50.021 -AI

Alex

Week 12: Probabilistic Graphical Models

[The following notes are compiled from various sources such as textbooks, lecture materials, Web resources and are shared for academic purposes only, intended for use by students registered for a specific course. In the interest of brevity, every source is not cited. The compiler of these notes gratefully acknowledges all such sources.]

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

Recap

Recap: How to use Bayesian nets

- you have a BN modeling $P(X_1, \ldots, X_n)$. You are interested in a model for a joint distribution over $X_{set1} \subset (X_1, \ldots, X_n)$
- You have observed events in $X_{set2} \subset (X_1, \ldots, X_n)$. You can condition on all events that you have observed.
- Usually one then computes $P(X_{set1}|X_{set2})$ - this involves summing out/marginalizing out all variables in $(X_1, \ldots, X_n) \setminus X_{set1 \cup set2}$, that is all variables which are not in X_{set1} or X_{set2}
- conditioning $P(\cdot|X_{set2})$ is inputting information about observed events.

In this class you should see – only for discrete variables

- how to read off what conditional independences are implied in a bayesian network, that is: what relationships you can or cannot model

The question is: under what conditions two variables are independent?

In Bayesian Nets there are more independences than you have putted in!

We do encode conditional independence in Bayesian networks by making design choices like

$$P(X_i|X_{i-1}, X_{i-2}, \ldots, X_1) = P(X_i|parents(X_i))$$

where $parents(X_i)$ refers to the BN structure, namely the incoming arcs to the node X_i

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

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

holds for all values of these variables. This is equivalent to

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

Three disjoint sets $X_{Set1}, X_{Set2}, X_{Set3}$ of variables are independent, if

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

holds for all values of these variables.

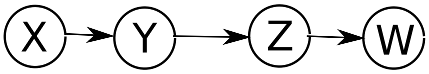
We have conditional independence between two subsets X_{Set2}, X_{Set1} conditioned on a set X_{Set5} if

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

This is equivalent to

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

The BN has often more independencies in it!



$$P(X, Y, Z, W) = P(W|X, Y, Z)P(Z|X, Y)P(Y|X)P(X)$$
$$P(X, Y, Z, W) = P(W|Z)P(Z|Y)P(Y|X)P(X)$$

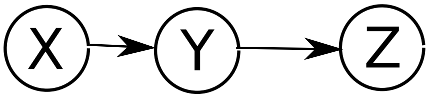
so: $W \perp (X, Y)|Z$, $Z \perp X|Y$,ok.

Suppose you observe Y and condition on Y, there is one more, not directly encoded!

Question: When are two nodes independendent given certain evidence (that is encoded by conditioning on $Variable = value$, if evidence is that this variable takes this value)? -

Causal Chain

A causal chain is a graph like that



Lets consider it only for X-Y-Z

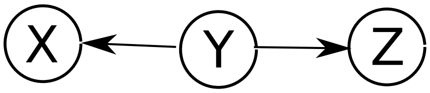
- X is independent of Z? yes or no. Consider the case of: thick clouds → rain → traffic jams
- X is independent of Z conditioned on Y!

to see the second one, consider: $P(Z|X,Y)$

$$\begin{aligned} P(Z|X,Y) &= \frac{P(X,Y,Z)}{P(X,Y)} \\ P(X,Y,Z) &= P(Z|Y)P(Y|X)P(X) \\ \Rightarrow P(Z|X,Y) &= \frac{P(Z|Y)P(Y|X)P(X)}{P(Y|X)P(X)} = P(Z|Y) \end{aligned}$$

so $P(Z|X,Y) = P(Z|Y)$!

Common Cause



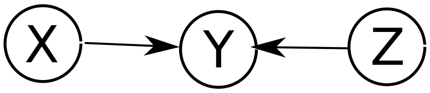
- X is independent of Z? yes or no. Consider the case of: emails written at night← capstone deadline is near → students have huge eyesacks
- X is independent of Z conditioned on Y!

to see the second one, consider: $P(Z|X,Y)$

$$\begin{aligned} P(Z|X,Y) &= \frac{P(X,Y,Z)}{P(X,Y)} \\ P(X,Y,Z) &= P(Z|Y)P(X|Y)P(Y) \\ \Rightarrow P(Z|X,Y) &= \frac{P(Z|Y)P(X|Y)P(Y)}{P(X|Y)P(Y)} = P(Z|Y) \end{aligned}$$

- Two effects can influence each other/ be correlated via a common cause.
- but: Observing the cause blocks influence between effects.

Common Effect



- X is independent of Z? yes or no. Consider the case of: cook sucks → soup too salty ← freshly in love
- X is **not independent of Z conditioned on Y!!!**

to see the first one, consider: $P(X,Z)$

$$\begin{aligned} &P(X,Y,Z) = P(Y|Z,X)P(X)P(Z) \\ &\Rightarrow P(X,Z) = \sum_y P(Y|Z,X)P(X)P(Z) = P(X)P(Z) \end{aligned}$$

Why does the second thing hold?

Telling you that the cook is actually good at cooking tells you nothing whether he is in love but ...

You observe the soup is way too salty. Telling that the cook is actually good at cooking when you know that the soup is too salty tells you now something about the node "freshly in love", namely that it must be true.

to see the second one formally:

$$\begin{aligned} P(X,Y,Z) &= P(Y|Z,X)P(X)P(Z) \\ \Rightarrow P(Y) &= \sum_{z,x} P(Y|Z=z,X=x)P(X=x)P(Z=z) \end{aligned}$$

$$P(X,Z|Y) = \frac{P(Y|Z,X)P(X)P(Z)}{\sum_{z',x'} P(Y|Z,X=x')P(X=x')P(Z=z')}$$

This does not factorize into an X-term and an Z-Term
can also check

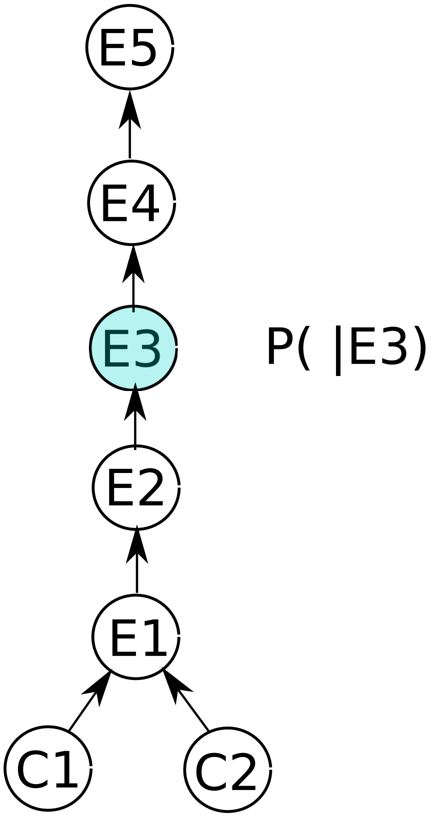
$$\begin{aligned} P(X=x|Z=z,Y=y) &= \frac{P(X=x,Y=y,Z=z)}{P(Y=y,Z=z)} \\ &= \frac{P(Y=y|Z=z,X=x)P(X=x)P(Z=z)}{\sum_{x'} P(Y=y|Z=z,X=x')P(X=x')P(Z=z)} \\ &= \frac{P(Y=y|Z=z,X=x)P(X=x)}{\sum_{x'} P(Y=y|Z=z,X=x')P(X=x')} \\ \neq P(X=x|Y=y) \end{aligned}$$

This still depends on the value z of variable Z! So **no independence**

- seeing an effect puts the causes in competition for explaining it!!
- Observing an effects activates influence between causes

Common Effect generalized

If you observe a common effect indirectly, then it changes probabilities observed down the chain, and weakly, but still existingly can put common causes into competition.



an example:

you know that colleague X is a very good motorbike driver. So no matter what, if his motorbike starts, then he will be at work **not late** in 99.99% of all cases. If his motorbike does not start, then he is late to work in 50% of all cases. His bike starts in 99% of all cases.

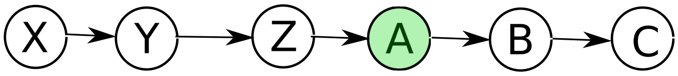
You know also a network like:

dead battery → motorbike will not start ← out of fuel

If you observe that he is late to work it means: a very high probability that his motorbike did not start $\frac{0.5 \cdot 0.01}{0.5 \cdot 0.01 + 0.0001 \cdot 0.99} = \frac{0.5}{0.5+0.01} \approx$ 0.99.

if you additionally know that his battery could not have been dead, and he is late to work (which in above example means with 99% prob that his bike did not start, and only 1% due to other causes), then very likely he was out of fuel.

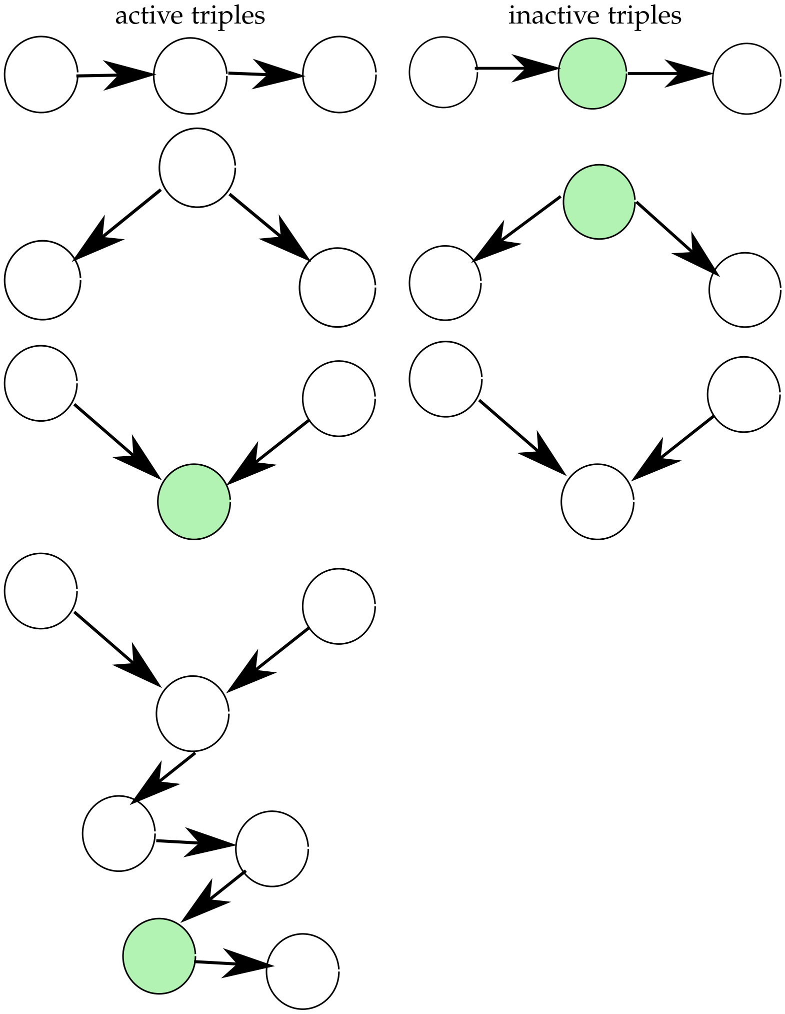
this holds all for longer chains



It is clear that observing a breaks the causal chain between X and C: X and C are conditionally independent on A.

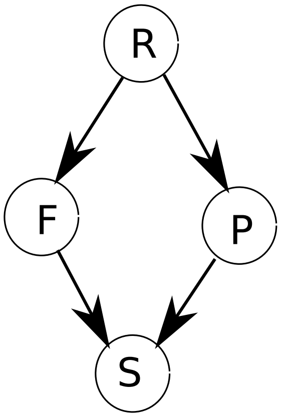
D-separation algorithm

- are X and Y conditionally independent given evidence in variables $Z_1, \ldots Z_k$?
- consider all paths in the undirected graphs from X to Y
- if all paths are inactive, then above is true


active triples

inactive triples

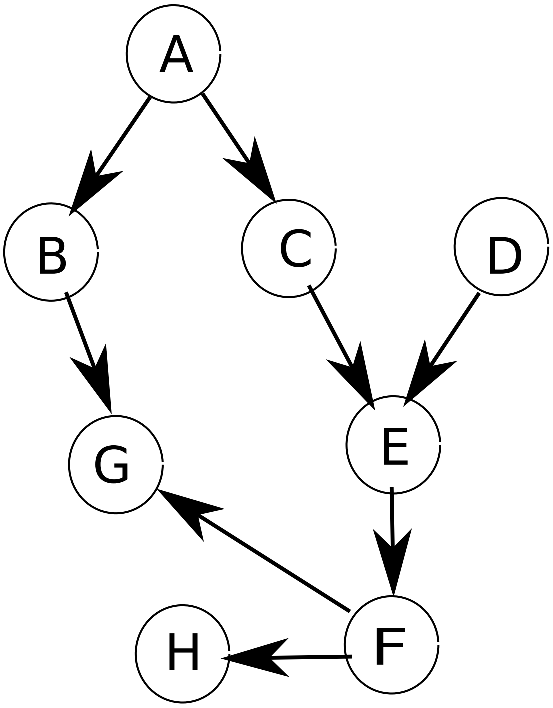
Example 1



R - heavy flood type rain, F - ponding outside of my home, P - power outage, S - me sad

- $F \perp P$? ✗
- $F \perp P | R$? ✓
- $F \perp P | R, S$? ✗

Example 2



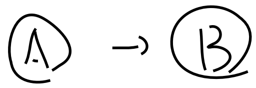
- $B \perp C | A$?, $A \perp F | E$?, $C \perp D | F$?, $A \perp G | B, F$?

  ✓ ✓ ✗ ✓

Can it happen that two nodes are not d-separated but still independent?

Yes

But the reverse is confirmed



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

How to fit a model - the simplest method

We have designed a Bayesian network. We have data. How to fit the parameters?

Suppose we have a node $P(c|s,b)$ - the probability of lung cancer c, conditioned on smoking (s) and asbestos exposure (a). Now we have a dataset of triples $x_i = (a_i, s_i, c_i), i = 1, \ldots, 10$.

What is a natural way to compute it? Empirical estimate

$$\hat{P}(c = 1|a = 1, b = 0) = \frac{\#\{(a,b,c)|a = 1, b = 0, c = 1\}}{\#\{(a,b,c)|a = 1, b = 0, c \in \{0,1\}\}}$$

Cannot be used to compute events such that conditioning has no data. That is: if $a = 1, b = 1$ has no samples, then this is not usable → divide by zero

Drawback: Unobserved data obtains zero probability.

This might be also undesirable, even when we can compute conditional probabilities:

Suppose one learns probabilities of words over a large vocabulary with many words. Unobserved words get zero probability. This is a kind of overfitting. It might be useful to have for every word a small non-zero probability, even if you never observed it, simply because your set of words might be too small.

Laplace[1] correction is one possible remedy: Suppose a random variable has K outcomes $p_1, \ldots, p_K$, then use

$$\hat{p}_i = \frac{\text{\#observed outcomess of type } i + 1}{\text{total number of all observations} + K}$$

What happened here? One pretends to have observed for every outcome one more sample. A ghost sample. In total we have K additional observations, thats why we must divide by the number of original observations $+K$,

This can be generalized to have more smoothing s with s observed ghost samples for every outcome:

$$\hat{p}_i = \frac{\text{\#observed outcomess of type } i + s}{\text{total number of all observations} + s \cdot K}$$

As $s \to \infty$, the smoothed estimates converges towards the uniform distribution, where every outcome has same probability.

[1] https://en.wikipedia.org/wiki/Pierre-Simon_Laplace

Laplace correction can be used with $s > 0$, even if $s > 0$ is a small **non-integer** such as 0.09 :).

Can we use this idea for $\hat{P}(c = 1|a = 1, b = 0)$ above ? Yes!

$$P(c = 1|a = 1, b = 0) = \frac{P(c = 1, a = 1, b = 0)}{P(a = 1, b = 0)}$$

Can use laplace correction for estimates of $P(c = 1, a = 1, b = 0)$ and $P(a = 1, b = 0)$. No zero prob failures anymore.

pro : no zero prob failure

con : bias

Laplace correction is theoretically justifiable as a maximum a posteriori estimate, that is maximum likelihood with a prior distribution on the variable to be modeled - we will see it on Monday.