Bayes Nets
9.19: Computational Psycholinguistics
Fall 2021

Roger Levy

Massachusetts Institute of Technology

18 October 2021

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

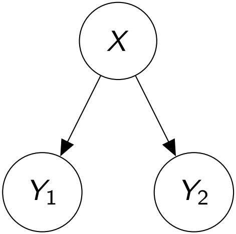
- A lot of the interesting joint probability distributions in the study of language involve *conditional independencies* among the variables
- So next we'll introduce you to a general framework for specifying conditional independencies among collections of random variables
- It won't allow us to express *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!

A non-linguistic example, redux

- 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**
- 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	$P(Y_1 = \mathrm{H}\|X)$	$P(Y_1 = \mathrm{T}\|X)$
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?*
- That is, is it the case that $Y_1 \perp Y_2 | \{\}$?
- **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

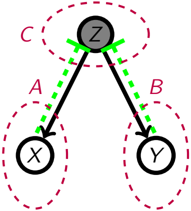
- The comprehensive criterion for assessing conditional independence is known as D-separation.
- A path between two disjoint node sets A and B is a sequence of edges connecting some node in A with some node in B
- Any node on a given path has converging arrows if two edges on the path connect to it and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

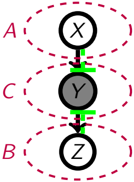
1. there is some node N on the path whose arrows do not converge and which *is* in C; or
2. there is some node N on the path with converging arrows, and neither N nor any of its descendants is in C.

Common-cause d-separation (from knowing Z)

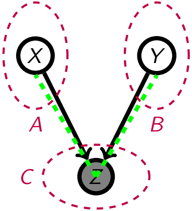


(Shaded node=in C)

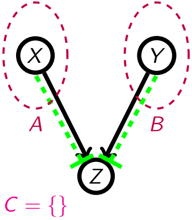
Intervening d-separation (from knowing Y)



Explaining away: knowing Z prevents d-separation



D-separation in the absence of knowledge of Z



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.

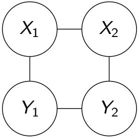
- If C d-separates A and B, then

$$A \perp B | C$$

- **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 expressable 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:

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

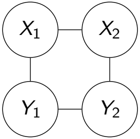
(Most probable outcomes, each with prob. 0.195: either all 0s, or all 1s)

- 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.

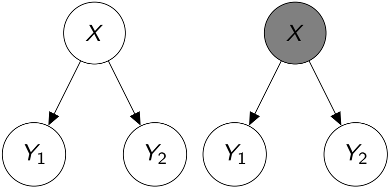
Conditional independencies not expressable in a Bayes net



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

- 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



- *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

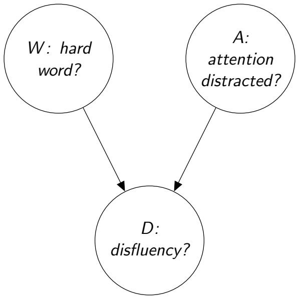
An example of explaining away

I saw an exhibition about the, uh. . .

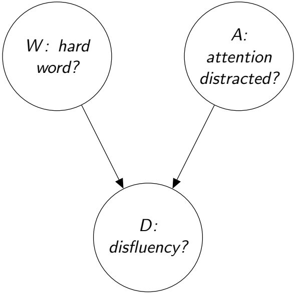
There are several causes of disfluency, including:

- 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:

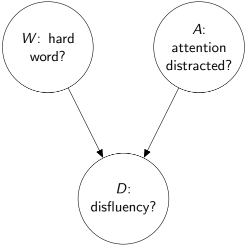


An example of explaining away



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



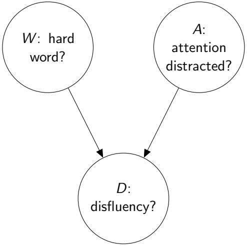
- Let's suppose that both hard words and distractions are unusual, the latter more so

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

- 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 $P(W = \text{hard}|D = \text{disfluent})$?
- 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":

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

- 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!

$$\begin{array}{l} P(W) \\ P(A) \\ P(D|W, A) \end{array}$$

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

$$
\begin{aligned}
P(\text{hard}|\text{disfl}, \text{distr}) &= \frac{P(\text{disfl}|\text{hard}, \text{distr})P(\text{hard}|\text{distr})}{P(\text{disfl}|\text{distr})} && \text{(Bayes' Rule)} \\
&= \frac{P(\text{disfl}|\text{hard}, \text{distr})P(\text{hard})}{P(\text{disfl}|\text{distr})} && \text{(Independence from the DAG)} \\
P(\text{disfl}|\text{distr}) &= \sum_{w'} P(\text{disfl}|W = w')P(W = w') && \text{(Marginalization)} \\
&= P(\text{disfl}|\text{hard})P(\text{hard}) + P(\text{disfl}|\text{easy})P(\text{easy}) \\
&= 0.6 \times 0.25 + 0.3 \times 0.75 \\
&= 0.375 \\
P(\text{hard}|\text{disfl}, \text{distr}) &= \frac{0.6 \times 0.25}{0.375} \\
&= 0.4
\end{aligned}
$$

An example of the disfluency model

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

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

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.

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