Directed Acyclic Graphical Models, and Causal Models
9.S918: Statistical Inference in Brain and Cognitive Sciences
Spring 2024

Roger Levy

Massachusetts Institute of Technology

11 April 2024

Today's content

- Conditional Independence
- Bayes Nets (a.k.a. directed acyclic graphical models, DAGs)

(Conditional) Independence

Events A and B are said to be Conditionally Independent given information C if

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

Conditional independence of A and B given C is often expressed as

$$A \perp B|C$$

Directed graphical models

- A lot of the interesting joint probability distributions that arise in science and practical applications alike involve *conditional independencies* among the variables

Directed graphical models

- A lot of the interesting joint probability distributions that arise in science and practical applications alike involve *conditional independencies* among the variables
- So next is an introduction to a general framework for specifying conditional independencies among collections of random variables

Directed graphical models

- A lot of the interesting joint probability distributions that arise in science and practical applications alike involve *conditional independencies* among the variables
- So next is an introduction to a general framework for specifying conditional independencies among collections of random variables
- It won't allow us to express *all possible* independencies that may hold, but it goes a long way

Directed graphical models

- A lot of the interesting joint probability distributions that arise in science and practical applications alike involve *conditional independencies* among the variables
- So next is an introduction to a general framework for specifying conditional independencies among collections of random variables
- It won't allow us to express *all possible* independencies that may hold, but it goes a long way
- And I hope that you'll agree that the framework is intuitive too!

Directed graphical models

- A lot of the interesting joint probability distributions that arise in science and practical applications alike involve *conditional independencies* among the variables
- So next is an introduction to a general framework for specifying conditional independencies among collections of random variables
- It won't allow us to express *all possible* independencies that may hold, but it goes a long way
- And I hope that you'll agree that the framework is intuitive too!
- The intuitiveness is because a causal interpretation of the framework is natural—and, indeed, this is formalized in the causal treatment of Bayes nets

The coin factory

- Imagine a factory that produces three types of coins in equal volumes:

The coin factory

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;

The coin factory

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - 2-headed coins;

The coin factory

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - 2-headed coins;
  - 2-tailed coins.

The coin factory

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - 2-headed coins;
  - 2-tailed coins.
- Generative process:

The coin factory

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - 2-headed coins;
  - 2-tailed coins.
- Generative process:
  - The factory produces a coin of type X and sends it to you;

The coin factory

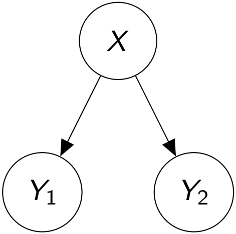
- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - 2-headed coins;
  - 2-tailed coins.
- Generative process:
  - The factory produces a coin of type X and sends it to you;
  - You receive the coin and flip it twice, with H(eads)/T(ails) outcomes Y_1 and Y_2

The coin factory

- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - 2-headed coins;
  - 2-tailed coins.
- Generative process:
  - The factory produces a coin of type X and sends it to you;
  - You receive the coin and flip it twice, with H(eads)/T(ails) outcomes Y_1 and Y_2
- Receiving a coin from the factory and flipping it twice is **sampling** (or **taking a sample**) from the joint distribution $P(X, Y_1, Y_2)$

This generative process is a Bayes Net

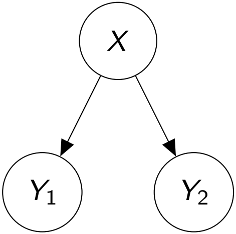
The directed acyclic graphical model (DAG), or Bayes net:



- Semantics of a Bayes net: the joint distribution can be expressed as the product of the conditional distributions of each variable **given only its parents**

This generative process is a Bayes Net

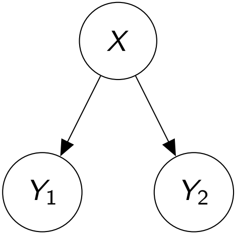
The directed acyclic graphical model (DAG), or Bayes net:



- Semantics of a Bayes net: the joint distribution can be expressed as the product of the conditional distributions of each variable **given only its parents**
- In this DAG, $P(X, Y_1, Y_2) = P(X)P(Y_1|X)P(Y_2|X)$

This generative process is a Bayes Net

The directed acyclic graphical model (DAG), or Bayes net:

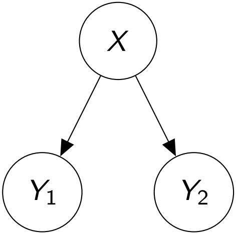


- Semantics of a Bayes net: the joint distribution can be expressed as the product of the conditional distributions of each variable **given only its parents**
- In this DAG, $P(X, Y_1, Y_2) = P(X)P(Y_1|X)P(Y_2|X)$

X	$P(X)$
Fair	$\frac{1}{3}$
2-H	$\frac{1}{3}$
2-T	$\frac{1}{3}$

This generative process is a Bayes Net

The directed acyclic graphical model (DAG), or Bayes net:



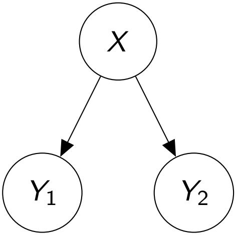
- Semantics of a Bayes net: the joint distribution can be expressed as the product of the conditional distributions of each variable **given only its parents**
- In this DAG, $P(X, Y_1, Y_2) = P(X)P(Y_1|X)P(Y_2|X)$

X	$P(X)$
Fair	$\frac{1}{3}$
2-H	$\frac{1}{3}$
2-T	$\frac{1}{3}$

X	$P(Y_1 = \mathrm{H}|X)$	$P(Y_1 = \mathrm{T}|X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

This generative process is a Bayes Net

The directed acyclic graphical model (DAG), or Bayes net:



- Semantics of a Bayes net: the joint distribution can be expressed as the product of the conditional distributions of each variable **given only its parents**
- In this DAG, $P(X, Y_1, Y_2) = P(X)P(Y_1|X)P(Y_2|X)$

X	$P(X)$
Fair	$\frac{1}{3}$
2-H	$\frac{1}{3}$
2-T	$\frac{1}{3}$

X	$P(Y_1 = \mathrm{H}\|X)$	$P(Y_1 = \mathrm{T}\|X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

X	$P(Y_2 = \mathrm{H}\|X)$	$P(Y_2 = \mathrm{T}\|X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

Conditional independence in Bayes nets

X	$P(X)$
Fair	$\frac{1}{3}$
2-H	$\frac{1}{3}$
2-T	$\frac{1}{3}$

X	$P(Y_1 = \mathrm{H} \mid X)$	$P(Y_1 = \mathrm{T} \mid X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

X	$P(Y_2 = \mathrm{H} \mid X)$	$P(Y_2 = \mathrm{T} \mid X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

Question:

- *Conditioned on not having any further information, are the two coin flips Y_1 and Y_2 in this generative process independent?*

Conditional independence in Bayes nets

X	$P(X)$
Fair	$\frac{1}{3}$
2-H	$\frac{1}{3}$
2-T	$\frac{1}{3}$

X	$P(Y_1 = \mathrm{H}\vert X)$	$P(Y_1 = \mathrm{T}\vert X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

X	$P(Y_2 = \mathrm{H}\vert X)$	$P(Y_2 = \mathrm{T}\vert X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

Question:

- *Conditioned on not having any further information, are the two coin flips Y_1 and Y_2 in this generative process independent?*
- "Independent" needs further interpretation! It might mean: is it the case that $Y_1 \perp Y_2 | \{\}$?

Conditional independence in Bayes nets

X	$P(X)$
Fair	$\frac{1}{3}$
2-H	$\frac{1}{3}$
2-T	$\frac{1}{3}$

X	$P(Y_1 = \mathrm{H}\vert X)$	$P(Y_1 = \mathrm{T}\vert X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

X	$P(Y_2 = \mathrm{H}\vert X)$	$P(Y_2 = \mathrm{T}\vert X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

Question:

- *Conditioned on not having any further information, are the two coin flips Y_1 and Y_2 in this generative process independent?*
- "Independent" needs further interpretation! It might mean: is it the case that $Y_1 \perp Y_2 | \{\}$?
- The answer to this question is **No!**

Conditional independence in Bayes nets

X	$P(X)$
Fair	$\frac{1}{3}$
2-H	$\frac{1}{3}$
2-T	$\frac{1}{3}$

X	$P(Y_1 = \mathrm{H}\|X)$	$P(Y_1 = \mathrm{T}\|X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

X	$P(Y_2 = \mathrm{H}\|X)$	$P(Y_2 = \mathrm{T}\|X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

Question:

- *Conditioned on not having any further information, are the two coin flips Y_1 and Y_2 in this generative process independent?*
- "Independent" needs further interpretation! It might mean: is it the case that $Y_1 \perp Y_2 | \{\}$?
- The answer to this question is **No!**
  - $P(Y_2 = H) = \frac{1}{2}$ (you can see this by symmetry)

Conditional independence in Bayes nets

X	$P(X)$
Fair	$\frac{1}{3}$
2-H	$\frac{1}{3}$
2-T	$\frac{1}{3}$

X	$P(Y_1 = \text{H}\vert X)$	$P(Y_1 = \text{T}\vert X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

X	$P(Y_2 = \text{H}\vert X)$	$P(Y_2 = \text{T}\vert X)$
Fair	$\frac{1}{2}$	$\frac{1}{2}$
2-H	1	0
2-T	0	1

Question:

- *Conditioned on not having any further information, are the two coin flips Y_1 and Y_2 in this generative process independent?*
- "Independent" needs further interpretation! It might mean: is it the case that $Y_1 \perp Y_2 | \{\}$?
- The answer to this question is **No!**
  - $P(Y_2 = H) = \frac{1}{2}$ (you can see this by symmetry)
  - But $P(Y_2 = H | Y_1 = H) = \overbrace{\frac{1}{3} \times \frac{1}{2}}^{\text{Coin was fair}} + \overbrace{\frac{2}{3} \times 1}^{\text{Coin was 2-H}} = \frac{5}{6}$

Formally assessing conditional independence in Bayes Nets

- The comprehensive criterion for assessing conditional independence is known as D-separation.

Formally assessing conditional independence in Bayes Nets

- The comprehensive criterion for assessing conditional independence is known as D-separation.
- A path between two disjoint node sets A and B is a sequence of edges connecting some node in A with some node in B

Formally assessing conditional independence in Bayes Nets

- The comprehensive criterion for assessing conditional independence is known as D-separation.
- A path between two disjoint node sets A and B is a sequence of edges connecting some node in A with some node in B
- Any node on a given path has converging arrows if two edges on the path connect to it and point to it.

Formally assessing conditional independence in Bayes Nets

- The comprehensive criterion for assessing conditional independence is known as D-separation.
- A path between two disjoint node sets A and B is a sequence of edges connecting some node in A with some node in B
- Any node on a given path has converging arrows if two edges on the path connect to it and point to it.
- A node on the path has non-converging arrows if two edges on the path connect to it, but at least one does not point to it.

Formally assessing conditional independence in Bayes Nets

- The comprehensive criterion for assessing conditional independence is known as D-separation.
- A path between two disjoint node sets A and B is a sequence of edges connecting some node in A with some node in B
- Any node on a given path has converging arrows if two edges on the path connect to it and point to it.
- A node on the path has non-converging arrows if two edges on the path connect to it, but at least one does not point to it.
- A third disjoint node set C d-separates A and B if for every path between A and B, either:
    1. there is some node N on the path whose arrows do not converge and which *is* in C; or
    2. there is some node N on the path with converging arrows, and neither N nor any of its descendants is in C.

Major types of d-separation

A node set C d-separates A and B if for every path between A and B, either:

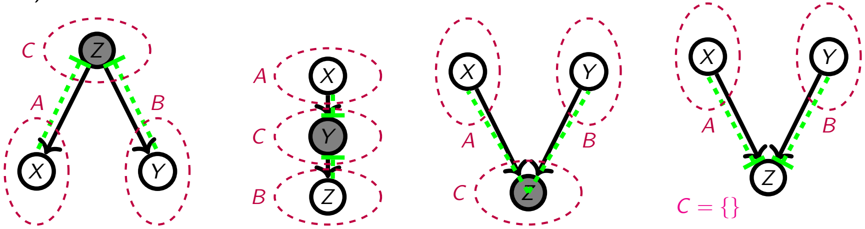
1. there is some node N on the path whose arrows do not converge and which *is* in C; or
2. there is some node N on the path with converging arrows, and neither N nor any of its descendants is in C.

Common-cause d-separation (from knowing Z)

Mediating d-separation (from knowing Y)

Explaining away: knowing Z prevents d-separation

D-separation in the absence of knowledge of Z



(Shaded node=in C)

D-separation and conditional independence

A node set C d-separates A and B if for every path between A and B, either:

1. there is some node N on the path whose arrows do not converge and which *is* in C; or
2. there is some node N on the path with converging arrows, and neither N nor any of its descendants is in C.

- If C d-separates A and B, then

$$A \perp B | C$$

D-separation and conditional independence

A node set C d-separates A and B if for every path between A and B, either:

1. there is some node N on the path whose arrows do not converge and which *is* in C; or
2. there is some node N on the path with converging arrows, and neither N nor any of its descendants is in C.

- If C d-separates A and B, then

$$A \perp B | C$$

- **Caution:** the converse is *not* the case: $A \perp B | C$ does not necessarily imply that the joint distribution on all the random variables in $A \cup B \cup C$ can be represented with a Bayes Net in which C d-separates A and B.

D-separation and conditional independence

A node set C d-separates A and B if for every path between A and B, either:

1. there is some node N on the path whose arrows do not converge and which *is* in C; or
2. there is some node N on the path with converging arrows, and neither N nor any of its descendants is in C.

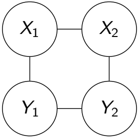
- If C d-separates A and B, then

$$A \perp B | C$$

- **Caution:** the converse is *not* the case: $A \perp B | C$ does not necessarily imply that the joint distribution on all the random variables in $A \cup B \cup C$ can be represented with a Bayes Net in which C d-separates A and B.
  - **Example:** let X_1, X_2, Y_1, Y_2 each be 0/1 random variable, and let the joint distribution reflect the constraint that $Y_1 = (X_1 == X_2)$ and $Y_2 = \text{xor}(X_1, X_2)$. This gives us $Y_1 \perp Y_2 | \{X_1, X_2\}$, but you won't be able to write a Bayes net involving these four variables such that $\{X_1, X_2\}$ d-separates Y_1 and Y_2.

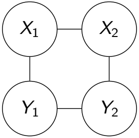
Conditional independencies not expressible in a Bayes net

- **Example:** let X_1, X_2, Y_1, Y_2 each be binary 0/1 random variables, in the following arrangement on an **undirected** graph:



Conditional independencies not expressible in a Bayes net

- **Example:** let X_1, X_2, Y_1, Y_2 each be binary 0/1 random variables, in the following arrangement on an **undirected** graph:



$$\begin{aligned}
f_1(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq X_2) \\
f_2(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq Y_1) \\
f_3(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_2 \neq Y_2) \\
f_4(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(Y_1 \neq Y_2)
\end{aligned}$$

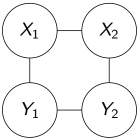
- Suppose the joint distribution is determined entirely by adjacent nodes "liking" to have the same value. Formally, for example:

$$P(X_1, X_2, Y_1, Y_2) \propto \prod_{i=1}^{4} \left(\frac{1}{2}\right)^{f_i(X_1, X_2, Y_1, Y_2)}$$

(Most probable outcomes, each with prob. 0.195: either all 0s, or all 1s)

Conditional independencies not expressible in a Bayes net

- **Example:** let X_1, X_2, Y_1, Y_2 each be binary 0/1 random variables, in the following arrangement on an **undirected** graph:



$$\begin{aligned} f_1(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq X_2) \\ f_2(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq Y_1) \\ f_3(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_2 \neq Y_2) \\ f_4(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(Y_1 \neq Y_2) \end{aligned}$$

- Suppose the joint distribution is determined entirely by adjacent nodes "liking" to have the same value. Formally, for example:

$$P(X_1, X_2, Y_1, Y_2) \propto \prod_{i=1}^{4} \left(\frac{1}{2}\right)^{f_i(X_1, X_2, Y_1, Y_2)}$$

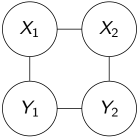
(Most probable outcomes, each with prob. 0.195: either all 0s, or all 1s)

- In this model, both the following conditional independencies hold:

$$X_1 \perp Y_2 | \{X_2, Y_1\} \qquad X_2 \perp Y_1 | \{X_1, Y_2\}$$

Conditional independencies not expressible in a Bayes net

- **Example:** let X_1, X_2, Y_1, Y_2 each be binary 0/1 random variables, in the following arrangement on an **undirected** graph:



$$
\begin{aligned}
f_1(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq X_2) \\
f_2(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq Y_1) \\
f_3(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_2 \neq Y_2) \\
f_4(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(Y_1 \neq Y_2)
\end{aligned}
$$

- Suppose the joint distribution is determined entirely by adjacent nodes "liking" to have the same value. Formally, for example:

$$P(X_1, X_2, Y_1, Y_2) \propto \prod_{i=1}^{4} \left(\frac{1}{2}\right)^{f_i(X_1, X_2, Y_1, Y_2)}$$

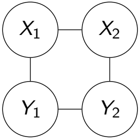
(Most probable outcomes, each with prob. 0.195: either all 0s, or all 1s)

- In this model, both the following conditional independencies hold:

$$X_1 \perp Y_2 | \{X_2, Y_1\} \qquad\qquad X_2 \perp Y_1 | \{X_1, Y_2\}$$

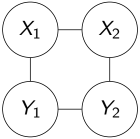
- But this set of conditional independencies cannot be expressed in a Bayes Net.

Conditional independencies not expressible in a Bayes net



$$
\begin{aligned}
f_1(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq X_2) \\
f_2(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq Y_1) \\
f_3(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_2 \neq Y_2) \\
f_4(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(Y_1 \neq Y_2)
\end{aligned}
$$

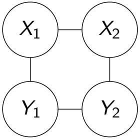
Conditional independencies not expressible in a Bayes net



$$
\begin{aligned}
f_1(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq X_2) \\
f_2(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq Y_1) \\
f_3(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_2 \neq Y_2) \\
f_4(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(Y_1 \neq Y_2)
\end{aligned}
$$

- This example is an instance of an Ising model, the prototypical case of a Markov random field, a model class that can be represented as undirected graphs

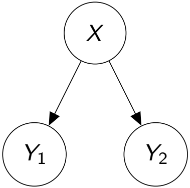
Conditional independencies not expressible in a Bayes net



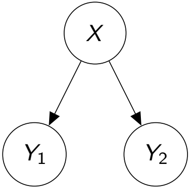
$$
\begin{aligned}
f_1(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq X_2) \\
f_2(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_1 \neq Y_1) \\
f_3(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(X_2 \neq Y_2) \\
f_4(X_1, X_2, Y_1, Y_2) &= \mathbf{I}(Y_1 \neq Y_2)
\end{aligned}
$$

- This example is an instance of an Ising model, the prototypical case of a Markov random field, a model class that can be represented as undirected graphs
- We won't look at these further, but you can read about them in books and papers about graphical models (e.g., (Bishop, 2006, Section 8.3)

Back to our example

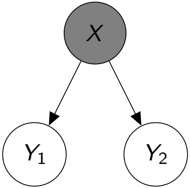


Back to our example



- *Without looking at the coin before flipping it*, the outcome Y_1 of the first flip gives me information about the type of coin, and affects my beliefs about the outcome of Y_2

Back to our example



- *Without looking at the coin before flipping it*, the outcome Y_1 of the first flip gives me information about the type of coin, and affects my beliefs about the outcome of Y_2

- But if I *look* at the coin before flipping it, Y_1 and Y_2 are rendered independent

An example of explaining away

I saw an exhibition about the, uh...

An example of explaining away

I saw an exhibition about the, uh. . .

There are several causes of disfluency, including:

An example of explaining away

I saw an exhibition about the, uh...

There are several causes of disfluency, including:

- An upcoming word is difficult to produce (e.g., low frequency, *astrolabe*)

An example of explaining away

I saw an exhibition about the, uh...

There are several causes of disfluency, including:

- An upcoming word is difficult to produce (e.g., low frequency, *astrolabe*)
- The speaker's attention was distracted by something in the non-linguistic environment

An example of explaining away

I saw an exhibition about the, uh...

There are several causes of disfluency, including:

- An upcoming word is difficult to produce (e.g., low frequency, *astrolabe*)
- The speaker's attention was distracted by something in the non-linguistic environment

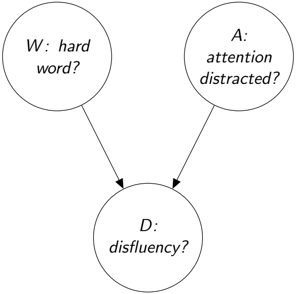
An example of explaining away

I saw an exhibition about the, uh...

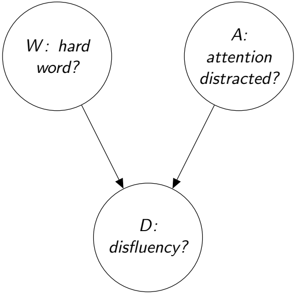
There are several causes of disfluency, including:

- An upcoming word is difficult to produce (e.g., low frequency, *astrolabe*)
- The speaker's attention was distracted by something in the non-linguistic environment

A reasonable graphical model:

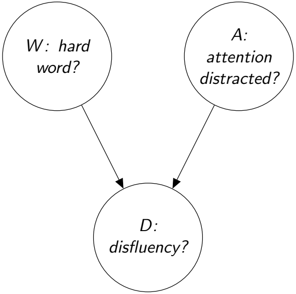


An example of explaining away



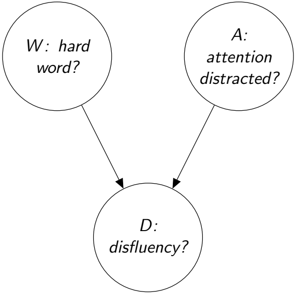
- Without knowledge of D, there's no reason to expect that W and A are correlated

An example of explaining away



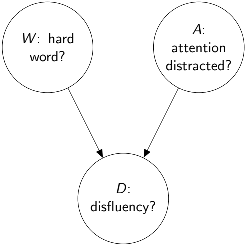
- Without knowledge of D, there's no reason to expect that W and A are correlated
- But hearing a disfluency *demands a cause*

An example of explaining away



- Without knowledge of D, there's no reason to expect that W and A are correlated
- But hearing a disfluency *demands a cause*
- Knowing that there was a distraction *explains away* the disfluency, reducing the probability that the speaker was planning to utter a hard word

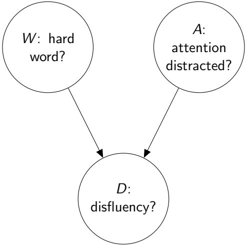
An example of the disfluency model



- Let's suppose that both hard words and distractions are unusual, the latter more so

$$P(W = \text{hard}) = 0.25$$
$$P(A = \text{distracted}) = 0.15$$

An example of the disfluency model



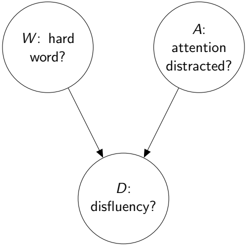
- Let's suppose that both hard words and distractions are unusual, the latter more so

$$P(W = \text{hard}) = 0.25$$
$$P(A = \text{distracted}) = 0.15$$

- Hard words and distractions both induce disfluencies; having both makes a disfluency *really* likely

W	A	D=no disfluency	D=disfluency
easy	undistracted	0.99	0.01
easy	distracted	0.7	0.3
hard	undistracted	0.85	0.15
hard	distracted	0.4	0.6

An example of the disfluency model

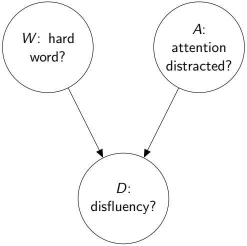


$$P(W = \text{hard}) = 0.25$$
$$P(A = \text{distracted}) = 0.15$$

W	A	D=no disfluency	D=disfluency
easy	undistracted	0.99	0.01
easy	distracted	0.7	0.3
hard	undistracted	0.85	0.15
hard	distracted	0.4	0.6

- Suppose that we observe the speaker uttering a disfluency. What is $P(W = \text{hard}|D = \text{disfluent})$?

An example of the disfluency model

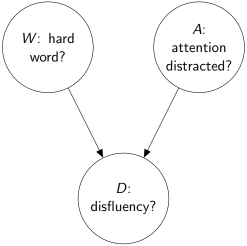


$$P(W = \text{hard}) = 0.25$$
$$P(A = \text{distracted}) = 0.15$$

W	A	D=no disfluency	D=disfluency
easy	undistracted	0.99	0.01
easy	distracted	0.7	0.3
hard	undistracted	0.85	0.15
hard	distracted	0.4	0.6

- Suppose that we observe the speaker uttering a disfluency. What is $P(W = \text{hard} | D = \text{disfluent})$?
- Now suppose we also learn that her attention is distracted. What does that do to our beliefs about W

An example of the disfluency model



$$P(W = \text{hard}) = 0.25$$
$$P(A = \text{distracted}) = 0.15$$

W	A	D=no disfluency	D=disfluency
easy	undistracted	0.99	0.01
easy	distracted	0.7	0.3
hard	undistracted	0.85	0.15
hard	distracted	0.4	0.6

- Suppose that we observe the speaker uttering a disfluency. What is $P(W = \text{hard}|D = \text{disfluent})$?
- Now suppose we also learn that her attention is distracted. What does that do to our beliefs about W
- That is, what is $P(W = \text{hard}|D = \text{disfluent}, A = \text{distracted})$?

An example of the disfluency model

Fortunately, there is automated machinery to "turn the Bayesian crank":

$$P(W = \text{hard}) = 0.25$$

An example of the disfluency model

Fortunately, there is automated machinery to “turn the Bayesian crank”:

$$P(W = \text{hard}) = 0.25$$
$$P(W = \text{hard} | D = \text{disfluent}) = 0.57$$

An example of the disfluency model

Fortunately, there is automated machinery to "turn the Bayesian crank":

$$
\begin{aligned}
P(W = \text{hard}) &= 0.25 \\
P(W = \text{hard} | D = \text{disfluent}) &= 0.57 \\
P(W = \text{hard} | D = \text{disfluent}, A = \text{distracted}) &= 0.40
\end{aligned}
$$

An example of the disfluency model

Fortunately, there is automated machinery to "turn the Bayesian crank":

$$
\begin{aligned}
P(W=\text{hard}) &= 0.25 \\
P(W=\text{hard}|D=\text{disfluent}) &= 0.57 \\
P(W=\text{hard}|D=\text{disfluent}, A=\text{distracted}) &= 0.40
\end{aligned}
$$

- Knowing that the speaker was distracted (A) *decreased* the probability that the speaker was about to utter a hard word (W)—A **explained** D **away**.

An example of the disfluency model

Fortunately, there is automated machinery to "turn the Bayesian crank":

$$
\begin{aligned}
P(W = \text{hard}) &= 0.25 \\
P(W = \text{hard}|D = \text{disfluent}) &= 0.57 \\
P(W = \text{hard}|D = \text{disfluent}, A = \text{distracted}) &= 0.40
\end{aligned}
$$

- Knowing that the speaker was distracted (A) *decreased* the probability that the speaker was about to utter a hard word (W)—A **explained** D **away**.
- A caveat: the type of relationship among A, W, and D will depend on the values one finds in the probability table!

$$
\begin{aligned}
&P(W) \\
&P(A) \\
&P(D|W, A)
\end{aligned}
$$

Summary thus far

Key points:

- Bayes' Rule is a compelling framework for modeling inference under uncertainty
- DAGs/Bayes Nets are a broad class of models for specifying joint probability distributions with conditional independencies
- Classic Bayes Net references: Pearl (1988, 2000); Jordan (1998); Russell and Norvig (2003, Chapter 14); Bishop (2006, Chapter 8).

An example of the disfluency model

$$\boxed{P(W = \text{hard}|D = \text{disfluent}, A = \text{distracted})}$$

hard	W=hard
easy	W=easy
disfl	D=disfluent
distr	A=distracted
undistr	A=undistracted

$$P(\text{hard}|\text{disfl, distr}) = \frac{P(\text{disfl}|\text{hard, distr})P(\text{hard}|\text{distr})}{P(\text{disfl}|\text{distr})} \quad \text{(Bayes' Rule)}$$

$$= \frac{P(\text{disfl}|\text{hard, distr})P(\text{hard})}{P(\text{disfl}|\text{distr})} \quad \text{(Independence from the DAG)}$$

$$P(\text{disfl}|\text{distr}) = \sum_{w'} P(\text{disfl}|W = w')P(W = w') \quad \text{(Marginalization)}$$

$$= P(\text{disfl}|\text{hard})P(\text{hard}) + P(\text{disfl}|\text{easy})P(\text{easy})$$

$$= 0.6 \times 0.25 + 0.3 \times 0.75$$

$$= 0.375$$

$$P(\text{hard}|\text{disfl, distr}) = \frac{0.6 \times 0.25}{0.375}$$

$$= 0.4$$

An example of the disfluency model

$$\boxed{P(W = \text{hard} | D = \text{disfluent})}$$

$$
\begin{aligned}
P(\text{hard}|\text{disfl}) &= \frac{P(\text{disfl}|\text{hard})P(\text{hard})}{P(\text{disfl})} && \text{(Bayes' Rule)}\\
P(\text{disfl}|\text{hard}) &= \sum_{a'} P(\text{disfl}|A = a', \text{hard})P(A = a'|\text{hard})\\
&= P(\text{disfl}|A = \text{distr}, \text{hard})P(A = \text{distr}|\text{hard}) + P(\text{disfl}|\text{undistr}, \text{hard})P(\text{undistr}|\text{hard})\\
&= 0.6 \times 0.15 + 0.15 \times 0.85\\
&= 0.2175\\
P(\text{disfl}) &= \sum_{w'} P(\text{disfl}|W = w')P(W = w')\\
&= P(\text{disfl}|\text{hard})P(\text{hard}) + P(\text{disfl}|\text{easy})P(\text{easy})\\
P(\text{disfl}|\text{easy}) &= \sum_{a'} P(\text{disfl}|A = a', \text{easy})P(A = a'|\text{easy})\\
&= P(\text{disfl}|A = \text{distr}, \text{easy})P(A = \text{distr}|\text{easy}) + P(\text{disfl}|\text{undistr}, \text{easy})P(\text{undistr}|\text{easy})\\
&= 0.3 \times 0.15 + 0.01 \times 0.85\\
&= 0.0535\\
P(\text{disfl}) &= 0.2175 \times 0.25 + 0.0535 \times 0.75\\
&= 0.0945\\
P(\text{hard}|\text{disfl}) &= \frac{0.2175 \times 0.25}{0.0945}\\
&= 0.575396825396825
\end{aligned}
$$

Interventions

- Suppose we have a collection of random variables V that follow some joint probability distribution. We define an "intervention" operator that can be conditioned on in probabilistic queries (Pearl, 2009):

$$\mathrm{Do}(\cdot)$$

Interventions

- Suppose we have a collection of random variables V that follow some joint probability distribution. We define an "intervention" operator that can be conditioned on in probabilistic queries (Pearl, 2009):

$$\mathrm{Do}(\cdot)$$

- Intuitively, conditioning on $\mathrm{Do}(X = x)$, where $X \subseteq V$ and x are values for X, means **"intervening" exogeneously to the "system" constituted by** V, to "set" the value(s) of X to x.

Interventions

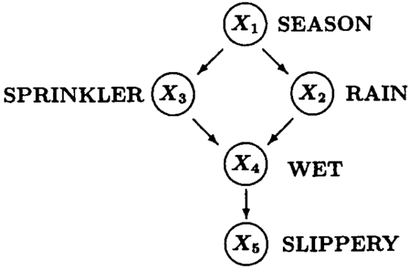
- Suppose we have a collection of random variables V that follow some joint probability distribution. We define an "intervention" operator that can be conditioned on in probabilistic queries (Pearl, 2009):

$$\mathrm{Do}(\cdot)$$

- Intuitively, conditioning on $\mathrm{Do}(X = x)$, where $X \subseteq V$ and x are values for X, means **"intervening" exogeneously to the "system" constituted by** V, to "set" the value(s) of X to x.
- In general, $P(V|X)$ and $P(V|\mathrm{Do}(X = x))$ will **NOT** be the same distribution. $P(V|\mathrm{Do}(X = x))$, also notated as $P_x(V)$, is sometimes called an interventional distribution.

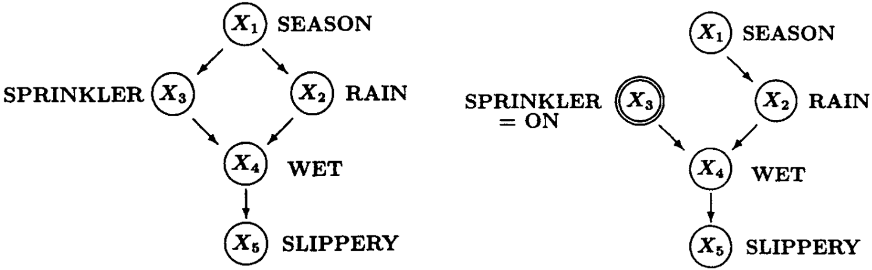
Causal Bayes Nets and interventions as "graph surgery"

- If V can be organized into a causal Bayes Net G, then the relationship between the base joint distribution (no interventions) and the set of interventional distributions can be characterized succinctly.



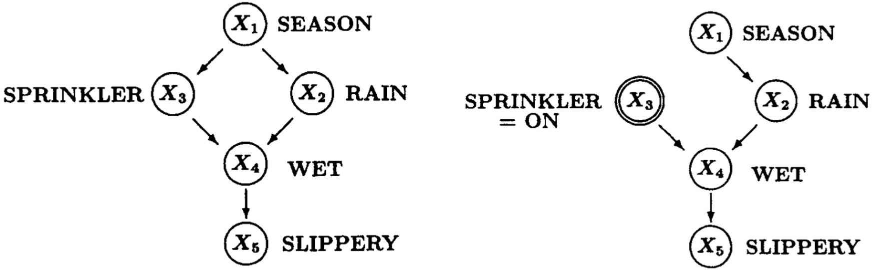
Causal Bayes Nets and interventions as “graph surgery”

- If V can be organized into a causal Bayes Net G, then the relationship between the base joint distribution (no interventions) and the set of interventional distributions can be characterized succinctly.
- To find $P(V|\mathrm{Do}(X = x)$, simply “cut” all the links in G between each variable in X and its parents to create a new graph G', and then do ordinary probabilistic conditioning $P(V|X = x)$ within G'.



Causal Bayes Nets and interventions as "graph surgery"

- If V can be organized into a causal Bayes Net G, then the relationship between the base joint distribution (no interventions) and the set of interventional distributions can be characterized succinctly.
- To find $P(V|\mathrm{Do}(X = x)$, simply "cut" all the links in G between each variable in X and its parents to create a new graph G', and then do ordinary probabilistic conditioning $P(V|X = x)$ within G'.
- This is sometimes called "graph surgery" (Spirtes et al., 1993)



Association versus causation

$$X \rightarrow Y \qquad\qquad X \leftarrow Y$$

- These two Bayes Nets encode identical constraints on the joint distribution $P(X, Y)$ **(review: what constraints are these?)**

Association versus causation

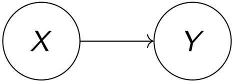
- These two Bayes Nets encode identical constraints on the joint distribution $P(X, Y)$ **(review: what constraints are these?)**
- (Answer: no constraints; any joint distribution is allowed!)

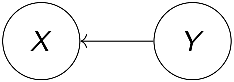
Association versus causation

$$X \rightarrow Y \qquad\qquad X \leftarrow Y$$

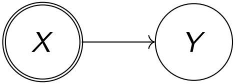
- These two Bayes Nets encode identical constraints on the joint distribution $P(X, Y)$ **(review: what constraints are these?)**
- (Answer: no constraints; any joint distribution is allowed!)
- However, if they are **causal** Bayes Nets, they are substantively different

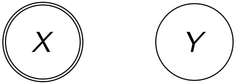
Association versus causation



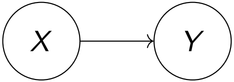


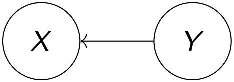
- These two Bayes Nets encode identical constraints on the joint distribution $P(X, Y)$ **(review: what constraints are these?)**
- (Answer: no constraints; any joint distribution is allowed!)
- However, if they are **causal** Bayes Nets, they are substantively different
- Intervening to set X to some value x has different consequences in the two causal Bayes Nets:



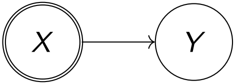


Association versus causation

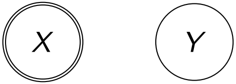




- These two Bayes Nets encode identical constraints on the joint distribution $P(X, Y)$ **(review: what constraints are these?)**
- (Answer: no constraints; any joint distribution is allowed!)
- However, if they are **causal** Bayes Nets, they are substantively different
- Intervening to set X to some value x has different consequences in the two causal Bayes Nets:

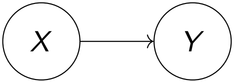


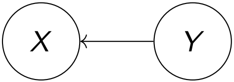
X = smoking
Y = lung cancer



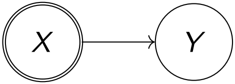
- This is a simple instance of distinguishing **association** from **causation**

Association versus causation

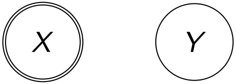




- These two Bayes Nets encode identical constraints on the joint distribution $P(X, Y)$ **(review: what constraints are these?)**
- (Answer: no constraints; any joint distribution is allowed!)
- However, if they are **causal** Bayes Nets, they are substantively different
- Intervening to set X to some value x has different consequences in the two causal Bayes Nets:



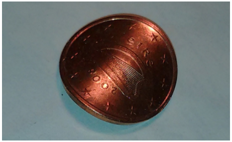
X = smoking
Y = lung cancer

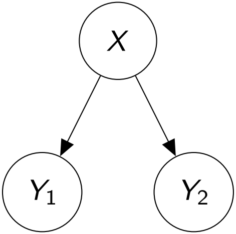


X = thermometer reading
Y = ambient temperature

- This is a simple instance of distinguishing **association** from **causation**

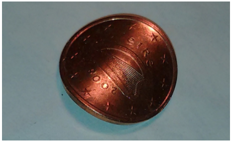
Back to our previous example, a bit modified

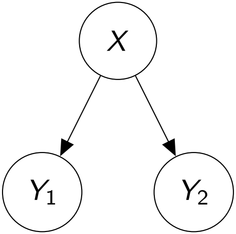




- Imagine a factory that produces three types of coins in equal volumes:

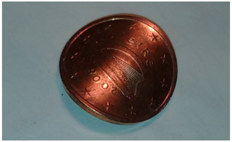
Back to our previous example, a bit modified

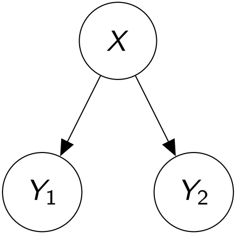




- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;

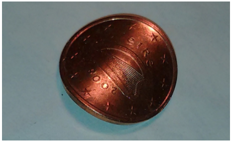
Back to our previous example, a bit modified

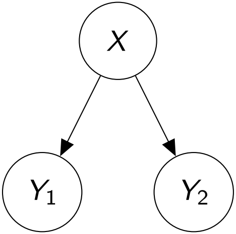




- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - A slightly **bent** coin that lands heads with $^3/_5$ probability;

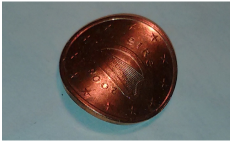
Back to our previous example, a bit modified

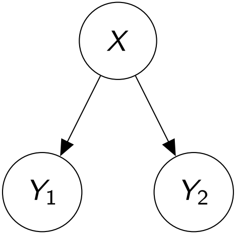




- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - A slightly **bent** coin that lands heads with $^3/_5$ probability;
  - A slightly bent coin that lands tails with $^3/_5$ probability.

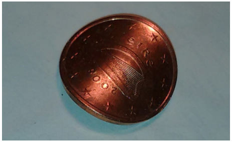
Back to our previous example, a bit modified

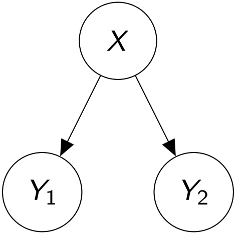




- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - A slightly **bent** coin that lands heads with $3/5$ probability;
  - A slightly bent coin that lands tails with $3/5$ probability.
- Generative process:

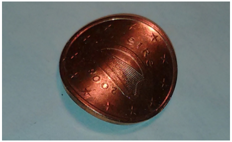
Back to our previous example, a bit modified

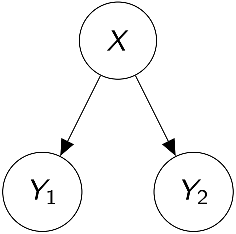




- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - A slightly **bent** coin that lands heads with $^3/_5$ probability;
  - A slightly bent coin that lands tails with $^3/_5$ probability.
- Generative process:
  - The factory produces a coin of type X and sends it to you;

Back to our previous example, a bit modified



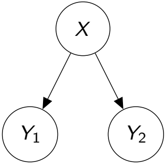


- Imagine a factory that produces three types of coins in equal volumes:
  - Fair coins;
  - A slightly **bent** coin that lands heads with $^3/_5$ probability;
  - A slightly bent coin that lands tails with $^3/_5$ probability.
- Generative process:
  - The factory produces a coin of type X and sends it to you;
  - You receive the coin and flip it twice, with H(eads)/T(ails) outcomes Y_1 and Y_2

Predictive value $\neq$ influence through intervention

Three types of coins in equal volumes:

- Fair coins;
- A slightly **bent** coin that lands heads with $3/5$ probability;
- A slightly bent coin that lands tails with $3/5$ probability.

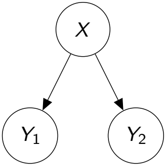


- The outcome of the first coin flip Y_1 has **predictive value** for the outcome of the second coin flip Y_2, and vice versa

Predictive value $\neq$ influence through intervention

Three types of coins in equal volumes:

- Fair coins;
- A slightly **bent** coin that lands heads with $^3/_5$ probability;
- A slightly bent coin that lands tails with $^3/_5$ probability.

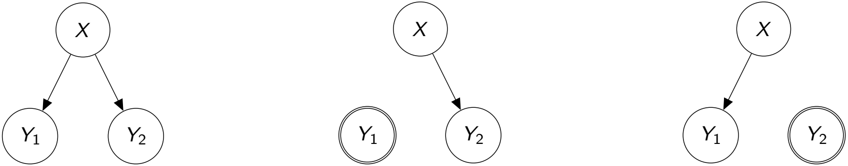


- The outcome of the first coin flip Y_1 has **predictive value** for the outcome of the second coin flip Y_2, and vice versa
- I could learn this **association** from observing pairs of flips of coins from the coin factory

Predictive value $\neq$ influence through intervention

Three types of coins in equal volumes:

- Fair coins;
- A slightly **bent** coin that lands heads with $^3/_5$ probability;
- A slightly bent coin that lands tails with $^3/_5$ probability.



- The outcome of the first coin flip Y_1 has **predictive value** for the outcome of the second coin flip Y_2, and vice versa
- I could learn this **association** from observing pairs of flips of coins from the coin factory
- But I cannot intervene on either variable to influence the other, because neither is causally upstream of the other!

Implications

- Throughout this class, we will endeavor to organize our information as much as possible into models that represent plausible causal chains of influence

Implications

- Throughout this class, we will endeavor to organize our information as much as possible into models that represent plausible causal chains of influence
- Typically, organizing information this way will help ensure that our statistical inferences actually answer our scientific questions of interest

Implications

- Throughout this class, we will endeavor to organize our information as much as possible into models that represent plausible causal chains of influence
- Typically, organizing information this way will help ensure that our statistical inferences actually answer our scientific questions of interest
- Traditional statistical tools are *associational*, so we need this top-down machinery (here, the mind of the scientist!) to ensure that they're being deployed appropriately

Implications

- Throughout this class, we will endeavor to organize our information as much as possible into models that represent plausible causal chains of influence
- Typically, organizing information this way will help ensure that our statistical inferences actually answer our scientific questions of interest
- Traditional statistical tools are *associational*, so we need this top-down machinery (here, the mind of the scientist!) to ensure that they're being deployed appropriately
- We must also stay cognizant of possible "unseen" latent causes, and that we may be uncertain about the true causal relationship among our observable variables

References I

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer.

Jordan, M. I., editor (1998). *Learning in Graphical Models*. Cambridge, MA: MIT Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, 2 edition.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*. Cambridge, second edition.

Russell, S. and Norvig, P. (2003). *Artificial Intelligence: a Modern Approach*. Prentice Hall, second edition edition.

Spirtes, P., Glymour, C. N., and Scheines, R. (1993). *Causation, prediction, and search*. MIT press.