(B)$? Lets assume $P(C=0) = 0.5$

We know

$$P(A) = \sum_{c \in dom(C)} P(A|C)P(C)$$

Then:

$$P(A=0) =? \quad \frac{0.6}{0.6+1.4} = 0.3 \qquad \sum_{c \in \{0,1\}} P(A=0|C=c)P(C=c) = 0.4 \times 0.5 + 0.2 \times 0.5 = 0.3$$
$$P(B=0) =? \quad \frac{0.7}{0.7+1.3} = 0.35$$
$$P(A=0,B=0) =? \quad \sum_c P(A=0|C=c) \cdot P(B=0|C=c) P(C=c)$$

Definition Conditional independence

We have conditional independence between two subsets (remember you can do more than two subsets for independence!) X_{Set2}, X_{Set1} conditioned on a set X_{Set5} if

$$P(X_{Set1 \cup Set2}|X_{Set5}) = P(X_{Set1}|X_{Set5})P(X_{Set2}|X_{Set5})$$

How can that be useful? Suppose you want to model something like a common cause for two effects: if it rains, then you take an umbrella, but you also take longer to work. knowing whether it rains, they become independent - umbrella and delays on the way to work. However, in general taking an umbrella and delays on the way to work are not independent, and modeling them as independent would omit valuable information.[3]

[3] Do you know that lack of sleep and stress at work is one major cause of more coding errors and weight gains? Singapore is a city that never sleeps. It is not all about sports or hungering oneself down.

What is the value of that: models with unconditionally independent variables make a very strong assumption. See the section why independence is limiting below!

But: conditional independence still allow to model processes where variables are independent when the conditioning variable is summed away!

Note: all the notions of independence applies, just with the addition that one must condition on the conditioning variables, that is: if A is independent B conditional on C, then

$$P(A|B,C) = P(A|C)$$
$$P(B|A,C) = P(B|C)$$

This is equivalent to:

$$P(X_{Set2}|X_{Set1}, X_{Set5}) = P(X_{Set2}|X_{Set5})$$
$$P(X_{Set1}|X_{Set2}, X_{Set5}) = P(X_{Set1}|X_{Set5})$$

and now to modeling with Bayesian Networks

Alarm example:

We have an alarm system, this can be activated either by a real burglary, or cause a false alarm due to earthquakes. We have 4 binary variables

- A - alarm sounds or not
- B - burglary occured or not
- E - Earthquake
- R - radio report of earthquakes

Joint probability is $P(A, R, E, B)$. How to use model assumptions to reduce parameters?

it can be always factorized by the chain rule:

$$P(A,R,E,B) = P(A|R,E,B)P(R,E,B)$$
$$P(A,R,E,B) = P(A|R,E,B)P(R|E,B)P(E,B)$$
$$P(A,R,E,B) = P(A|R,E,B)P(R|E,B)P(E|B)P(B)$$

chain rule factorization in general:

see $P(X_1, X_{set2}) = P(X_1|X_{set2})P(X_{set2})$

$$P(X_1, X_2, \ldots, X_n) = P(X_1|X_2, \ldots, X_n)P(X_2, \ldots, X_n)$$
$$P(X_1, X_2, \ldots, X_n) = P(X_1|X_2, \ldots, X_n)P(X_2|X_3 \ldots, X_n)P(X_3, \ldots, X_n)$$
$$P(X_1, X_2, \ldots, X_n) = P(X_1|X_2, \ldots, X_n)P(X_2|X_3 \ldots, X_n)P(X_3|X_4 \ldots, X_n)P(X_4, \ldots, X_n)$$
$$\ldots$$
$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_{i-1}, \ldots, X_1)$$

This can be done along any ordering of variables, not just $\{1, 2, 3, \ldots, n\}$. How many different factorizations can be written down?

Lets get back to modeling:

Alarm is not directly influenced by the radio report, if we assume that we would know whether burglary or earthquake happened

So it should be independent, **conditioned on knowledge of values of the variables for burglary and earthquake**:
$P(A|R,E,B) = P(A|E,B)$
Radio reports about earthquakes is not directly influenced by thieves, so it should be independent:
$P(R|B) = P(R)$.
Applying conditioning on the variable E in our factorization we obtain:
$P(R|E,B) = P(R|E)$

Earthquake is not directly influenced by the burglar:
$P(E|B) = P(E)$

The resulting model under these assumptions is:

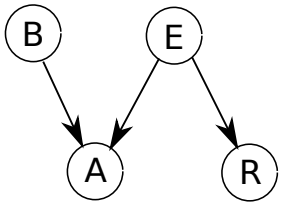
$$P(A,R,E,B) = P(A|R,E,B)P(R|E,B)P(E|B)P(B)$$
$$P(A,R,E,B) = P(A|E,B)P(R|E)P(E)P(B)$$

How many parameters does the second formultation have? $P(A|E,B)$ - 4 cases times 1 probability ($P(A = 1|\ldots) = 1 - P(A = 0|\ldots)$)
$P(R|E)$ - 2

$P(E)$ -1
$P(B)$ -1
In total 8 parameters, not 15. One can see here the reduction of parameters that need to be fitted or learned due to independence assumptions

We have used here independence and conditional independence as model inputs.

We can visualize the representation of
$P(A, R, E, B) = P(A|R, E)P(R|E)P(E)P(B)$ by a Bayesian network which is shown in the figure below



A bayesian network consists of

- a directed graph
- one conditional probability per node.
  - a node stands for the probability variable to be modeled. A node for variable R models $P(R|conditioning\ variables)$
  - incoming arcs denote dependency on conditioning variables.

The probability of R is conditioned on E and only on E because it has only one incoming arc. So it must be $P(R|conditioning\ variables) = P(R|E)$
Same: The probability of A has two arcs from B,E, so it is conditioned on B and E, and it must be $P(A|B, E)$

Bayesian networks is a way to make independence assumptions and model joint probabilities with less parameters.

Bayesian network representation:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|parents(X_i))$$

where parents are given by the arc structure in the Bayesian network.

Compare to the general chaining case:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i|X_{i-1}, \ldots, X_1)$$

The difference is: due to independence assumptions we do not condition the probability for X_i anymore on the whole chain $X_{i-1}, \ldots, X_1$ - but only on a few selected parents. We condition on those parents where we think that independence with respect to the parents **does not hold**.

Pro: more efficient to compute due to independence assumptions

Con: independence assumptions limit what you can model - independence means cannot model a relationship

Why independence is limiting

Lets see this in an example:

Assume $X \in \{0, 1, 2, 3, 4, 5\}$ is independent of $Y \in \{0, 1, 2, 3, 4, 5\}$, that is $P(X, Y) = P(X)P(Y)$. Suppose you want to train a regression function $y = f(x)$ between x and y. You aim to minimize square loss.

$$L(f(x), y) = \mathbb{E}_{(x,y)\sim P}[(y - f(x))^2]$$

What would be the best model?

Consider $p(Y|X)$

$$\begin{aligned} L(f(x), y) &= \mathbf{E}[(y - f(x))^2] \\ &= \sum_{x_i, y_k} (y_k - f(x_i))^2 p(x_i, y_k) \\ = \sum_{x_i} p(x_i) \sum_{y_k} p(y_k)(y_k - f(x_i))^2 \end{aligned}$$

So you need to find for every x_i the value $f_i = f(x_i)$ which minimizes

$$\sum_{x_i} p(x_i) \sum_{y_k} p(y_k)(y_k - f_i)^2$$

Note that solving

$$\frac{\partial}{\partial f_i} \sum_{x_i} p(x_i) \sum_{y_k} p(y_k)(y_k - f_i)^2 = 0$$

results in:

$$\begin{aligned} 0 &= -p(x_i) \sum_{y_k} p(y_k)(y_k - f_i) \\ \Leftrightarrow 0 &= 1 \cdot \sum_{y_k} p(y_k)(y_k - f_i) \end{aligned}$$

which is not the same solution for all f_i!!!

$$f_i = \frac{\sum_{y_k} p(y_k) y_k}{\sum_{y_k} p(y_k)} = \sum_{y_k} p(y_k) y_k$$

This solution is for every x_i the same and does NOT depend on x_i at all. So independence means: cannot model any meaningful relation between the independent variables.

We have seen: Independence is too limiting for meaningful modeling, but we can use conditional independence, which does NOT imply independence.

Burglary example in math, Explaining Away

Tasks:

- An alarm is observed, what is the probability of burglary given the bayesian net? $P(B=1 \mid A=1) = \frac{P(B=1, A=1)}{P(A=1)} = \frac{\sum_{E=e} P(B=1, A=1 \mid E=e) \cdot P(E=e)}{P(A=1)}$ → get from table)
  We want what probability term?

- An alarm is observed, and we hear a radio report about the earth quake, what is the probability of burglary given the bayesian net? We want what probability term?

$P(A=1|B,E) =$

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

$P(A=1) = 0.9999 \times 0.01 \times (1-0.000001) + 0.99 \times 0.01 \times 0.000001 + \ldots$

$P(R|E) =$

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

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

Compare these two probabilities: Knowing that an earthquake occurred, "explains away" to an extent the fact that the alarm is ringing. The latter is lower

Practical take away on how to use bayes nets

Explaining away is the fact that probabilities may change if we condition on an event. Regarding this event: without knowledge

whether it happened or not– we have to marginalize its variable out instead of conditioning on it.

The point is here: We want to compute some probability of something. if we know nothing about a variable, then we must marginalize out over it (sum it out/integrate it out if has density)

If we know that an event happened regarding this variable, we can condition on this variable

How to use bayesian networks (once you have probabilities):

The general idea: you need a distribution over a marginal subset of variables

- if you know nothing else, then you must marginalize out all other variables.
- if you know additional events whether they happened or not (as with the radio report), then you can condition on those variable and by this way input knowledge about events.

Using conditioning does NOT imply causality

Common mistake: causation can be used to build BNs but **Conditioning does not imply causality of effects!**

Consider the probability of rain (R) and traffic jams (T). Once we have a joint distribution $P(R, T)$, we can write
We can compute both $P(T|R)$ and $P(R|T)$ from the joint $P(R, T)$, and it gives the same joint back:

$$P(R) = \sum_T P(R, T)$$
$$P(R, T) = P(T|R)P(R)$$
$$P(T) = \sum_R P(R, T)$$
$$P(R, T) = P(R|T)P(T)$$

A model $P(R, T) = P(R|T)P(T)$ does not imply that traffic jams cause rain. It is merely based on the probabilities $P(R|T)$ of rain given that we observed traffic jams $T = 1$ (or given that we observed a free road $T = 0$).

R	T	$P(R,T)$
1	1	0.4
0	1	0.25
1	0	0.2
0	0	0.15

Compute $P(T|R)$ and $P(R|T)$!

An important application

We can use Bayes Nets for **marginalization** in Bayesian networks

Since

$$P(A,B) = P(A|B)P(B)$$

we can compute

$$P(A) = \sum_B P(A,B) = \sum_B P(A|B)P(B)$$

This holds in general (provided that X_{Set1}, X_{Set2} are disjoint sets):
marginalization with conditional probabilities

- input: $P(X_{Set1}|X_{Set2})$
- goal: need the value for $P(X_{Set1})$
- then use formula below:

$$P(X_{Set1}) = \sum_{X_{Set2}} P(X_{Set1}|X_{Set2})P(X_{Set2})$$

Of course, this holds also, when we apply conditioning by another set X_{Set3} to both sides of the equation aboce

$$\begin{aligned} P(X_{Set1}|X_{Set3}) &= \sum_{X_{Set2}} P(X_{Set1}|X_{Set2}, X_{Set3})P(X_{Set2}|X_{Set3}) \\ &= \sum_{X_{Set2}} P(X_{Set1}|X_{Set2 \cup Set3})P(X_{Set2}|X_{Set3}) \end{aligned}$$