Belief nets

Today:

1. understand the relationship between conditional independence, factorization, and the structure of a belief net
2. “explaining away”

"it's all in the joint"

Suppose we observe x_1, and want to know the probability distribution $p(x_3|x_1)$, a vector with n elements.

$$p(x_3|x_1) = \frac{p(x_1, x_3)}{p(x_1)} = \frac{\sum_{x_2} p(x_1, x_2, x_3)}{\sum_{x_2} \sum_{x_3} p(x_1, x_2, x_3)}$$

Joint probability table

x_1	x_2	x_3	$p(\mathbf{x})$
0	0	0	0.30
0	0	1	0.02
0	1	0	0.21
0	1	1	0.01
1	0	0	0.11
1	0	1	0.20
1	1	0	0.09
1	1	1	0.06
			1.00

Note the denominator is numerator summed over the alternatives for x_3

Any query about a conditional distribution can be found by summing up probabilities in the joint

factors

By the product rule we have

$$\begin{aligned} p(x_1, x_2, x_3) &= p(x_2, x_3 | x_1)\, p(x_1) \\ &= p(x_3 | x_1, x_2)\, p(x_2 | x_1)\, p(x_1) \end{aligned}$$

known as *factorizing* the joint.

The three terms are termed *factors*.

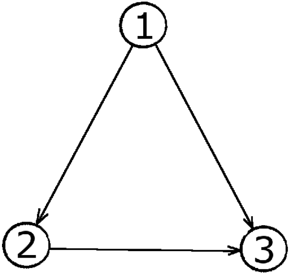
Here each factor consists of a conditional probability table.

factorizations are graphs

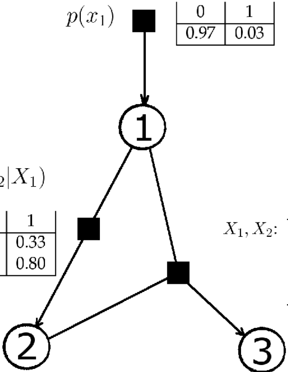
A factoring of the joint arrived at via the product rule can be represented as a graph called a *directed graphical model* or *belief net*.

$$p(x_1, x_2, x_3) = p(x_3|x_1, x_2)\; p(x_2|x_1)\; p(x_1)$$

belief net



factor graph



$p(x_1)$

0	1
0.97	0.03

$p(x_2|X_1)$

X_1:		0	1
	0	0.67	0.33
	1	0.20	0.80

$p(x_3|x_1, x_2)$

X_1, X_2:		0	1
	00	0.06	0.94
	01	0.33	0.67
	10	0.10	0.90
	11	0.99	0.01

Figure 3.4: Two BNs for a 4 variable distribution. Both graphs (a) and (b) represent the *same* distribution $p(x_1, x_2, x_3, x_4)$. Strictly speaking they represent the same (lack of) independence assumptions – the graphs say nothing about the content of the tables. The extension of this 'cascade' to many variables is clear and always results in a Directed Acyclic Graph.

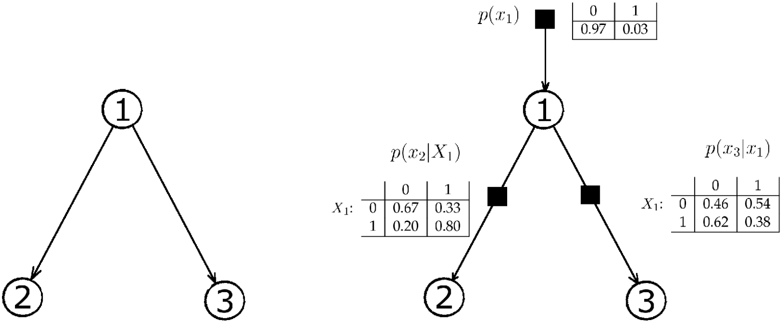
- as it stands, several graphs could encode the *same* joint
- we always get a directed, acyclic graph (DAG)
- any such graph has at least one *ancestral ordering*: an ordering of nodes such that every node's ancestors in the graph precede that node in the ordering
- intractable in two ways

We will **make assumptions**. This will **buy tractability**.

delete 2→3

Deleting the $2 \to 3$ link leaves the belief net and factor graph shown below, and the joint can be read off as

$$p(x_1, x_2, x_3) \quad = \quad p(x_3|x_1) \;\; p(x_2|x_1) \;\; p(x_1)$$



Compare this to the original joint: the x_3 term no longer depends explicitly on x_2 but only on x_1.

Look at the probability of node 2 given the other two:

$$\begin{aligned}
p(x_2|x_1, x_3) &= \frac{p(x_1, x_2, x_3)}{p(x_1, x_3)} && \text{by product rule} \\
&= \frac{p(x_3|x_1)\, p(x_2|x_1)\, p(x_1)}{p(x_3|x_1)\, p(x_1)} && \text{using the joint} \\
&= p(x_2|x_1)
\end{aligned}$$

Thus x_2 and x_3 are *conditionally independent given* x_1.

You could write this as $x_2 \perp\!\!\!\perp x_3 \mid x_1$

Now the conditional probability tables contain only 5 entries (was 7), and so

- the conditional independence constrains the values that the full look-up table can contain,
- the conditional probability tables implicitly specify the full joint but in fewer numbers.

In the fully connected graph, $x_2 \not\perp\!\!\!\perp x_3$ in the joint, and $x_2 \not\perp\!\!\!\perp x_3 \mid x_1$

In this new graph, $x_2 \not\perp\!\!\!\perp x_3$ in the joint as before, but now $x_2 \perp\!\!\!\perp x_3 \mid x_1$.

Deleting any link in the graph directly implies new independencies between variables, and vice versa.

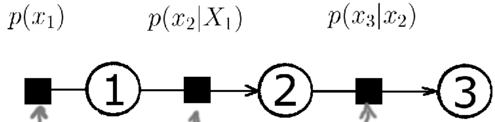
So a belief net is an encoding of a set of independence relationships between variables.

delete 1→3

$$p(x_1, x_2, x_3) \;=\; p(x_3|x_2)\;\; p(x_2|x_1)\;\; p(x_1)$$

belief net

factor graph



0	1
0.97	0.03

X_1:	0	1
0	0.67	0.33
1	0.20	0.80

X_2:	0	1
0	0.16	0.84
1	0.25	0.75

Now the x_3 term no longer depends explicitly on x_1, and again the conditional table for node 3 has fewer numbers to specify. It's easy to show that nodes 1 and 3 are now conditionally independent given a state of node 2:

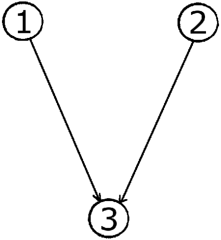
$$
\begin{aligned}
p(x_1|x_2, x_3) &= \frac{p(x_1, x_2, x_3)}{p(x_2, x_3)} \\
&= \frac{p(x_1)\, p(x_2|x_1)\, p(x_3|x_2)}{p(x_3|x_2)\, p(x_2)} \\
&= p(x_1|x_2)
\end{aligned}
$$

So $x_1 \not\perp\!\!\!\perp x_3$ in the joint, but $x_1 \perp\!\!\!\perp x_3 \mid x_2$.

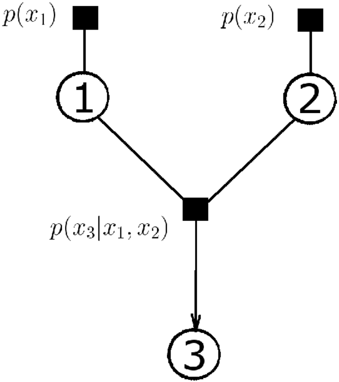
delete 1→2

$$p(x_1, x_2, x_3) = p(x_3|x_1, x_2)\ p(x_2)\ p(x_1)$$

belief net



factor graph



explaining away

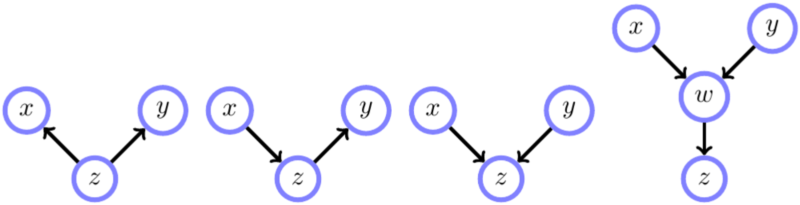
In this case nodes 1 and 2 are independent *before* node 3 is known:

$$\begin{aligned} p(x_1, x_2) &= \sum_{x_3} p(x_1, x_2, x_3) \\ &= \sum_{x_3} p(x_3|x_1, x_2)\, p(x_2)\, p(x_1) \\ &= p(x_2)\, p(x_1) \sum_{x_3} p(x_3|x_1, x_2) \\ &= p(x_2)\, p(x_1) \end{aligned}$$

but they become *dependent* if X_3 is observed - to see this, try reducing $p(x_1|x_2, x_3)$ to $p(x_1|x_3)$ and convince yourself that it can't be done!

The classic example (from Pearl) is the "burglar alarm problem"...

examples



(a) (b) (c) (d)

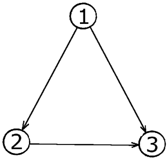
In (a) and (b), observing variable z doesn't generate new dependencies.

In (c) it does.

Question: in (d), would finding out z affect $P(x, y)$?

some motifs in belief nets

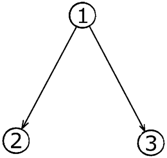
fully connected



$$x_i \not\perp\!\!\!\perp x_j$$

$$x_i \not\perp\!\!\!\perp x_j \mid x_k$$

naive Bayes



$$x_2 \not\perp\!\!\!\perp x_3$$

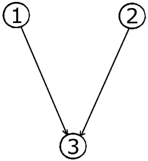
$$x_2 \perp\!\!\!\perp x_3 \mid x_1$$

a chain

$$x_1 \not\perp\!\!\!\perp x_3$$

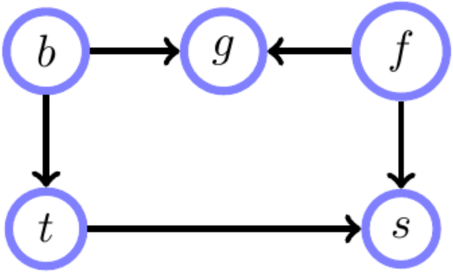
$$x_1 \perp\!\!\!\perp x_3 \mid x_2$$

explaining away



$x_1 \perp\!\!\!\perp x_2,$ and yet

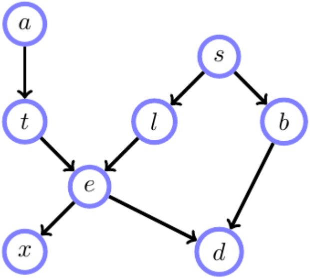
$$x_1 \not\perp\!\!\!\perp x_2 \mid x_3$$



Are t and f unconditionally independent? ie. is $t \perp\!\!\!\perp f \mid \emptyset$?

What about $t \perp\!\!\!\perp f \mid g$?

directed graphical models (belief networks)



$x =$ Positive X-ray
$d =$ Dyspnea (Shortness of breath)
$e =$ Either Tuberculosis or Lung Cancer
$t =$ Tuberculosis
$l =$ Lung Cancer
$b =$ Bronchitis
$a =$ Visited Asia
$s =$ Smoker

Given a particular belief net, the fundamental question we want to be able to ask is “what is the posterior probability distribution of variable j?”. That is, what is the vector

$$p(x_j \mid \text{variables for which we have observed values})$$

e.g. ”What is the probability the patient has lung cancer, given they have bronchitis and a positive X-ray result, and they visited Asia (but with all other variables unknown)?”

conditional independence relationships

Suppose we've assigned nodes indices in an ancestral ordering. The completely unconstrained factorization of the joint is

$$p(x_1, \ldots, x_K) \; = \; \prod_{i=1}^{K} p(x_i | x_1, \ldots, x_{i-1})$$

but our belief net is using this instead:

$$\prod_{i=1}^{K} p(x_i | \mathrm{parents}_i)$$

belief net factorization:

$$p(x_i | x_1, \ldots, x_{i-1}) \; = \; p(x_i | \mathrm{parents}_i)$$

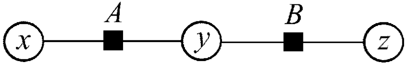
undirected PGMs (a.k.a. Markov random fields)

Graphical models describe joint probability distributions that *factor*. As we've seen, one way a distribution can factor is via application of the product rule to the joint as in, say, $p(x, y, z) = p(x)\ p(y|x)\ p(z|x, y)$, which corresponds to a directed graph called a Belief Net. However other factorisations exist. For example we could have

$$p(x, y, z) = \frac{1}{Z}\ \ \phi_A(x, y)\ \ \phi_B(y, z)$$

where Z is a normalisation factor. The ϕ are usually called "potentials".

Eg. if x, y, x are binary, ϕ_A and ϕ_B are 2x2 tables.



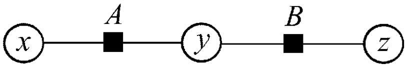
The potentials ϕ need only be positive.

One way to ensure this positivity is to use exponentials of another function: $\phi_A = e^{E_A}$. That way the function E is free to roam over *any* values. Then we have

$$p(x, y, z) \;=\; \frac{1}{Z} e^{E_A(x,y)} e^{E_B(y,z)} \;=\; \frac{1}{Z} e^{E_A(x,y)+E_B(y,z)}.$$

Physicists note: the E are completely analogous to (negative) energies in a physical system with Boltzmann distribution p

Note that the potentials ϕ *don't* need to be normalised along either their rows or columns.



Are x and z conditionally independent given y?

$$p(x,z) = \sum_{Y} p(x,y,z) \qquad \propto \sum_{Y} \phi_A(x,y)\ \phi_B(y,z)$$

It seems clear that they won't de-couple if we don't know y. But:

$$\begin{aligned} p(x,z|y) &= \frac{p(x,y,z)}{\sum_x \sum_Z p(x,y,z)} \qquad \propto \phi_A(x,y)\, \phi_B(y,z) \\ &= p(x|y)\, p(z|y) \end{aligned}$$

Once we know y, the distribution $p(x,z|y)$ factors.

In an undirected graph, a variable becomes conditionally independent of *all other* variables, given its neighbours.

In *all* PGMS

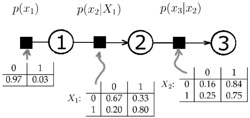
Markov blanket

a variable is conditionally independent of everything else, given the values in its Markov blanket

(*ie.* the variables one step away in the factor graph)

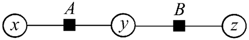
PGM summary

directed



- each factor is normalised
- product of all factors is automatically normalised
- can exhibit “explaining away”
- arrows are suggestive of a “causal” interpretation

undirected



- factors aren’t normalised
- product of all factors is not normalised
- no “causal” interpretation?
- seem to be a superset of directed models in fact...

Graphical models **simplify** the full joint by making **assumptions** about conditional independencies between variables.

inference in PGMs

Denote the nodes in a graphical model that we know the values of by "$\texttt{obs}$" (short for "observations").

three tasks:

1. infer $p(x|\texttt{obs})$, for any query variable x in the light of any set of observed variables $\texttt{obs}$.
2. infer the *most likely joint state* $p(\mathbf{x}|\texttt{obs})$, for all nodes simultaneously.
3. improve the tables (or learn from scratch) using a data set.

The first task is solved by the SUM-PRODUCT algorithm, a.k.a. "probability propagation", "belief propagation", the "forward-backward algorithm", and "turbo decoding".