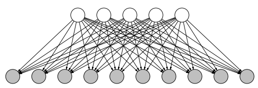
Bayesian networks: definitions



- In this module, I'll present the formal definition of Bayesian networks, give a few examples, and talk about an important property called explaining away.

Review: probability

Random variables: sunshine $S \in \{0, 1\}$, rain $R \in \{0, 1\}$

Joint distribution (probabilistic database):

$\mathbb{P}(S, R) =$

s	r	$\mathbb{P}(S = s, R = r)$
0	0	0.20
0	1	0.08
1	0	0.70
1	1	0.02

Marginal distribution: (aggregate rows)

$\mathbb{P}(S) =$

s	$\mathbb{P}(S = s)$
0	0.28
1	0.72

Conditional distribution: (select rows, normalize)

$\mathbb{P}(S \mid R = 1) =$

s	$\mathbb{P}(S = s \mid R = 1)$
0	0.8
1	0.2

- Before introducing Bayesian networks, let's review some basic probability. We start with an example about the weather. Suppose we have two boolean random variables, S and R representing whether there is sunshine and whether there is rain, respectively. Think of an assignment to (S, R) as representing a possible state of the world.
- The **joint distribution** specifies a probability for each assignment to (S, R) (state of the world). We use lowercase letters (e.g., s and r) to denote values and uppercase letters (e.g., S and R) to denote random variables. Note that $\mathbb{P}(S = s, R = r)$ is a probability (a number) while $\mathbb{P}(S, R)$ is a distribution (represented by a table of probabilities). We don't know what state of the world we're in, but we know what the probabilities are (there are no unknown unknowns). Think of the joint distribution as one giant (probabilitsic) database that contains full information about how the world works.
- Sometimes, we might only be interested in a subset of the variables, e.g., sunshine S. From the joint distribution, we can derive a **marginal distribution** over that. In the case of S, we get this by summing the probabilities of the rows in the joint distribution table that share the same value of S. The interpretation is that we are interested in (the marginal probability of) S. We don't explicitly care about R, but we still need to take into account R's effect on S. We say in this case that R is **marginalized out**.
- Sometimes, we might observe evidence; for example, suppose we know that there's rain ($R = 1$). Again from the joint distribution, we can derive a **conditional distribution** of the remaining variables (S) given this evidence $R = 1$. We do this by selecting rows of the table matching the condition and then normalizing the remaining probabilities so that they sum to 1. Note that this normalization constant is exactly $\mathbb{P}(R = 1)$.

Review: probability

Variables: S (sunshine), R (rain), T (traffic), A (autumn)

Joint distribution (probabilistic database):

$$\mathbb{P}(S, R, T, A)$$

Marginal conditional distribution (probabilistic inference):

- **Condition** on evidence (traffic, autumn): $T = 1, A = 1$
- Interested in **query** (rain?): R

$$\boxed{\mathbb{P}(\underbrace{R}_{\text{query}} \mid \underbrace{T = 1, A = 1}_{\text{condition}})}$$

(S is **marginalized out**)

- Let us augment our running example with two other random variables, T (whether there is traffic) and A (whether it's autumn).
- We have a joint distribution, which again can be thought of as a probabilistic database that tells us how the world works.
- Probabilistic inference is the process of answering questions against this database. In general, we can both condition on evidence and be interested in a subset of the remaining variables at the same time.
- For example, we might **condition** on there being traffic and the fact that it's autumn.
- And we might be interested in whether there is rain (called the **query** variable), marginalizing out **sunshine**.
- The set of conditioning variables, query variables, and variables that are marginalized out should form a partitioning of all the variables.

A puzzle

Problem: earthquakes, burglaries, and alarms

Earthquakes and **burglaries** are independent events (probability ϵ).

Either will cause an **alarm** to go off.

Suppose you get an **alarm**.

Does hearing that there's an **earthquake** increase, decrease, or keep constant the probability of a **burglary**?

Joint distribution:

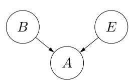
$$\mathbb{P}(E, B, A)$$

Questions:

$$\mathbb{P}(B = 1 \mid A = 1) \quad \textbf{?} \quad \mathbb{P}(B = 1 \mid A = 1, E = 1)$$

- Let's consider a classic puzzle, which we will tackle with Bayesian networks. Suppose that in the world, earthquakes and burglaries are independent (and hopefully rare) events, and for the sake of simplicity, assume that each one has a probability ϵ (say 0.05) of happening. You have installed an alarm that will notify you if either one happens.
- Now suppose you are away on vacation and you get an alarm notification on your phone. You would expect at this point that the probability of your home being burglarized has gone up. But suppose then you see breaking news saying that there was an earthquake near your home. How does that change your beliefs about the burglary?
- One could try to intuit the answer, but this is risky because sometimes the right answer is counterintuitive. In this case, you might think since earthquakes and burglaries are independent, that the probability shouldn't change. But that would be wrong. So let's use Bayesian networks instead to perform this type of **reasoning under uncertainty** in a principled way.
- Let us try to write down this question using the language of probability. The first step is to always figure out the variables of interest, which in this case are earthquake E, burglary B, and alarm A.
- We then have a joint distribution over these variables, which we will define later. But first the questions. We are interested in comparing the probability of a burglary given an alarm only versus given alarm and earthquake.

Bayesian network (alarm)



b	$p(b)$
1	ϵ
0	$1-\epsilon$

e	$p(e)$
1	ϵ
0	$1-\epsilon$

b	e	a	$p(a \mid b, e)$
0	0	0	1
0	0	1	0
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	0
1	1	1	1

$p(b) = \epsilon \cdot [b = 1] + (1 - \epsilon) \cdot [b = 0]$

$p(e) = \epsilon \cdot [e = 1] + (1 - \epsilon) \cdot [e = 0]$

$p(a \mid b, e) = [a = (b \vee e)]$

$$\mathbb{P}(B = b, E = e, A = a) \stackrel{\text{def}}{=} p(b)p(e)p(a \mid b, e)$$

- Now let us define the joint distribution. Recall the first step was just to define the three variables, B (burglary), E (earthquake), and A (alarm).
- Second, we connect up the variables to model the dependencies. Unlike in factor graphs, these dependencies are represented as **directed** edges. You can intuitively think about the directionality as representing causality, though what this actually means is a more complex issue and beyond the scope of this module.
- Third, for each variable, we specify a **local conditional distribution** of that variable given its parent variables. In this example, B and E have no parents while A has two parents, B and E. This local conditional distribution is what governs how a variable is generated.
- Fourth, we define the joint distribution over all the random variables as the product of all the local conditional distributions.
- Note that we write the local conditional distributions using p, while $\mathbb{P}$ is reserved for the joint distribution over all random variables, which is defined as the product.

Probabilistic inference (alarm)

Joint distribution

b	e	a	$\mathbb{P}(B = b, E = e, A = a)$
0	0	0	$(1-\epsilon)^2$
0	0	1	0
0	1	0	0
0	1	1	$(1-\epsilon)\epsilon$
1	0	0	0
1	0	1	$\epsilon(1-\epsilon)$
1	1	0	0
1	1	1	ϵ^2

Questions:

$$\mathbb{P}(B = 1) = \epsilon(1 - \epsilon) + \epsilon^2 = \epsilon$$

$$\mathbb{P}(B = 1 \mid A = 1) = \frac{\epsilon(1-\epsilon)+\epsilon^2}{\epsilon(1-\epsilon)+\epsilon^2+(1-\epsilon)\epsilon} = \frac{1}{2-\epsilon}$$

$$\mathbb{P}(B = 1 \mid A = 1, E = 1) = \frac{\epsilon^2}{\epsilon^2+(1-\epsilon)\epsilon} = \epsilon$$

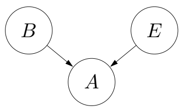
[demo]

News flash: earthquakes decrease burglarlies!*

*This is not a causal statement!

- We multiply all the local conditional distributions together to produce the joint distribution. Recall this is the probabilistic that is the source of all truth, and from it we can answer all sorts of questions.
- Let us start with the simplest query, $\mathbb{P}(B = 1)$: what is the probability of burglary without any evidence? We can sum up all the rows with $B = 1$ to get ϵ.
- Now suppose we hear the alarm $A = 1$. Let us first filter out all the rows where $A = 1$ does not hold. Then we look at the sum of the probabilities of rows where $B = 1$ over the sum of all the probabilities. The resulting probability of burglary is now $\mathbb{P}(B = 1 \mid A = 1) = \frac{1}{2-\epsilon}$.
- Now let us condition on alarm ($A = 1$) and earthquake ($E = 1$). Filter out rows that don't satisfy the condition, and look at the fraction of probabilities of remaining rows on $B = 1$. The resulting probability of burglary goes **down** to $\mathbb{P}(B = 1 \mid A = 1, E = 1) = \epsilon$ again.
- So in the end, observing that there's an earthquake does actually decrease the probability of the burglary. This might be counterintuitive because we said that burglaries and earthquakes are independent. But it's important to not interpret this causally. Creating more earthquakes clearly will not make the burglars disappear. When dealing with slippery questions such as these, we need a sound mathematical framework like Bayesian networks to ensure that we get the right answers.

Explaining away



Key idea: explaining away

Suppose two causes positively influence an effect. Conditioned on the effect, further conditioning on one cause reduces the probability of the other cause.

$$\mathbb{P}(B = 1 \mid A = 1, E = 1) < \mathbb{P}(B = 1 \mid A = 1)$$

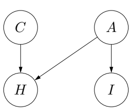
Note: happens even if causes are independent!

- This last phenomenon is so important for reasoning under uncertainty that it has a special name: **explaining away**. Suppose we have two **cause** variables B and E, which are parents of an **effect** variable A. Futher, assume the causes influence the effect positively (e.g., through the OR function).
- Let us condition on the evidence $A = 1$. We are trying to seek an explanation for $A = 1$ (what caused the alarm to go off?).
- Further conditioning on one of the causes ($E = 1$) decreases the probability of the other cause, because $E = 1$ alone **explains away** $A = 1$, and there's no more pressure on B.
- Note that in our setting, the probability of $B = 1$ returns to the original $\mathbb{P}(B = 1)$, but this need not be the case in general.
- Conditioning on $A = 1$ is important for explaining away. If you didn't, then the probability of $B = 1$ would not change. You can verify for yourself that $\mathbb{P}(B = 1 \mid E = 1) = \mathbb{P}(B = 1)$, which just follows from the definition of B and E being independent.

Medical diagnosis

Problem: cold or allergies?

You are coughing and have itchy eyes. Do you have a cold?



Random variables:

cold C, allergies A, cough H, itchy eyes I

Joint distribution:

$$\mathbb{P}(C = c, A = a, H = h, I = i) = p(c)p(a)p(h \mid c, a)p(i \mid a)$$

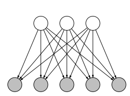
Questions:

$$\mathbb{P}(C = 1 \mid H = 1) = 0.28 \qquad \mathbb{P}(C = 1 \mid H = 1, I = 1) = 0.13$$

[demo]

- Here is another example (a cartoon version of Bayesian networks for medical diagnosis).
- Step 1: identify all the relevant variables.
- Step 2: draw arrows between them, using prior knowledge. Using our simplistic medical knowledge, suppose that a cough can be either because of a cold or because of allergies, but itchy eyes are generally only caused by allergies.
- Step 3: define a local conditional distribution for each variable.
- Step 4: multiply all the local conditional distributions to form the joint distribution.
- Now we have our probabilistic database and we can ask questions about it. Our motivating question is $\mathbb{P}(C, A \mid H = 1, I = 1)$.
- You can try the demo to get a quantitative answer. Note that $\mathbb{P}(C = 1 \mid H = 1) = 0.28$, which is another example of explaining away. Observing itchy eyes provides evidence for A, which explains away the cough ($H = 1$), resulting in a reduced probability of cold ($C = 1$).
- Note that even qualitatively reasoning about even a four-node Bayesian network can be quite subtle, let alone getting quantitative answers on large Bayesian networks. But we can rest at ease since the laws of probability make sure that all these calculations are internally consistent provided we defined the Bayesian network correctly (which in practice is an admittedly hard modeling task).

Bayesian network (definition)



Definition: Bayesian network

Let $X = (X_1, \ldots, X_n)$ be random variables.

A **Bayesian network** is a directed acyclic graph (DAG) that specifies a joint distribution over X as a product of local conditional distributions, one for each node:

$$\mathbb{P}(X_1 = x_1, \ldots, X_n = x_n) \stackrel{\text{def}}{=} \prod_{i=1}^{n} p(x_i \mid x_{\text{Parents}(i)})$$

- Without further ado, let's define a Bayesian network formally. A Bayesian network defines a joint distribution over a set of random variables.
- Second, we have a directed **acyclic** graph over the variables that captures the qualitative dependencies.
- Third, we specify a local conditional distribution for each variable X_i, which is a function that specifies a distribution over X_i given an assignment $x_{\text{Parents}(i)}$ to its parents in the graph (possibly no parents).
- Finally, the joint distribution is simply **defined** to be the product of all of the local conditional distributions.
- Notationally, we use lowercase p (in $p(x_i \mid x_{\text{Parents}(i)})$) to denote a local conditional distribution, and uppercase $\mathbb{P}$ to denote the induced joint distribution over all variables. While we will see that the two coincide, it is important to keep these things separate in your head!

Probabilistic inference (definition)

Input

Bayesian network: $\mathbb{P}(X_1, \dots, X_n)$
Evidence: $E = e$ where $E \subseteq X$ is subset of variables
Query: $Q \subseteq X$ is subset of variables

Output

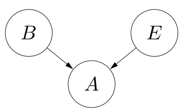
$\mathbb{P}(Q \mid E = e) \longleftrightarrow \mathbb{P}(Q = q \mid E = e)$ for all values q

Example: if coughing and itchy eyes, have a cold?

$$\mathbb{P}(C \mid H = 1, I = 1)$$

- Now given a Bayesian network representing a probabilistic database, we can answer questions on it.
- In particular, we are given a set of evidence variables E and values e. We are also given a set of query variables Q. What a probabilistic inference algorithm should output given this is the marginal conditional distribution $\mathbb{P}(Q \mid E = e)$.
- Note that this output is a table that specifies a probability for each assignment of values to Q.
- So far, we have shown examples of probabilistic inference on small Bayesian networks. The bad news is that in general, answering arbitrary probabilistic inference questions on arbitrary Bayesian networks is computationally intractable. The good news it that the core probabilistic inference in Bayesian networks is identical to Markov networks (which we will see later).

Summary



- Random variables capture state of world
- Directed edges between variables represent dependencies
- Local conditional distributions $\Rightarrow$ joint distribution
- Probabilistic inference: ask questions about world
- Captures reasoning patterns (e.g., explaining away)

- In summary, we have introduced Bayesian networks.
- It's important to think about an assignment to random variables as capturing the state of the world.
- Directed edges represent qualitative (sometimes causal) dependencies.
- Quantitatively, we specify a local conditional distribution for each variable conditioned on its parents, and multiply them together to get a joint distribution.
- Now we have our probabilistic database on which we can ask all sorts of questions, marginal conditional probabilities.
- Hopefully through the alarm and medical diagnosis examples, you are able to appreciate that the framework can capture intuitive or counter-intuitive reasoning patterns such as explaining away in a mathematically sound way so you can sleep well at night.