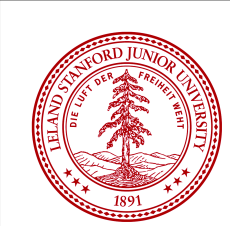
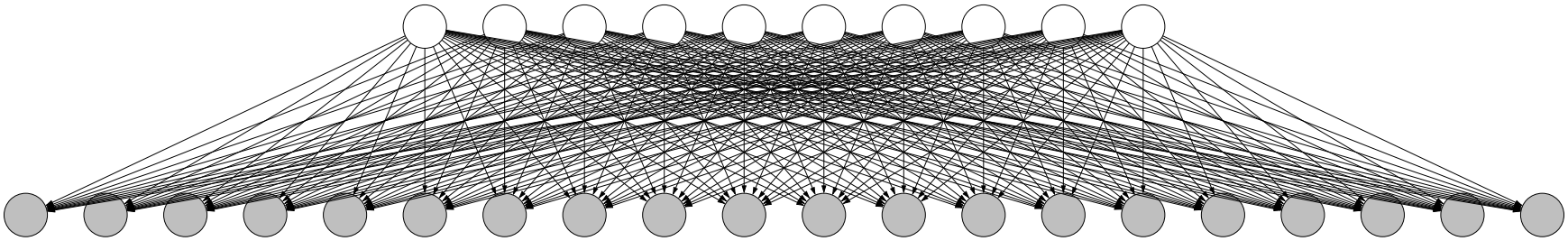
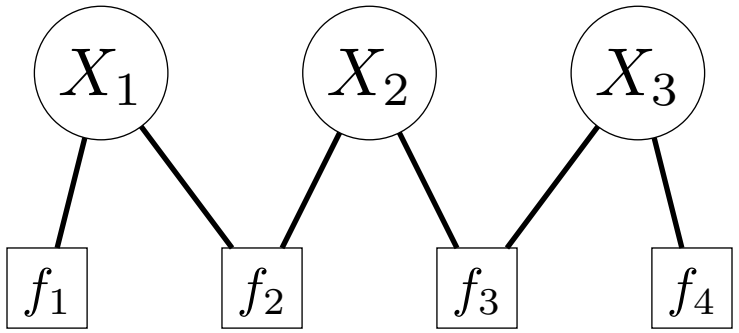
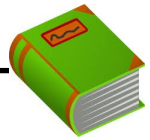


Lecture 13: Bayesian networks I



Review: definition





Definition: factor graph

Variables:

$X = (X_1, \ldots, X_n)$, where $X_i \in \text{Domain}_i$

Factors:

$f_1, \ldots, f_m$, with each $f_j(X) \geq 0$

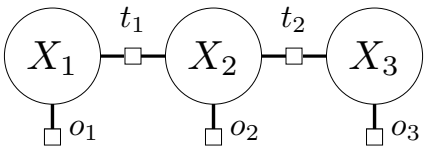
$$\text{Weight}(x) = \prod_{j=1}^{m} f_j(x)$$

- Last week, we talked about factor graphs, which uses local factors to specify a weight Weight(x) for each assignment x in a compact way. The stated objective was to find the maximum weight assignment.
- Given any factor graph, we saw a number of algorithms (backtracking, beam search, Gibbs sampling, variable elimination) for (approximately) optimizing this objective.

Review: person tracking

Problem: person tracking

Sensors reports positions: $0, 2, 2$. Objects don't move very fast and sensors are a bit noisy. What path did the person take?



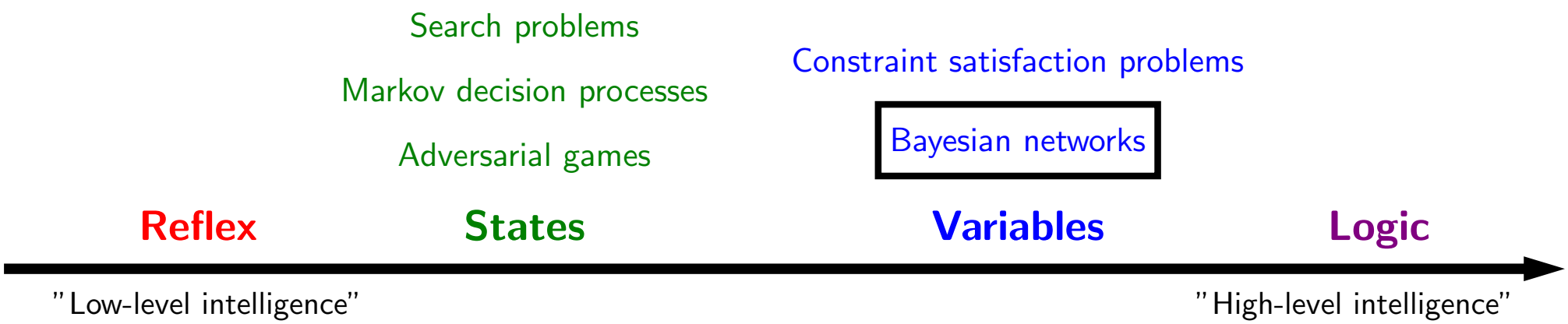
- Variables X_i: location of object at position i
- Transition factors $t_i(x_i, x_{i+1})$: incorporate physics
- Observation factors $o_i(x_i)$: incorporate sensors

[demo: `maxVariableElimination()`]

What do the factors **mean**?

- As an example, recall the object tracking example. We defined observation factors to capture the fact that the true object position is close to the sensor reading, and the transition factors to capture the fact that the true object positions across time are close to each other.
- We just set them rather ad-hocly. Is there a more principled way to think about these factors beyond being non-negative functions?

Course plan



Machine learning

- Much of this class has been on developing modeling frameworks. We started with state-based models, where we cast real-world problems as finding paths or policies through a state graph.
- Then, we saw that for a large class of problems (such as scheduling), it was much more convenient to use the language of factor graphs.
- While factor graphs could be reduced to state-based models by fixing the variable ordering, we saw that they also led to notions of treewidth and variable elimination, which allowed us to understand our models much better.
- In this lecture, we will introduce another modeling framework, Bayesian networks, which are factor graphs imbued with the language of probability. This will give probabilistic life to the factors of factor graphs.

Roadmap

Basics

Independence

Examples

- Bayesian networks were popularized in AI by Judea Pearl in the 1980s, who showed that having a coherent probabilistic framework is important for **reasoning under uncertainty**.
- There is a lot to say about the Bayesian networks (CS228 is an entire course about them and their cousins, Markov networks). So we will devote an entire lecture focusing exclusively on modeling. Any algorithms you see are a figment of your imagination (the next lecture will be for that).

Review: probability (example)

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

Joint distribution:

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

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:

$\mathbb{P}(S) =$

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

(aggregate rows)

Conditional distribution:

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

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

(select rows, normalize)

- Before introducing Bayesian networks, let's review our probability (at least the relevant parts). We start with an example about the weather. Suppose we have two boolean random variables, S and R representing sunshine and rain. 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 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 (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). The joint distribution contains all the information and acts as the central source of truth.
- From it, we can derive a **marginal distribution** over a subset of the variables. This is gotten by aggregating the rows that share the same value of S. The interpretation is that we are interested in S. We don't explicitly care about R, but we want to take into account R's effect on S. We say that R is **marginalized out**. This is a special form of elimination. In the last lecture, we leveraged max-elimination, where we took the max over the eliminated variables; here, we are taking a sum.
- The **conditional distribution** selects rows of the table matching the condition (right of the bar), and then normalizes the probabilities so that they sum to 1. The interpretation is that we observe the condition $(R = 1)$ and are interested in S. This is the conditioning that we saw for factor graphs, but where we normalize the selected rows to get probabilities.

Review: probability (general)

Random variables:

$X = (X_1, \ldots, X_n)$ partitioned into (A, B)

Joint distribution:

$$\mathbb{P}(X) = \mathbb{P}(X_1, \ldots, X_n)$$

Marginal distribution:

$$\mathbb{P}(A) = \sum_b \mathbb{P}(A, B = b)$$

Conditional distribution:

$$\mathbb{P}(A \mid B = b) = \frac{\mathbb{P}(A, B=b)}{\mathbb{P}(B=b)}$$

- In general, we have n random variables $X_1, \ldots, X_n$ and let X denote all of them. Suppose X is partitioned into A and B (e.g., $A = (X_1, X_3)$ and $B = (X_2, X_4, X_5)$ if $n = 5$).
- The marginal and conditional distributions can be defined over the subsets A and B rather than just single variables.
- Of course, we can also have a hybrid too: for $n = 3$, $\mathbb{P}(X_1 \mid X_3 = 1)$ marginalizes out X_2 and conditions on $X_3 = 1$.
- It is important to remember the types of objects here: $\mathbb{P}(A)$ is a table where rows are possible assignments to A, whereas $\mathbb{P}(A = a)$ is a number representing the probability of the row corresponding to assignment a.

Probabilistic inference

Random variables: unknown quantities in the world

$$X = (S, R, T, A)$$

In words:

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

In symbols:

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

(S is **marginalized out**)

- At this point, you should have all the definitions to compute any marginal or conditional distribution given access to a joint probability distribution. But what is this really doing and how is this useful?
- We should think about each assignment x as a possible state of the world (it's raining, it's not sunny, there is traffic, it is autumn, etc.). Think of the joint distribution as one giant database that contains full information about how the world works.
- In practice, we'd like to ask questions by querying this probabilistic database. First, we observe some evidence, which effectively fixes some of the variables. Second, we are interested in the distribution of some set of variables which we didn't observe. This forms a query, and the process of answering this query (computing the desired distribution) is called **probabilistic inference**.

Challenges

Modeling: How to specify a joint distribution $\mathbb{P}(X_1, \ldots, X_n)$ **compactly**?

Bayesian networks (factor graphs for probability distributions)

Algorithms: How to compute queries $\mathbb{P}(R \mid T = 1, A = 1)$ **efficiently**?

Variable elimination, Gibbs sampling, particle filtering (analogue of algorithms for finding maximum weight assignment)

- In general, a joint distribution over n variables has size exponential in n. From a modeling perspective, how do we even specify an object that large? Here, we will see that Bayesian networks, based on factor graphs, offer an elegant solution.
- From an algorithms perspective, there is still the question of how we perform probabilistic inference efficiently. In the next lecture, we will see how we can adapt all of the algorithms that we saw before for computing maximum weight assignments in factor graphs, essentially by replacing a max with a sum.
- The two desiderata are rather synergistic, and it is the same property — conditional independence — that makes both possible.

cs221.stanford.edu/q

Question

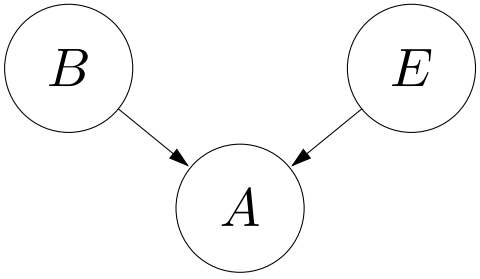
Earthquakes and burglaries are independent events that will cause an alarm to go off. Suppose you hear an alarm. How does hearing on the radio that there's an earthquake change your beliefs?

- it increases the probability of burglary
- it decreases the probability of burglary
- it does not change the probability of burglary

- Situations like these arise all the time in practice: we have a lot of unknowns which are all dependent on one another. If we obtain evidence on some of these unknowns, how does that affect our belief about the other unknowns?

Bayesian network (alarm)

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



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

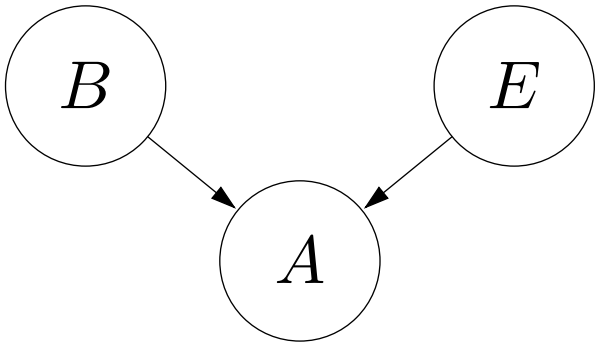
$$p(b) = \epsilon \cdot [b = 1] + (1 - \epsilon) \cdot [b = 0]$$

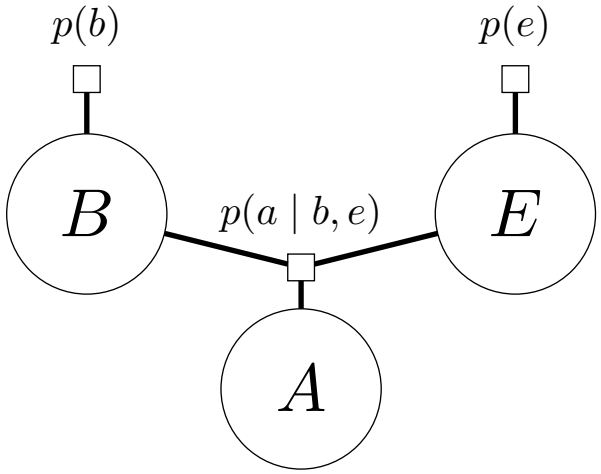
$$p(e) = \epsilon \cdot [e = 1] + (1 - \epsilon) \cdot [e = 0]$$

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

- Let us try to model the situation. First, we establish that there are three variables, B (burglary), E (earthquake), and A (alarm). Next, 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 suggesting causality, though what this actually means is a deeper question and beyond the scope of this class.
- 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.
- We are writing 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.

Bayesian network (alarm)





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

Bayesian networks are a special case of factor graphs!

- Note that the local conditional distributions (e.g., $p(a \mid b, e)$) are non-negative so they can be thought of simply as factors of a factor graph. The joint probability of an assignment is then the weight of that assignment.
- In this light, Bayesian networks are just a type of factor graphs, but with additional structure and interpretation.

Probabilistic inference (alarm)

Joint distribution:

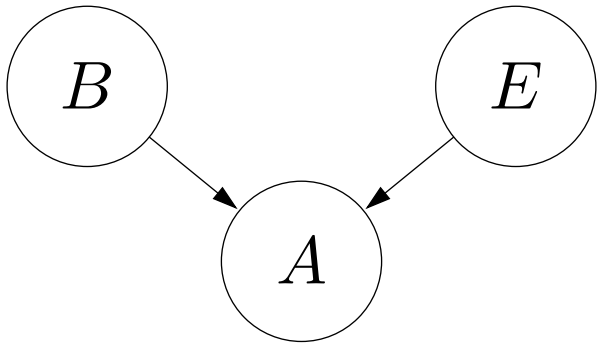
b	e	a	$\mathbb{P}(B=b, E=e, A=a)$
0	0	0	$(1-\epsilon)^2$
0	0	1	0
0	1	0	0
0	1	1	$(1-\epsilon)\epsilon$
1	0	0	0
1	0	1	$\epsilon(1-\epsilon)$
1	1	0	0
1	1	1	ϵ^2

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

[demo: $\epsilon = 0.05$]

- Bayesian networks can be used to capture common reasoning patterns under uncertainty (which was one of their first applications).
- Consider the following model: Suppose the probability of an earthquake is ϵ and the probability of a burglary is ϵ and both are independent. Suppose that the alarm always goes off if either an earthquake or a burglary occurs.
- In the prior, we can eliminate A and E and get that the probability of the burglary is ϵ.
- Now suppose we hear the alarm $A = 1$. The probability of burglary is now $\mathbb{P}(B = 1 \mid A = 1) = \frac{1}{2-\epsilon}$.
- Now suppose that you hear on the radio that there was an earthquake ($E = 1$). Then the probability of burglary goes down to $\mathbb{P}(B = 1 \mid A = 1, E = 1) = \epsilon$ again.

Explaining away

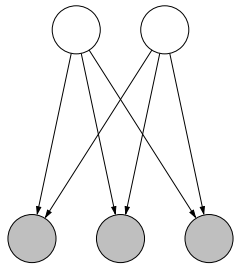


Key idea: explaining away

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

- This last phenomenon has a special name: **explaining away**. Suppose we have two **cause** variables B and E, which are parents of an **effect** variable A. Assume the causes influence the effect positively (e.g., through the OR function).
- Conditioned on the effect $A = 1$, there is some posterior probability of B. Conditioned on the effect $A = 1$ and the other cause $E = 1$, the new posterior probability is reduced. We then say that the other cause E has explained away B.

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 large joint distribution in a modular way, one variable at a time.
- First, the graph structure captures what other variables a given variable depends on.
- Second, we specify a local conditional distribution for 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 none). The joint distribution is simply **defined** to be the product of all of the local conditional distributions together.
- 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 the two can coincide, it is important to keep these things separate in your head!
- While formally, a Bayesian network just defines a probability distribution, it can be intuitive (although a bit dangerous) to think of the graph as capturing notions of **causality**. This is made more precise with causal networks, but this is beyond the scope of this class.

Special properties

Key difference from general factor graphs:

Key idea: locally normalized

All factors (local conditional distributions) satisfy:

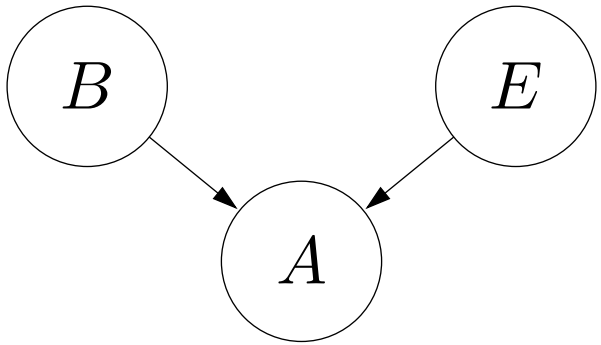
$$\sum_{x_i} p(x_i \mid x_{\text{Parents}(i)}) = 1 \text{ for each } x_{\text{Parents}(i)}$$

Implications:

- Consistency of sub-Bayesian networks
- Consistency of conditional distributions

- But Bayesian networks are more than that. The key property is that all the local conditional distributions, being distributions, sum to 1 over the first argument.
- This simple property results in two important properties of Bayesian networks that are not present in general factor graphs.

Consistency of sub-Bayesian networks



A short calculation:

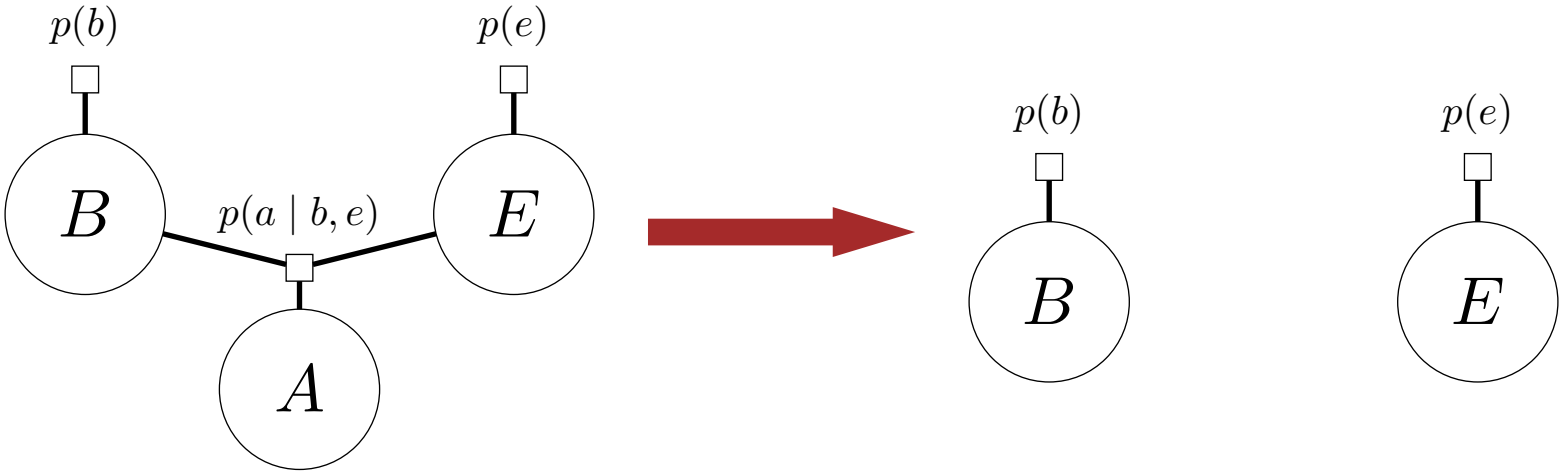
$$\begin{aligned}
\mathbb{P}(B = b, E = e) &= \textstyle\sum_a \mathbb{P}(B = b, E = e, A = a) \\
&= \textstyle\sum_a p(b)p(e)p(a \mid b, e) \\
&= p(b)p(e) \textstyle\sum_a p(a \mid b, e) \\
&= p(b)p(e)
\end{aligned}$$

- First, let's see what happens when we marginalize A (by performing algebra on the joint probability). We see that we end up with $p(b)p(e)$, which actually defines a sub-Bayesian network with one fewer variable, and the same local conditional probabilities.
- If one marginalizes out all the variables, then one gets 1, which verifies that a Bayesian network actually defines a probability distribution.
- The philosophical ramification of this property is that there could be many other variables that depend on the variables you've modeled (earthquakes also impacts traffic) but as long as you don't observe them, they can be ignored mathematically (ignorance is bliss). Note that this doesn't mean that knowing about the other things isn't useful.

Consistency of sub-Bayesian networks

Key idea: marginalization

Marginalization of a leaf node yields a Bayesian network without the node.

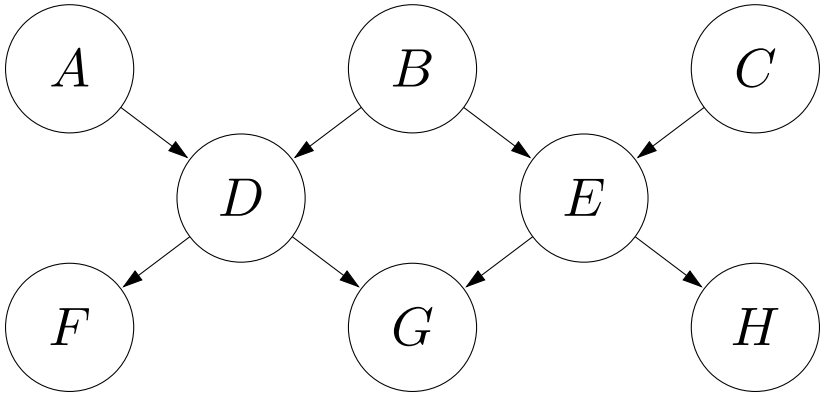


- This property is very attractive, because it means that whenever we have a large Bayesian network, where we don't care about some of the variables, we can just remove them (graph operations), and this encodes the same distribution as we would have gotten from marginalizing out variables (algebraic operations). The former, being visual, can be more intuitive.

Consistency of local conditionals

Key idea: local conditional distributions

Local conditional distributions (factors) are the true conditional distributions.



$$\underbrace{\mathbb{P}(D = d \mid A = a, B = b)}_{\text{from probabilistic inference}} = \underbrace{p(d \mid a, b)}_{\text{by definition}}$$

- Note that the local conditional distributions $p(d \mid a, b)$ are simply defined by the user. On the other hand, the quantity $\mathbb{P}(D = d \mid A = a, B = b)$ is not defined, but follows from probabilistic inference on the joint distribution defined by the Bayesian network.
- It's not clear a priori that the two have anything to do with each other. The second special property that we get from using Bayesian networks is that the two are actually the same.
- To show this, we can remove all the descendants of D by the consistency of sub-Bayesian networks, leaving us with the Bayesian network $\mathbb{P}(A = a, B = b, D = d) = p(a)p(b)p(d \mid a, b)$. By the chain rule, $\mathbb{P}(A = a, B = b, D = d) = \mathbb{P}(A = a, B = b)\mathbb{P}(D = d \mid A = a, B = b)$. If we marginalize out D, then we are left with the Bayesian network $\mathbb{P}(A = a, B = b) = p(a)p(b)$. From this, we can conclude that $\mathbb{P}(D = d \mid A = a, B = b) = p(d \mid a, b)$.
- This argument generalizes to any Bayesian network and local conditional distribution.

Medical diagnosis

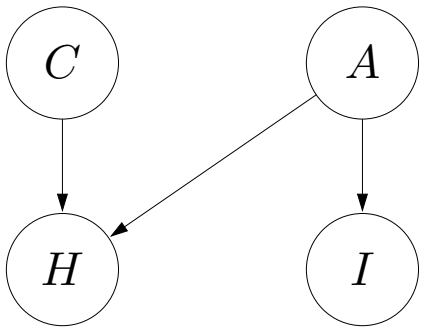
Problem: cold or allergies?

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

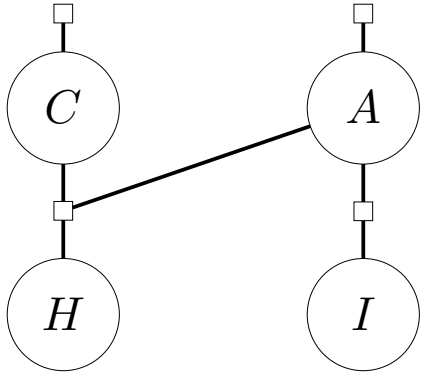
[demo]

Variables: **C**old, **A**llergies, Coug**h**, **I**tchy eyes

Bayesian network:



Factor graph:



- Here is another example (a cartoon version of Bayesian networks for medical diagnosis). Allergies and cold are the two hidden variables that we'd like to infer (we have some prior over these two). Cough and itchy eyes are symptoms that we observe as evidence, and we have some likelihood model of these symptoms given the hidden causes.
- We can use the demo to infer the hidden state given the evidence.

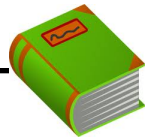
Bayes rule

Variables: **hidden** H, **evidence** E

True statements:

$$\mathbb{P}(H = h, E = e) = \mathbb{P}(H = h, E = e)$$

$$\mathbb{P}(H = h \mid E = e)\,\mathbb{P}(E = e) = \mathbb{P}(H = h)\,\mathbb{P}(E = e \mid H = h)$$



Definition: Bayes rule

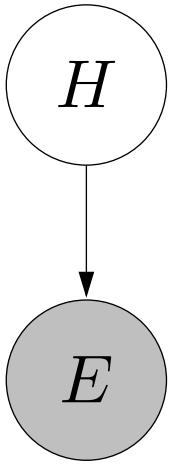
$$\underbrace{\mathbb{P}(H = h \mid E = e)}_{\text{posterior}} \propto \underbrace{\mathbb{P}(H = h)}_{\text{prior}} \underbrace{\mathbb{P}(E = e \mid H = h)}_{\text{likelihood}}$$

Significance: interpretation

- **Bayes rule** is an extremely important concept in probability, which follows essentially from two applications of the definition of conditional distributions.
- We assume a hidden variable H (e.g., true location) and an evidence variable E (e.g., sensor reading). We assume that there is a **prior** distribution over H, which is what one believes about H before having seen E. The **likelihood** is the probability of seeing a particular piece of evidence e given the hidden variable. The **posterior** is the distribution over the hidden variable given the evidence.
- Bayes rule allows us to "invert" the conditioning: if likelihood goes forwards, then posterior looks backwards.
- Notation: we will often say that a probability distribution $p(x)$ is proportional to another function $f(x)$ (written $p(x) \propto f(x)$) if there exists a normalization constant Z independent of x such that $p(x) = \frac{f(x)}{Z}$ for all x. Another way to think about it is that if we normalized $f(x)$, we would get $p(x)$.

Bayes rule

Bayesian network:



$$\mathbb{P}(H = h, E = e) = p(h)p(e \mid h)$$

$$\underbrace{\mathbb{P}(H = h \mid E = e)}_{\text{posterior (want)}} \propto \underbrace{p(h)}_{\text{prior (have)}} \quad \underbrace{p(e \mid h)}_{\text{likelihood (have)}}$$

- For a Bayesian network, Bayes rule is especially interesting because the prior and likelihood are local conditional distributions, which are specified by the user. The posterior is typically the query that one wants to answer.
- The mental picture is one of playing detective: we see some evidence, and want to figure out what happened.

Roadmap

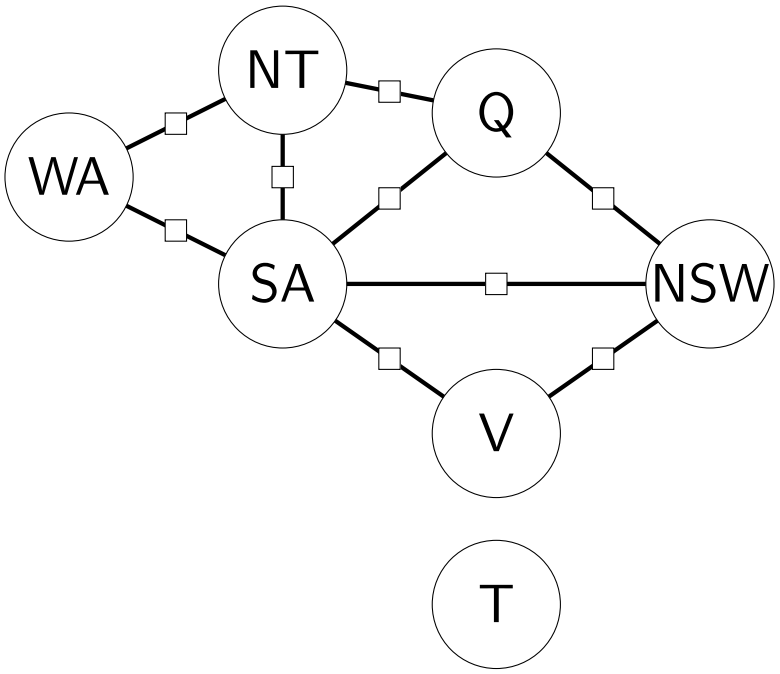
Basics

Independence

Examples

- Recall that independence (or more generally, conditional independence) is the key property that made things efficient for finding maximum weight assignments.
- But with Bayesian networks, things get a bit more complex and subtle...

Factor graph independence



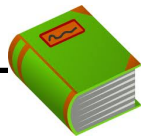
Definition: factor graph independence

Let A and B be two subsets of X.

We have $A \perp\!\!\!\perp B$ iff there is no path from A and B.

- Recall that we have defined independence purely in terms of the factor graph structure: let us call this **factor graph independence** for clarity.

Probabilistic independence



Definition: probabilistic independence

Let A and B be two (subsets of) random variables.

We have $A \perp\!\!\!\perp B$ iff

$$\mathbb{P}(A = a, B = b) = \mathbb{P}(A = a)\mathbb{P}(B = b) \text{ for all } a, b.$$

Example: probabilistic independence

x_1, x_2	$\mathbb{P}(X_1 = x_1, X_2 = x_2)$
0,0	0.06
0,1	0.04
1,0	0.54
1,1	0.36

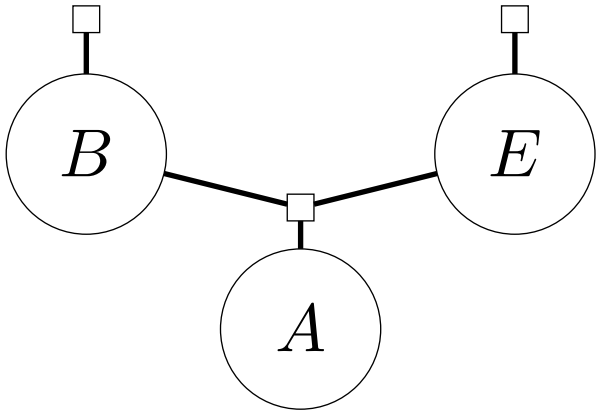
x_1	$\mathbb{P}(X_1 = x_1)$
0	0.1
1	0.9

x_2	$\mathbb{P}(X_2 = x_2)$
0	0.6
1	0.4

- Separately, there is the usual definition of independence from probability theory, which says that two (sets of) random variables A and B are independent iff the joint distribution is equal to the product of the marginals.

Tale of two independences

A Bayesian network is a factor graph **and** specifies a probability distribution...



$$\mathbb{P}(B = b, E = e, A = a)$$

Question: Are factor graph independence and probabilistic independence the same?

- A Bayesian network has its foot in both camps: it is equipped with a graph structure which can be used to define factor graph independence, but it also defines a joint distribution, which can be used to define probabilistic independence.
- Are these the same, and if not, what are the relationships between the two?

Independence

Factor graph independence (easy to check):

No path between A and B.

Probabilistic independence (semantically important):

$\mathbb{P}(A = a, B = b) = \mathbb{P}(A = a)\mathbb{P}(B = b)$ for all a, b.

Proposition: independence

Factor graph independence $\Rightarrow$ probabilistic independence.

- The first observation is that factor graph independence implies probabilistic independence. This should be clear intuitively: if two variables are not connected graphically, they have nothing to do with each other and thus should be independent.

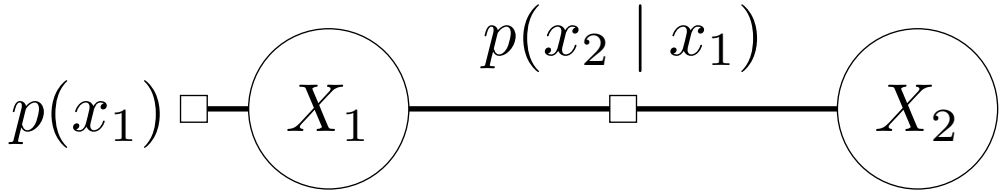
cs221.stanford.edu/q

Question

Does probabilistic independence imply factor graph independence?

yes
no
don't know

Independence revisited



Do not have factor graph independence.

Suppose:

x_1	x_2	$\mathbb{P}(X_1 = x_1, X_2 = x_2)$
0	0	0.25
0	1	0.25
1	0	0.25
1	1	0.25

$$p(x_2 \mid x_1) = 0.5$$

Still have X_1 and X_2 probabilistically independent...

Conclusion:

probabilistic independence $\not\Rightarrow$ factor graph independence

- However, the converse is not true. As a trivial example, one can have factors that depend on lots of variables but which always return 1. Two variables connected via this factor would not be factor graph independent. On the other hand, this factor isn't doing anything, and therefore the variables are still probabilistically independent.

Probabilities

Key idea: families of probability distributions

Each **factor graph structure** defines a family of different probability distributions:

$$\mathbb{P}(X = x) \propto \text{Weight}(x) = \prod_{j=1}^{m} f_j(x)$$

Adding edges grows the family of distributions:

$$\text{Distributions}\left(\; X_1 \quad X_2 \; \right) \subseteq \text{Distributions}\left(\; X_1 - X_2 \; \right)$$

In both:

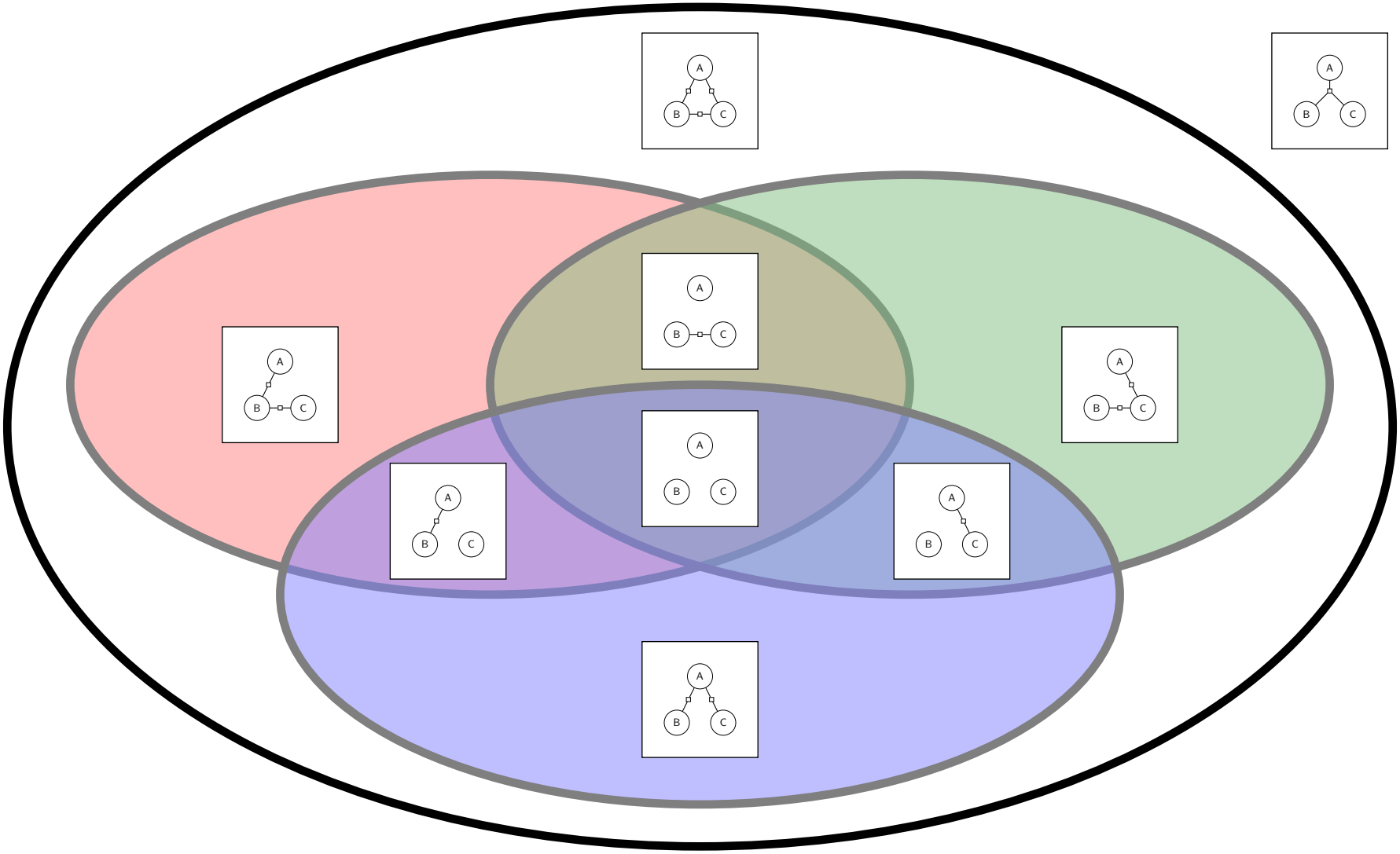
x_1, x_2	$\mathbb{P}(X_1 = x_1, X_2 = x_2)$
0,0	0.25
0,1	0.25
1,0	0.25
1,1	0.25

Only in more complex one:

x_1, x_2	$\mathbb{P}(X_1 = x_1, X_2 = x_2)$
0,0	0.5
0,1	0.0
1,0	0.0
1,1	0.5

- It turns out the right way to think about all this is that structural independence is really a property of **factor graph structures** (without the actual factor functions).
- A factor graph structure defines a family of probability distributions as we range over the actual factor functions.
- Adding new edges (factors) increases the set of distributions included in the family. For example, the uniform distribution is in both of the two factor graph structures over two nodes, but the distribution that forces $X_1 = X_2$ is only representable by the more complex factor graph.

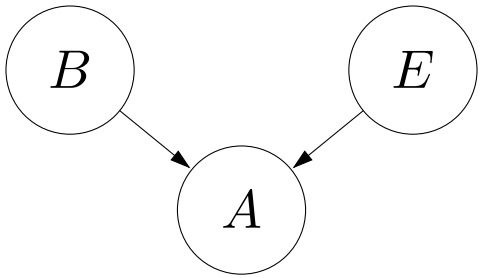
Families of distributions



green prohibits A — B red prohibits A — C blue prohibits B — C

- More generally, we can use a Venn diagram to represents the families of distributions represented by different factor graphs.
- Here, the red ellipse contains distributions where there is no factor between A and C; the green contains the ones where there is none between A and B; and the blue one prohibits B and C. The big ellipse contains distributions which only have binary factors.
- The smallest family covers exactly the distributions where all the variables are independent (structurally and probabilistically).
- The regions immediately around it and including the center region are distributions defined by a factor graph structure with one binary factor.
- Each of the three ellipses corresponds to structures with two binary factors.
- The big white region corresponds to structures with three binary factors.
- Finally, the set of all distributions (over three variables) is represented by the structure with a ternary factor.

Bayesian network independence

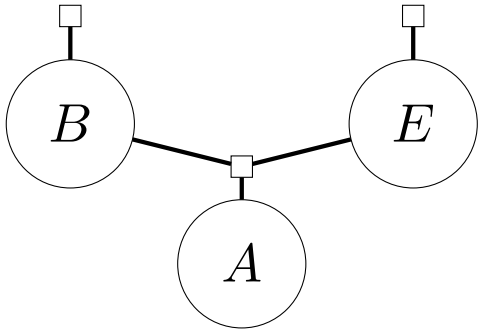


Are B and E independent?

- But there's more with Bayesian networks, which have additional structure beyond what's offered by a factor graph.

Bayesian network independence

Factor graph independence: no



Probabilistic independence: yes

(marginalize out A)

$$\mathbb{P}(B = b, E = e) = \mathbb{P}(B = b)\mathbb{P}(E = e)$$

- Based on the factor graph, we would say that B and E are not necessarily independent. However, if we use the consistency of sub-Bayesian networks, we can marginalize out A and are left with a smaller Bayesian network capturing the original marginal $\mathbb{P}(B = b, E = e)$. In this Bayesian network, it's clear that factor graph independence holds.

Bayesian network independence

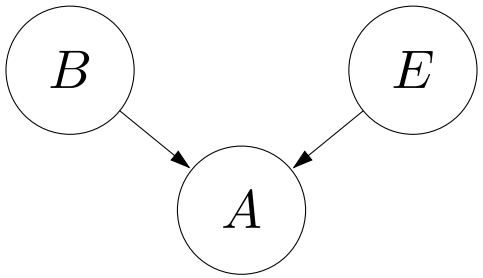
Bayesian networks have even more independences than specified by its factor graph...

Definition: Bayesian network independence

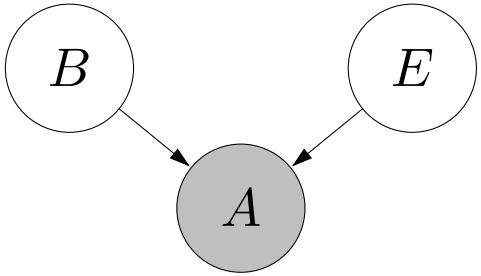
Given a Bayesian network structure, we say $A \perp\!\!\!\perp B$ if after marginalizing all descendants of A and B, the resulting factor graph has $A \perp\!\!\!\perp B$.

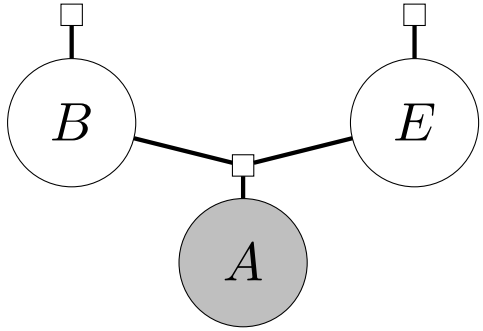
- Using the factor graph is not wrong: it's just overly conservative. To be more precise, we should take a Bayesian network of interest and marginalize out as many of the irrelevant variables as possible, which includes all descendants of A and B. Then we can appeal to factor graph independence.

Pattern 1: v-structure



Parents B and E are **conditionally dependent** (condition on A):

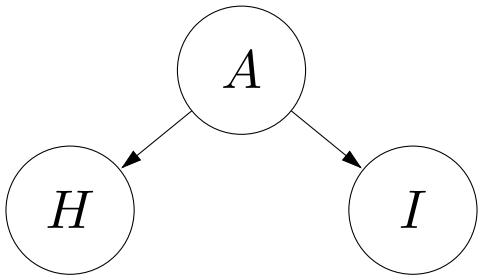




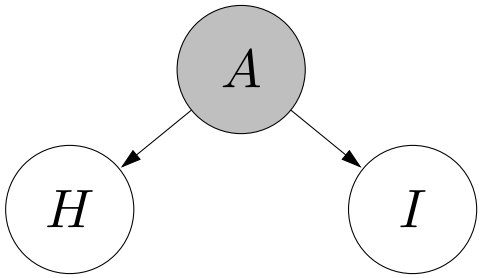
Parents B and E are **independent** (marginalize out A):

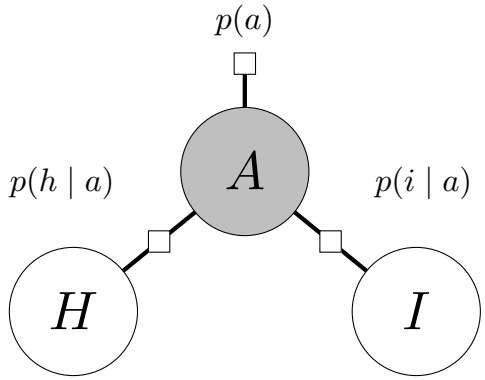
- For factor graphs, independence and conditional independence were straightforward graph properties. Two (sets of) variables B and E were conditionally independent given A if there was no path from B to E in the graph.
- Conditional independence in Bayesian networks warrant some discussion. First, suppose we condition on A. Are B and E independent? The answer is no. A common mistake is to look at the DAG and conclude that A separates B and E. It is important to look at the factor graph rather than the DAG. Recall that the factor graph corresponding to a Bayesian network includes a single factor that depends on both A and its two parents B and E. When we condition on A, that only removes A from the picture, but B and E still have a factor in common. This is intuitive: recall that given $A = 1$, E explained away B.
- Now let's not condition on A but ask whether B and E are independent? Here, we have to eliminate A. But remember that for Bayesian networks, eliminating variables that are not ancestors of the query variables simply amounts to removing them: the new factor produced is $f(b, e) = \sum_a p(a \mid b, e) = 1$, which can be safely ignored. As a result, B and E are actually independent. This is a special property of Bayesian networks: **eliminating children renders parents independent**. This is also intuitive if we think about the interpretation of Bayesian networks: Without any information (such as A), the parents B and E have nothing to do with each other (and thus are independent).
- Note: this three-node Bayesian network is called a **v-structure**. If two variables A, B are Bayesian network conditionally independent given C, then they are said to be **d-separated**.

Pattern 2: inverted v-structure

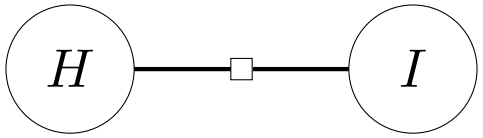


Children H and I are **conditionally independent** (condition on A):



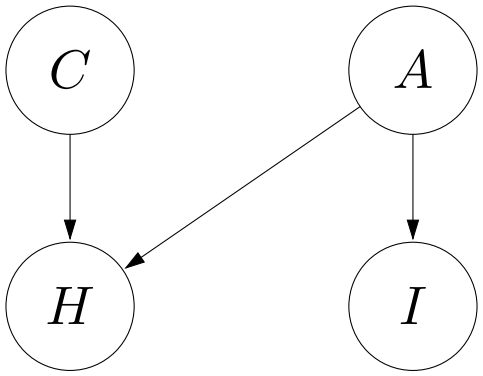


Children H and I are **dependent** (marginalize out A):



- The other structure to keep in mind is the inverted v-structure. Do not get these two confused! While the two structures look graphically similar, if you pause a moment to think about the semantics of Bayesian networks, you'll realize that the two are quite different. First, rather than having one ternary factor $p(a \mid b, e)$ and two unary factors, we now have two binary factors, $p(h \mid a)$ and $p(i \mid a)$, and one unary factor.
- In fact, the inverted v-structure is nothing special: all the independences follow from its factor graph structure.

Independence: examples



$C \perp\!\!\!\perp A$? yes

$C \perp\!\!\!\perp I$? yes

$C \perp\!\!\!\perp H$? no

$A \perp\!\!\!\perp I$? no

$A \perp\!\!\!\perp H$? no

$I \perp\!\!\!\perp H$? no

$C \perp\!\!\!\perp A \mid H$? no

$C \perp\!\!\!\perp I \mid H$? no

$A \perp\!\!\!\perp I \mid H$? no

$C \perp\!\!\!\perp H \mid A$? no

$C \perp\!\!\!\perp I \mid A$? yes

$I \perp\!\!\!\perp H \mid A$? yes

- Now let us look at the medical diagosis example from earlier, which combines both patterns.
- Which of these are Bayesian network independent? Go through each of these and convince yourself of the answer.
- First, marginalize out all the variables that do not have any descendants that we're conditioning on. This just means dropping those nodes by consistency properties.
- Next, we can either check the two patterns directly for a small enough factor graph, or else we convert the Bayesian network to a factor graph, where independence checking is easy (graph connectivity).

Summary of independences

Factor graph independence (on structures)

Bayesian network independence (on structures)

Probabilistic independence (on distributions)

- In summary, factor graph independence is the most conservative and the easiest to verify.
- If we have Bayesian networks, then we can marginalize out irrelevant variables first; this is essential to get tight results.
- Finally, the ultimate goal is probabilistic independence, which can be quite complex. We will always have independences which are not captured by the graph structure.

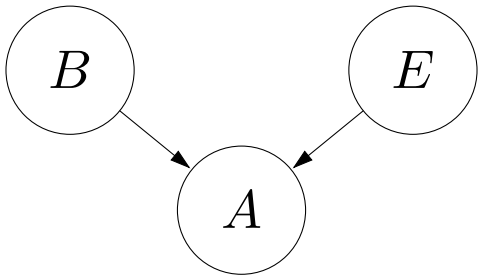
Roadmap

Basics

Independence

Examples

Probabilistic programs



Probabilistic program: alarm

$$B \sim \text{Bernoulli}(\epsilon)$$
$$E \sim \text{Bernoulli}(\epsilon)$$
$$A = B \vee E$$

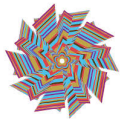
Key idea: probabilistic program

A randomized program that sets the random variables.

```
def Bernoulli(epsilon):  return random.random() < epsilon
```

- There is another way of writing down Bayesian networks other than graphically or mathematically, and that is as a probabilistic program. A **probabilistic program** is a randomized program that invokes a random number generator to make random choices. Executing this program will assign values to a collection of random variables $X_1, \ldots, X_n$; that is, generating an assignment.
- The probability (e.g., fraction of times) that the program generates that assignment is exactly the probability under the joint distribution specified by that program.
- We should think of this program as outputting the state of the world (or at least the part of the world that we care about for our task).
- Note that the probabilistic program is only used to define joint distributions. We usually wouldn't actually run this program directly.
- For example, we show the probabilistic program for alarm. $B \sim$ Bernoulli(ϵ) simply means that $\mathbb{P}(B = 1) = \epsilon$. Here, we can think about Bernoulli(ϵ) as a randomized function (`random() < epsilon`) that returns 1 with probability ϵ and 0 with probability $1 - \epsilon$.

Probabilistic program: example



Probabilistic program: object tracking

$X_0 = (0, 0)$

For each time step $i = 1, \ldots, n$:

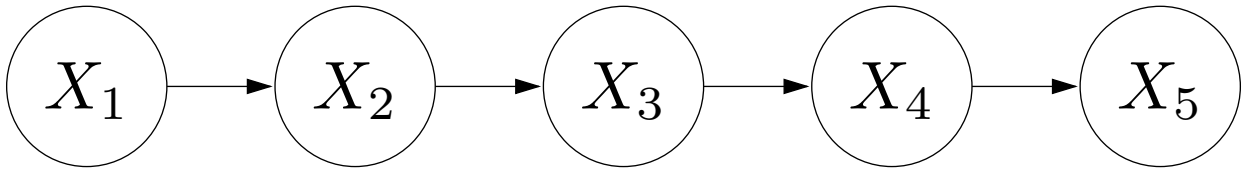
With probability α:

$X_i = X_{i-1} + (1, 0)$ [go right]

With probability $1 - \alpha$:

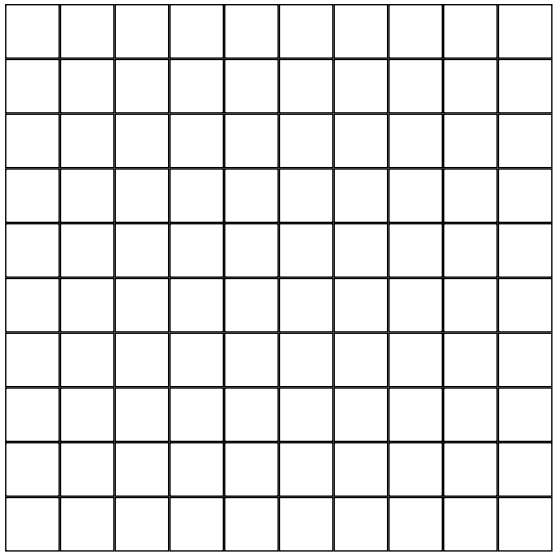
$X_i = X_{i-1} + (0, 1)$ [go down]

Bayesian network:



- This is a more interesting generative model since it has a for loop, which allows us to determine the distribution over a templatized set of n variables rather than just 3 or 4.
- In these cases, variables are generally indexed by something like time or location.
- We can also draw the Bayesian network. Each X_i only depends on X_{i-1}. This is a chain-structured Bayesian network, called a **Markov model**.

Probabilistic program: example



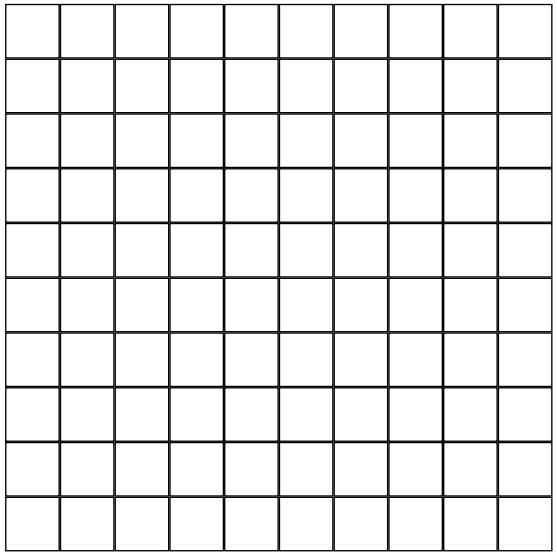
(press ctrl-enter to save)

Run

- Try clicking [Run]. Each time a new assignment of $(X_1, \ldots, X_n)$ is chosen.

Probabilistic inference: example

Query: what are possible trajectories given **evidence** $X_{10} = (8, 2)$?



(press ctrl-enter to save)

Run

- This program only serves for defining the distribution. Now we can query that distribution and ask the question: suppose the program set $X_{10} = (8, 2)$; what is the distribution over the other variables?
- In the demo, note that all trajectories are constrained to go through $(8, 2)$ at time step 10.

Probability distribution: example

Local conditional distribution:

$$p(x_i \mid x_{i-1}) = \alpha \cdot \underbrace{[x_i = x_{i-1} + (1,0)]}_{\text{right}} + (1-\alpha) \cdot \underbrace{[x_i = x_{i-1} + (0,1)]}_{\text{down}}$$

Joint distribution:

$$\mathbb{P}(X = x) = \prod_{i=1}^{n} p(x_i \mid x_{i-1})$$

Example: object tracking

Assignment:

$$x_1 = (1,0), x_2 = (1,1), x_3 = (2,1)$$

Joint distribution:

$$\mathbb{P}(X = x) = \alpha \cdot (1-\alpha) \cdot \alpha$$

- Let's ground out the probabilistic program in symbols.
- First, we formalize the (conditional) distribution over each variable X_i. In this example, this variable depends on only the previous position X_{i-1}, but in general, X_i could depend on any of the previous variables.
- The joint distribution over all the variables is simply the product of the conditional distribution over each of the variables.

Three equivalent representations

Probabilistic program:

Probabilistic program: object tracking

$X_0 = (0, 0)$

For each time step $i = 1, \ldots, n$:

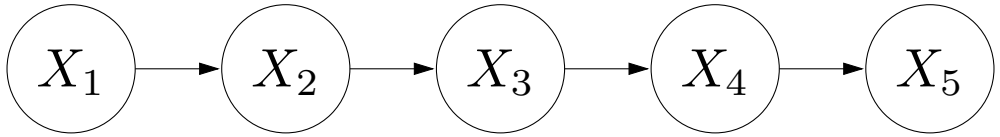
  With probability α:

    $X_i = X_{i-1} + (1, 0)$ [go right]

  With probability $1 - \alpha$:

    $X_i = X_{i-1} + (0, 1)$ [go down]

Bayesian network:



Mathematical definition:

$$p(x_i \mid x_{i-1}) = \alpha \cdot \underbrace{[x_i = x_{i-1} + (1, 0)]}_{\text{right}} + (1 - \alpha) \cdot \underbrace{[x_i = x_{i-1} + (0, 1)]}_{\text{down}}$$

- To summarize, we have seen three equivalent representations for the object tracking model: the probabilistic program, the graphical Bayesian network, and the mathematical definition. It's useful to be able to quickly switch back and forth between the different ones.

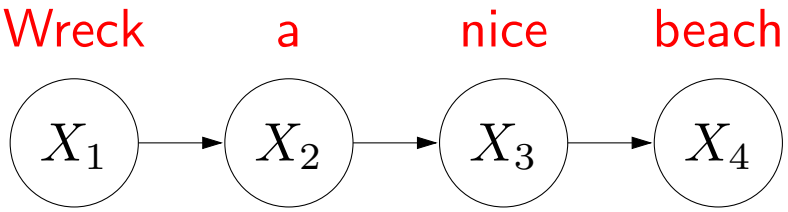
Application: language modeling

Question: what did I just say?

Probabilistic program: Markov model

For each position $i = 1, 2, \ldots, n$:

Generate word $X_i \sim p(X_i \mid X_{i-1})$

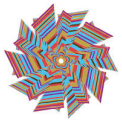


Applications:

- Speech recognition
- Machine translation

- In the context of natural language, a Markov model is known as a bigram model. A higher-order generalization of bigram models are n-gram models (more generally known as higher-order Markov models).
- Language models are often used to measure the "goodness" of a sentence, mostly within the context of a larger system such as speech recognition or machine translation.

Application: object tracking

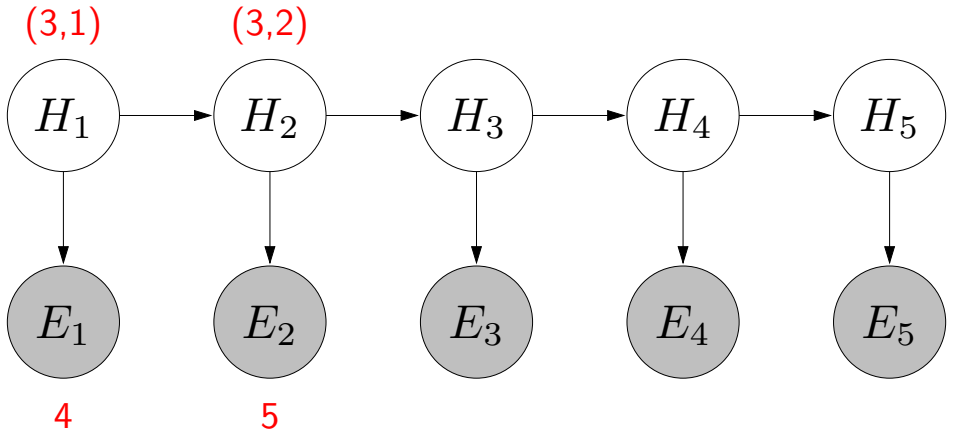


Probabilistic program: hidden Markov model (HMM)

For each time step $t = 1, \ldots, T$:

Generate object location $H_t \sim p(H_t \mid H_{t-1})$

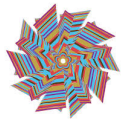
Generate sensor reading $E_t \sim p(E_t \mid H_t)$



Other applications: speech recognition

- Markov models are limiting because they do not have a way of talking about noisy evidence (sensor readings). They can be extended quite easily to hidden Markov models, which introduce a parallel sequence of observation variables.
- For example, in object tracking, H_t denotes the true object location, and E_t denotes the noisy sensor reading, which might be (i) the location H_t plus noise, or (ii) the distance from H_t plus noise, depending on the type of sensor.
- In speech recognition, H_t would be the phonemes or words and E_t would be the raw acoustic signal.

Application: multiple object tracking



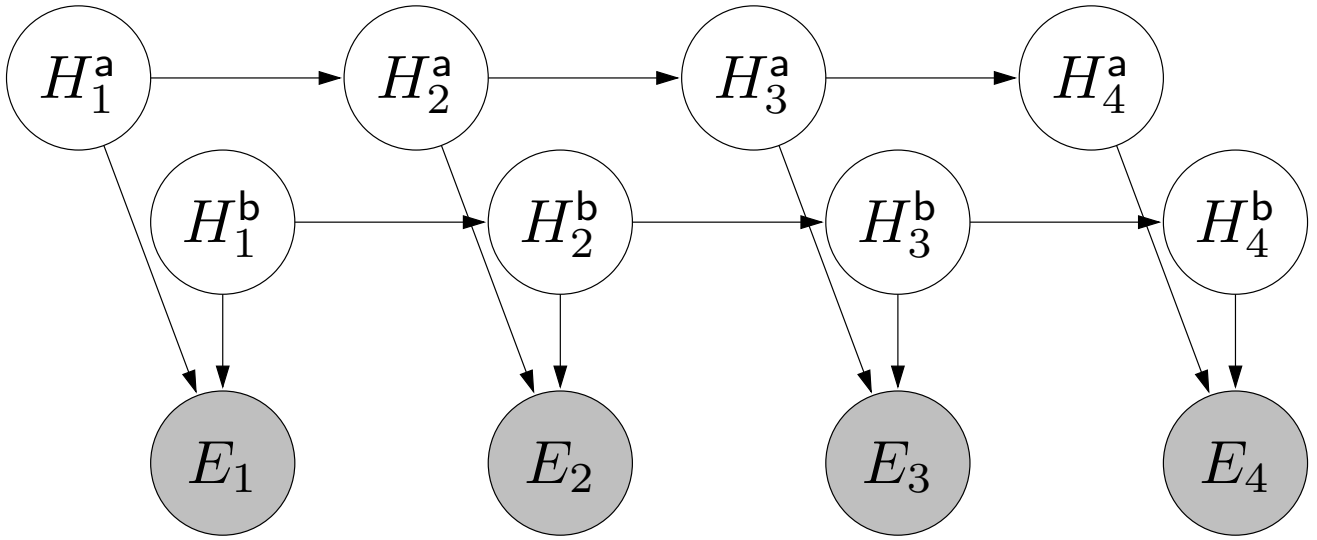
Probabilistic program: factorial HMM

For each time step $t = 1, \ldots, T$:

  For each object $o \in \{\mathsf{a}, \mathsf{b}\}$:

    Generate location $H_t^o \sim p(H_t^o \mid H_{t-1}^o)$

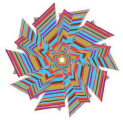
  Generate sensor reading $E_t \sim p(E_t \mid H_t^{\mathsf{a}}, H_t^{\mathsf{b}})$



- An extension of an HMM, called a **factorial HMM**, can be used to track multiple objects. We assume that each object moves independently according to a Markov model, but that we get one sensor reading which is some noisy aggregated function of the true positions.
- For example, E_t could be the set $\{H_t^{\mathsf{a}}, H_t^{\mathsf{b}}\}$, which reveals where the objects are, but doesn't say which object is responsible for which element in the set.

Application: document classification

Question: given a text document, what is it about?

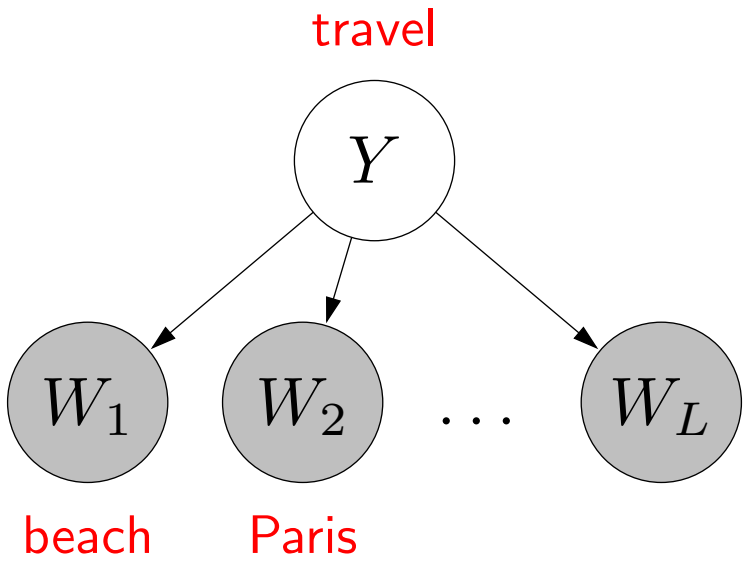


Probabilistic program: naive Bayes

Generate label $Y \sim p(Y)$

For each position $i = 1, \ldots, L$:

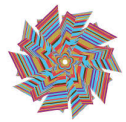
Generate word $W_i \sim p(W_i \mid Y)$



- Naive Bayes is a very simple model which can be used for classification. For document classification, we generate a label and all the words in the document given that label.
- Note that the words are all generated independently, which is not a very realistic model of language, but naive Bayes models are surprisingly effective for tasks such as document classification.

Application: topic modeling

Question: given a text document, what topics is it about?



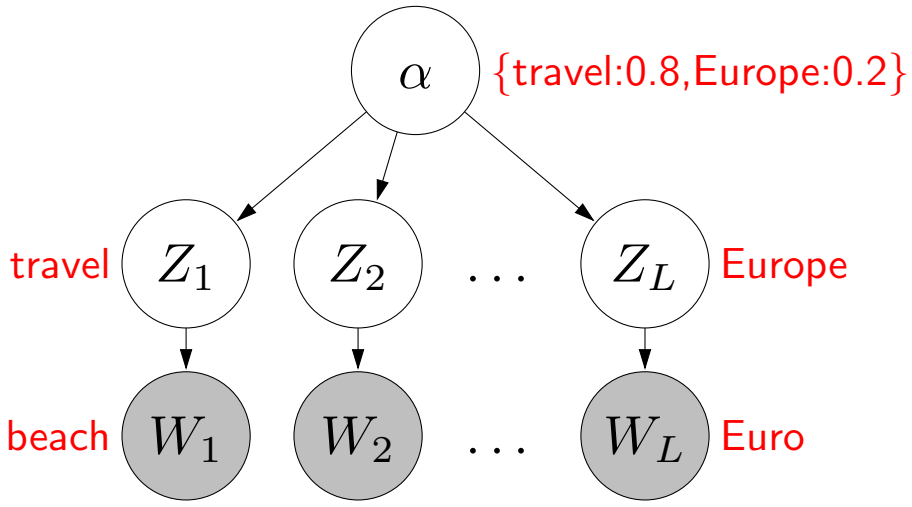
Probabilistic program: latent Dirichlet allocation

Generate a distribution over topics $\alpha \in \mathbb{R}^K$

For each position $i = 1, \ldots, L$:

Generate a topic $Z_i \sim p(Z_i \mid \alpha)$

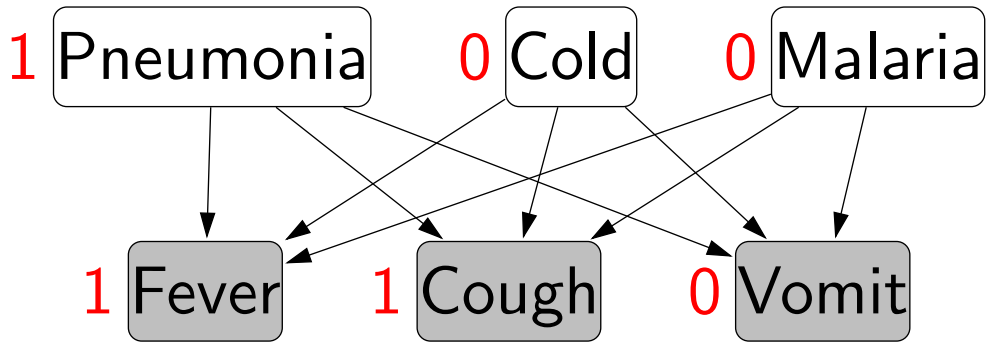
Generate a word $W_i \sim p(W_i \mid Z_i)$

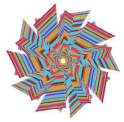


- A more sophisticated model of text is latent Dirichlet Allocation (LDA), which allows a document to not just be about one topic (which was true in naive Bayes), but about multiple topics.
- Here, the distribution over topics α is chosen per document from a Dirichlet distribution. Note that α is a continuous-valued random variable. For each position, we choose a topic according to that per-docuent distribution and generate a word given that topic.
- Latent Dirichlet Alloction (LDA) has been very infuential for modeling not only text but images, videos, music, etc.; any sort of data with hidden structure. It is very related to matrix factorization.

Application: medical diagnostics

Question: If patient has has a cough and fever, what disease(s) does he/she have?





Probabilistic program: diseases and symptoms

For each disease $i = 1, \ldots, m$:

Generate activity of disease $D_i \sim p(D_i)$

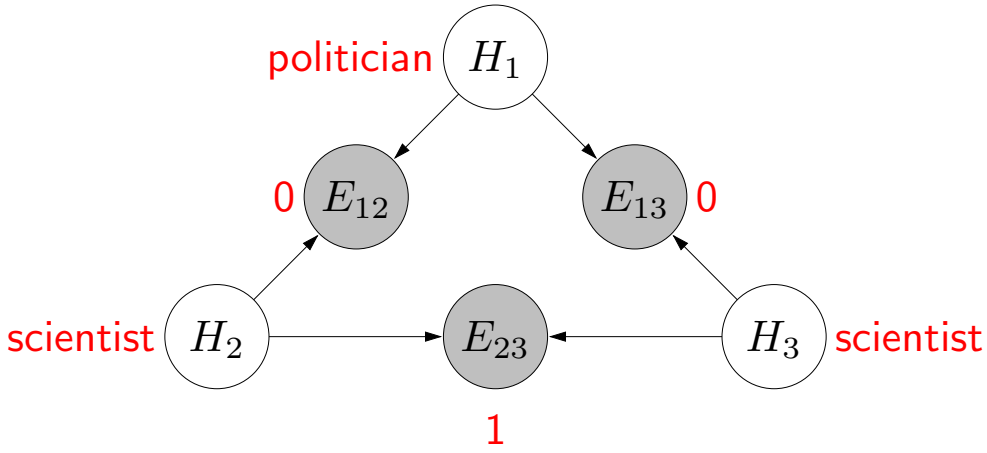
For each symptom $j = 1, \ldots, n$:

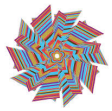
Generate activity of symptom $S_j \sim p(S_j \mid D_{1:m})$

- We already saw a special case of this model. In general, we would like to diagnose many diseases and might have measured many symptoms and vitals.

Application: social network analysis

Question: Given a social network (graph over n people), what types of people are there?





Probabilistic program: stochastic block model

For each person $i = 1, \ldots, n$:

Generate person type $H_i \sim p(H_i)$

For each pair of people $i \neq j$:

Generate connectedness $E_{ij} \sim p(E_{ij} \mid H_i, H_j)$

- One can also model graphs such as social networks. A very naive-Bayes-like model is that each node (person) has a "type". Whether two people interact with each other is determined solely by their types and random chance.
- Note: there are extensions called mixed membership models which, like LDA, allow each person to have multiple types.
- In summary, it is quite easy to come up with probabilistic programs that tell a story of how the world works for the domain of interest. These probabilistic programs define joint distributions over assignments to a collection of variables. Usually, these programs describe how some collection of hidden variables H that you're interested in behave, and then describe the generation of the evidence E that you see conditioned on H. After defining the model, one can do probabilistic inference to compute $\mathbb{P}(H \mid E = e)$.

Summary

Bayesian networks: modular definition of large joint distribution over variables

Probabilistic inference: condition on evidence, query variables of interest

Independence: restrictions on the set of possible distributions