CS 5100: Foundations of Artificial Intelligence

Bayesian Networks

Prof. Amy Sliva
November 10, 2011

Outline

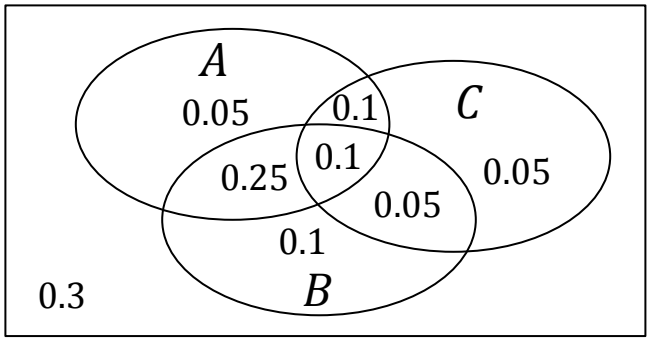
- Review joint and conditional probability
- Bayes rule
- Conditional independence
- Bayesian networks
- Naïve Bayes classification

- Thanks to Andrew Moore from CMU for some great slides (used with permission).

Joint probability distribution (JPD) review

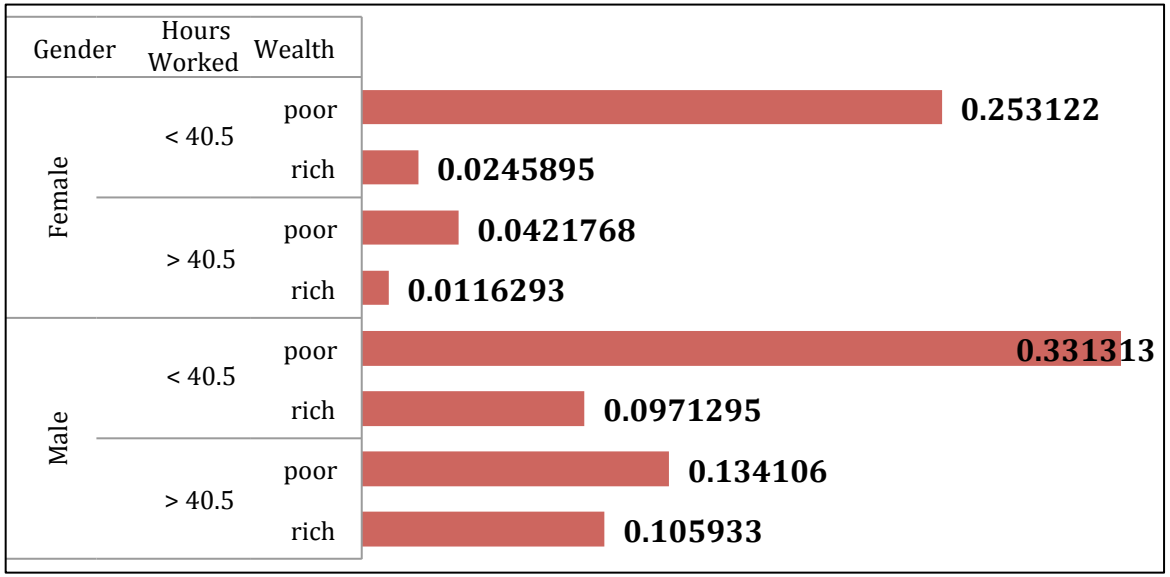
- Recipe for making a joint distribution of M variables:
  1. Make a truth table listing **all combinations of values** of the variables (if there are M boolean variables, then the table will have 2^M rows)
  2. For each combination of values, say how probable it is
  3. Following the **axioms of probability**, those numbers must sum to 1

A	*B*	*C*	Probability
F	F	F	0.3
F	F	T	0.05
F	T	F	0.1
F	T	T	0.05
T	F	F	0.05
T	F	T	0.1
T	T	F	0.25
T	T	T	0.1



Using the joint distribution

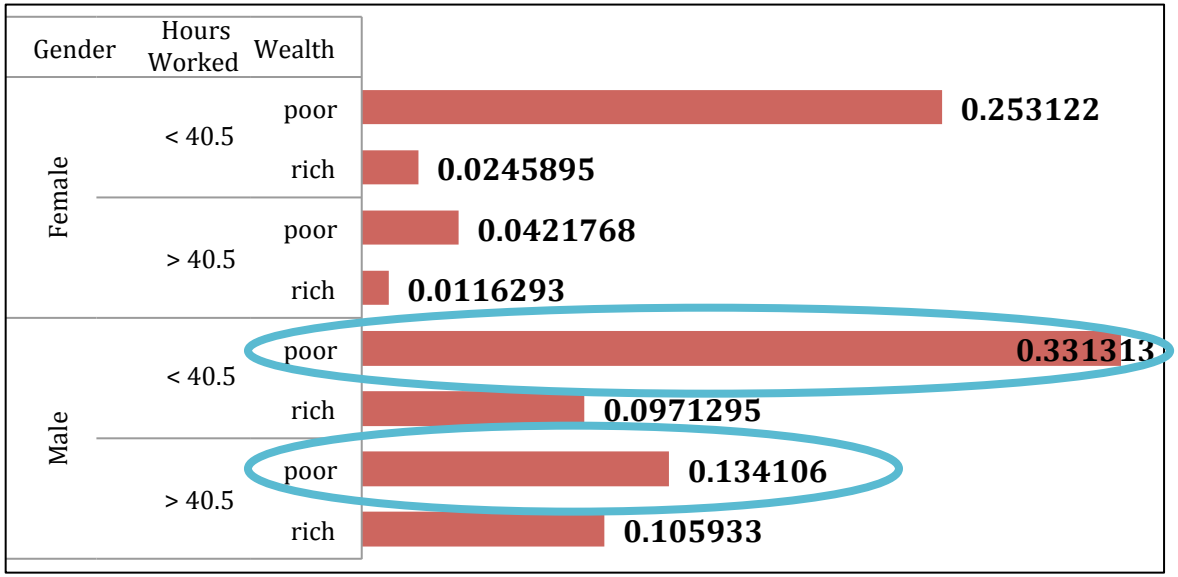
- Once you have the JPD you can calculate the probability of **any logical expression** involving your variables



$$\mathrm{P}(E) = \sum_{\text{rows matching } E} \mathrm{P}(row)$$

Using the joint distribution

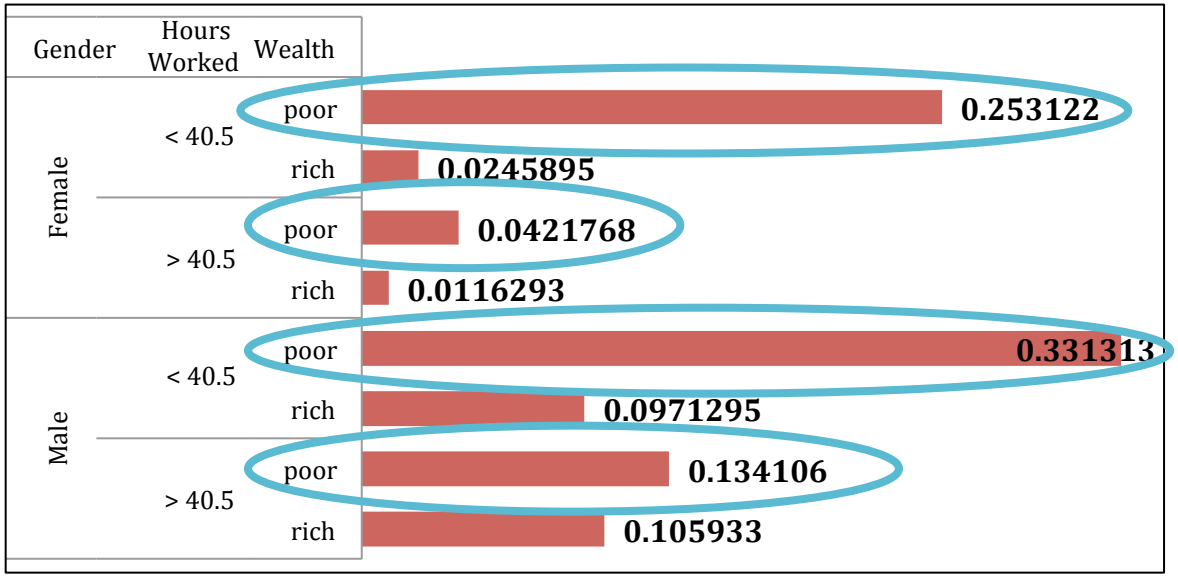
$$\mathrm{P}(\textit{Poor Male}) = 0.3313 + 0.1341 = 0.4654$$



$$\mathrm{P}(E) = \sum_{\text{rows matching } E} \mathrm{P}(row)$$

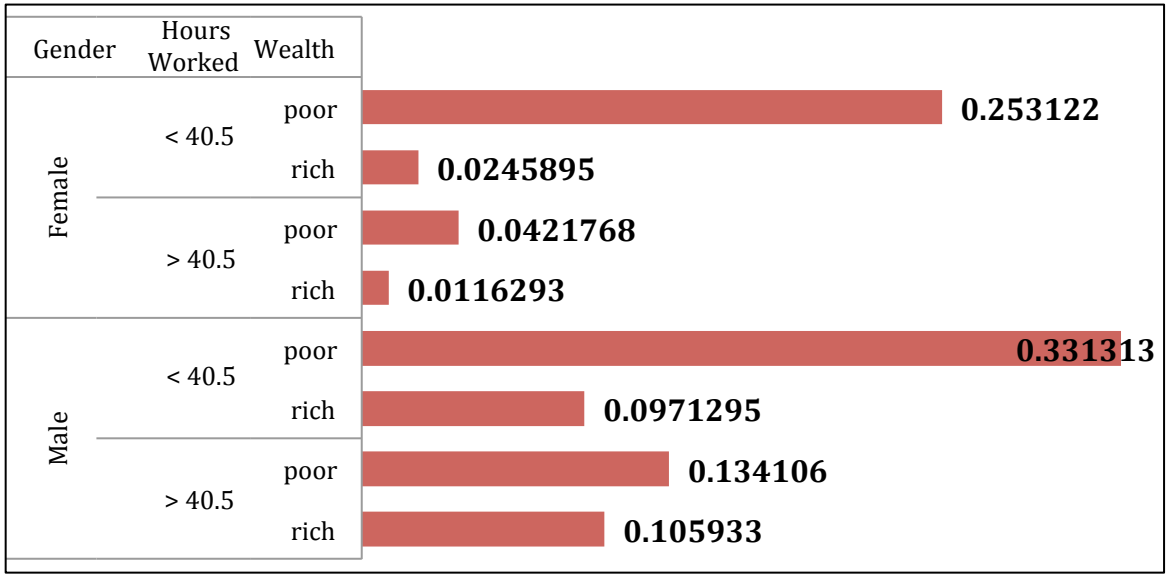
Using the joint distribution

$$\mathrm{P}(Poor)$$
$$= 0.2531 + 0.0421 + 0.3313 + 0.1341$$
$$= 0.7604$$



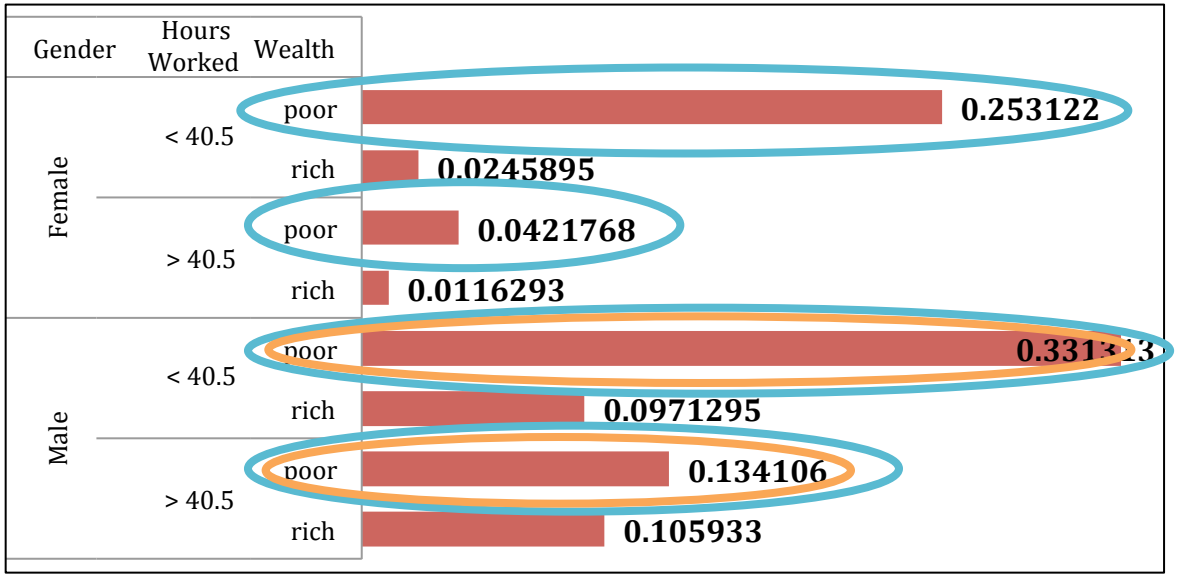
$$\mathrm{P}(E) = \sum_{\text{rows matching } E} \mathrm{P}(row)$$

Inference with the joint distribution



$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E2)} = \frac{\sum_{\text{rows matching } E1 \text{ and } E2} P(row)}{\sum_{\text{rows matching } E2} P(row)}$$

Inference with the joint distribution



$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E2)} = \frac{\sum_{\text{rows matching } E1 \text{ and } E2} P(row)}{\sum_{\text{rows matching } E2} P(row)}$$

P(*Male* | *Poor*) = 0.4654 / 0.7604 = 0.612

The good and the bad with JPDs

- **Good news ☺**
  Once you have a joint distribution, you can ask important questions about stuff that involves a lot of uncertainty!

- **Bad news ☹**
  Impossible to create for more than about 10 variables because there are so many numbers needed when you build the thing!

Solution—using fewer variables

- What **assumption** can we make to reduce the size of the problem?

- Suppose two events:
  A = Amy teaches class (otherwise it is cancelled!)
  S = it is sunny
- The joint pdf will have 4 entries

- To build the joint distribution we have to invent those 4 numbers….**OR WILL WE**?
  - We don't have to specify bottom level conjunctive events—i.e., P($\neg A \wedge S$) —IF…
  - …instead may be more convenient to specify things like P(A) or P(S)
- But just these marginal probs do not derive the joint distribution so cannot answer all questions

Independence

- The sunshine levels do not depend on and do not influence who is teaching

- This can be specified very simply:

$$P(S \mid A) = P(S)$$

- **This is a very powerful statement!**

- Requires **domain knowledge** and an understanding of causation, not just numerical probabilities

General definitions of independence

- From $P(S \mid A) = P(S)$, the axioms of probability imply:

1. $P(\neg S \mid A) = P(\neg S)$
2. $P(A \mid S) = P(A)$
3. $P(A \wedge S) = P(A)\,P(S)$
4. $P(\neg A \wedge S) = P(\neg A)\,P(S)$
5. $P(A \wedge \neg S) = P(A)\,P(\neg S)$
6. $P(\neg A \wedge \neg S) = P(\neg A)\,P(\neg S)$

- In **general** $P(A = u \wedge S = v) = P(A = u)\,P(S = v)$ for each combination of u = *True/False* and v = *True/False*

Joint distribution assuming independence

- We know:

  $P(A) = 0.6$

  $P(S) = 0.3$

  $P(S \mid A) = P(S)$

- From these statements we can derive the full joint distribution (**assuming independence**)

A	*S*	Probability
T	T	0.18
T	F	0.42
F	T	0.12
F	F	0.28

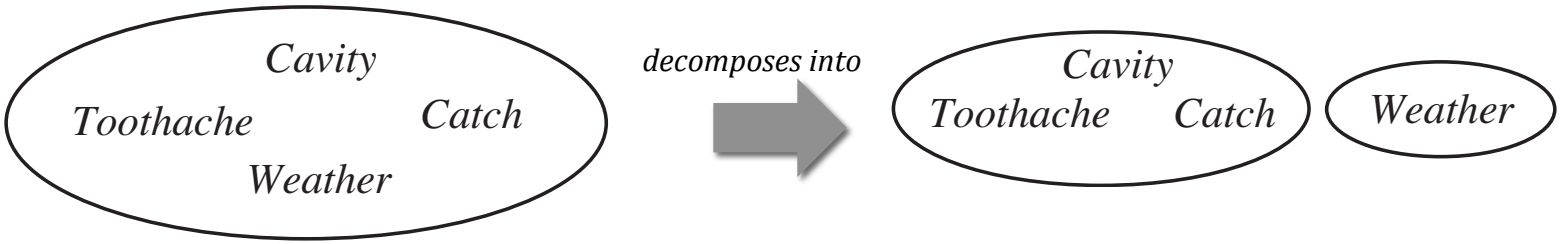
- Since we have the joint pdf, we can make any query!

Reduction in the joint pdf representation

- A and B are independent iff
  $\mathbf{P}(A \mid B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A,B) = \mathbf{P}(A)\mathbf{P}(B)$

  - E.g., $\mathbf{P}(Toothache, Catch, Cavity, Weather)$ has $2 \times 2 \times 2 \times 4$ entries in JPD
    $= \mathbf{P}(Toothache, Catch, Cavity)\mathbf{P}(Weather)$ has $2 \times 2 \times 2 + 4$ entries

    32 entries reduced to 12!



- **Absolute independence** powerful, but rare
  - In general, reduces exponential to linear complexity

Conditional probability review

- **Conditional** or **posterior** probabilities—based on known information

- Definition of **conditional probability**:
  $P(a \mid b) = \frac{P(a \wedge b)}{P(b)}$ if $P(b) > 0$

- Corollary: the **chain rule**
  $P(a \wedge b) = P(a \mid b)\, P(b)$
- Extend to $P(a \wedge b \wedge c \wedge \ldots) = ?$

Chain rule follows from this definition

- **Product rule** gives an alternative formulation:
  $P(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a)$

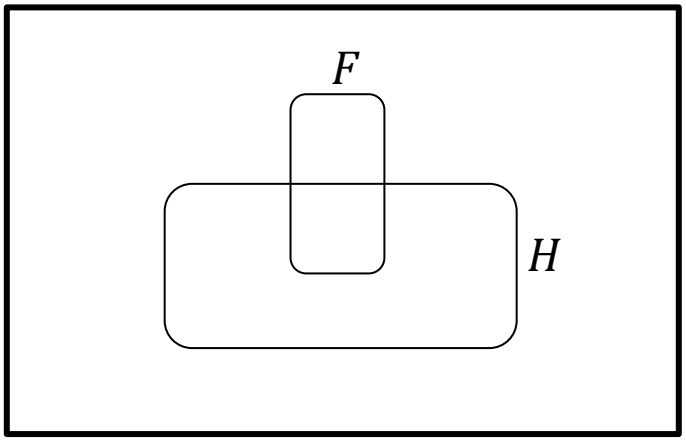
- **Chain rule** is derived by successive application of product rule

$$\begin{aligned}\mathbf{P}(X_1,\ldots,X_n) &= \mathbf{P}(X_1,\ldots,X_{n-1})\mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\ &= \mathbf{P}(X_1,\ldots,X_{n-2})\mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\ \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1}) \\ &= \ldots \\ &= \mathbf{P}(X_1)\mathbf{P}(X_2 \mid X_1)\mathbf{P}(X_3 \mid X_1, X_2)\ \ldots\ \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})\end{aligned}$$

OR $\Pi_{i = 1 \text{ to } n}\, \mathbf{P}(X_i \mid X_1,\ldots,X_{i-1})$

Conditional probability example

- P(A | B) = Fraction of worlds in which B is true that also have A true



H = Have a headache

F = Coming down with flu

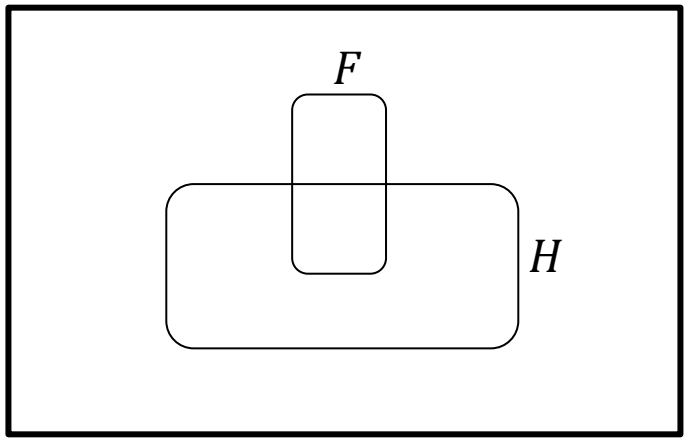
P(H) = 0.1

P(F) = 0.025

P(H | F) = 0.5

Headaches are rare and flu is rare, but if you are coming down with flu there is a 50-50 chance you'll have a headache.

Conditional probability example



H = Have a headache
F = Coming down with flu

$P(H) = 0.1$
$P(F) = 0.025$
$P(H \mid F) = 0.5$

$P(H \mid F)$ = **Fraction of flu-infected worlds in which you have a headache**

$$= \frac{\text{\# worlds with flu and headache}}{\text{\# worlds with flu}}$$

$$= \frac{\text{Area of “}H\text{ and }F\text{” region}}{\text{Area of “}F\text{” region}}$$

$$= \frac{P(H \wedge F)}{P(F)}$$

Conditional probability exercise

	Likes Football	Dislikes Football	Neutral
Male	0.25	0.1	0.15
Female	0.1	0.3	0.1

- Calculate:

  P(*Likes Football* | *Male*) = ?

  P(¬*Likes* | Female) = ?

Conditional probability exercise

	Likes Football	Dislikes Football	Neutral
Male	0.25	0.1	0.15
Female	0.1	0.3	0.1

- Calculate:

  P(*Likes Football* | *Male*) = **0.5**

  P(¬*Likes* | Female) = **0.8**

Conditional probability-based AI models

- Given a set of random variables $\boldsymbol{X}$ we are interested in the posterior joint distribution of the **query variables** $\boldsymbol{Y}$ given specific values $\boldsymbol{e}$ for the **evidence variables** $\boldsymbol{E}$

$$\mathbf{P}(\boldsymbol{Y} \mid E = e) = \alpha \mathbf{P}(\boldsymbol{Y} \wedge E = e)$$

  - Might have some hidden variables $\boldsymbol{H = X - Y - E}$ giving

$$\mathbf{P}(\boldsymbol{Y} \mid E = e) = \alpha \Sigma_h \mathbf{P}(\boldsymbol{Y} \wedge E = e \wedge H = h)$$

- **Note: what is α?**
- From the definition $\mathrm{P}(A \mid B) = \dfrac{\mathrm{P}(A \wedge B)}{\mathrm{P}(B)}$

  α is the denominator $1 / \mathbf{P}(\boldsymbol{E} = \boldsymbol{e})$. $\mathbf{P}(\boldsymbol{E} = \boldsymbol{e})$ can be computed from joint distribution as $\Sigma_h \mathbf{P}(\boldsymbol{E} = \boldsymbol{e} \wedge \boldsymbol{H} = \boldsymbol{h})$

Example (medical diagnosis)

- **Causal model:**
  $D \rightarrow I \rightarrow S \quad (Y \rightarrow H \rightarrow E)$

  Cancer → anemia → fatigue
  Kidney → disease anemia → fatigue

$$P(Y = cancer \mid E = fatigue) =$$
$$\alpha \, [\, P(Y = cancer \wedge E = fatigue \wedge anemia) +$$
$$P(Y = cancer \wedge E = fatigue \wedge \neg anemia) \,]$$

$$\alpha = 1/P(E = fatigue) \text{ or } 1/[P(E = fatigue \wedge anemia) +$$
$$P(E = fatigue \wedge \neg anemia) \,]$$

Conditional probability analysis

- $\mathbf{P}(\boldsymbol{Y} \mid E = e) = \alpha\mathbf{P}(\boldsymbol{Y} \wedge E = e) = \alpha\Sigma_h\mathbf{P}(\boldsymbol{Y} \wedge E = e \wedge H = h)$

- Terms in the summation are **joint distribution entries** because $\boldsymbol{Y}$, $\boldsymbol{E}$, and $\boldsymbol{H}$ together exhaust the full set of random variables

- **Obvious problems**
  1. Time and space complexity $O(d^n)$ where d is the largest arity
  2. Where do we get the numbers to solve real world problems?

- Independence assumption provides a solution to 1…

…but where do we find the numbers?

- Assuming independence, doctors may be able to **estimate** P(*symptom* | *disease*) for each *S/D* pair (causal reasoning)

- Hard to estimate what we really **need** to know: P(*disease* | *symptom*)

- This is why **Bayes rule** is so important in probabilistic AI!

Bayes Rule

- Product rule $P(a \wedge b) = P(a \mid b)P(b) = P(b \mid a)P(a)$

  $\Rightarrow$ **Bayes rule:** $P(a \mid b) = \dfrac{P(b \mid a)P(a)}{P(b)}$

- Or in distribution form:
  $\mathbf{P}(Y \mid X) = \dfrac{\mathbf{P}(X \mid Y)\mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\mathbf{P}(X \mid Y)\,\mathbf{P}(Y)$

- Useful for assessing **diagnostic** probability from **causal** probability:
  - $P(\text{Cause} \mid \text{Effect}) = \dfrac{P(\text{Effect} \mid \text{Cause})\,P(\text{Cause})}{P(\text{Effect})}$
  - E.g., Let M be meningitis, S be stiff neck
    $P(M \mid S) = P(S \mid M)\,P(M) / P(S) = 0.8 \times 0.0001 / 0.1 = 0.0008$
  - Note: posterior probability of meningitis is still very small!

More general forms of Bayes rule

- $$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B \mid A)P(A) + P(B \mid \neg A)P(\neg A)}$$

- $$P(A \mid B) = \frac{P(B \mid A, e)P(A \mid e)}{P(B \mid e)}$$

- $$P(A = v_i \mid B) = \frac{P(B \mid A = v_i)P(A = v_i)}{\sum_{k=1}^{n} P(B \mid A = v_k)P(A = v_k)}$$

Conditional independence

- **P**(*Toothache, Cavity, Catch*) has $2^3 - 1 = 7$ independent entries
  - If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
    - **P**(*catch* | *toothache*, *cavity*) = **P**(*catch* | *cavity*)
  - The same independence holds if I haven't got a cavity
    - **P**(*catch* | *toothache*, ¬*cavity*) = **P**(*catch* | ¬*cavity*)
- Catch is **conditionally independent** of Toothache given Cavity: **P**(*Catch* | *Toothache, Cavity*) = **P**(*Catch* | *Cavity*)
  - Equivalent statements (from original definitions of independence)
    - **P**(*Toothache* | *Catch, Cavity*) = **P**(*Toothache* | *Cavity*)
    - **P**(*Toothache, Catch* | *Cavity*) = **P**(*Toothache* | *Cavity*)**P**(*Catch* | *Cavity*)

Conditional independence (cont.)

- Write out full joint distribution using chain rule:

$$\begin{aligned}&\mathbf{P}(Toothache, Catch, Cavity)\\ &\quad = \mathbf{P}(Toothache \mid Catch, Cavity)\mathbf{P}(Catch, Cavity)\\ &\quad = \mathbf{P}(Toothache \mid Catch, Cavity)\mathbf{P}(Catch \mid Cavity)\mathbf{P}(Cavity)\\ &\quad = \mathbf{P}(Toothache \mid Cavity)\mathbf{P}(Catch \mid Cavity)\mathbf{P}(Cavity)\end{aligned}$$

i.e., 2 + 2+ 1 = 5 independent numbers

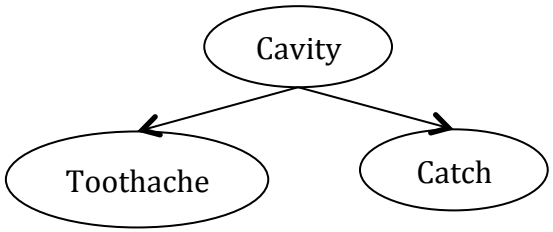
- In most cases, conditional independence reduces size of representation from **exponential** in n to **linear** in n
- Conditional independence is most basic and robust form of knowledge about uncertain environments

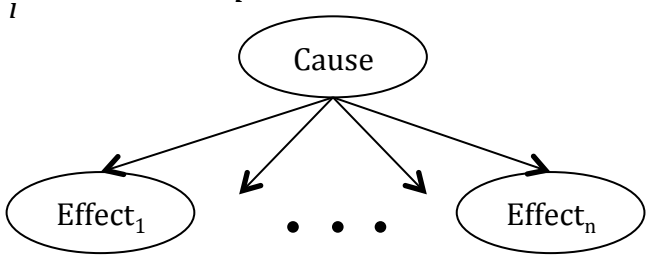
Bayes rule and conditional independence

- $\mathbf{P}(Cavity \mid toothache\ catch)$
  $= \mathbf{P}(toothache\ catch \mid Cavity)\mathbf{P}(Cavity)$
  $= \mathbf{P}(toothache \mid Cavity)\mathbf{P}(catch \mid Cavity)\mathbf{P}(Cavity)$

- We say: "toothache and catch are independent, given cavity"
  - Cavity **separates** Toothache and Catch because it is a direct cause of both
  - Example of a **naïve Bayes** model

- $\mathrm{P}(Cause, Effect_1, \ldots, Effect_n) = \mathrm{P}(Cause)\prod_i \mathrm{P}(Effect_i \mid Cause)$





- Total number of parameters is **linear** in n (number of systems)
- This is our first Bayesian inference net!

Revisiting our joint distribution example

- A: Amy teaches the class
- S: It is sunny
- L: The lecturer arrives *slightly* late

- Assume lecturers are sometimes delayed by bad weather

- Start by writing down knowledge we're happy about:
  - $P(S \mid A) = P(S)$, $P(S) = 0.3$, $P(A) = 0.6$
  - **Lateness is not independent of the weather and is not independent of the lecturer**
  - **Weather and lecturer are independent**

Reduce complexity with independence

- A: Amy teaches the class
- S: It is sunny
- L: The lecturer arrives *slightly* late

- Assume lecturers are sometimes delayed by bad weather

- Start by writing down knowledge we're happy about:
  - $P(S \mid A) = P(S)$, $P(S) = 0.3$, $P(A) = 0.6$
  - **Lateness is not independent of the weather and is not independent of the lecturer**
  - **Weather and lecturer are independent**

- Know the joint probability of S and A, so now need:
  - $\mathbf{P}(L \mid S, A)$ for the 4 cases where S and A are true/false

Reduce complexity with independence

- A: Amy teaches the class
- S: It is sunny
- L: The lecturer arrives *slightly* late

- Assume lecturers are sometimes delayed by bad weather

$$P(S \mid A) = P(S)$$
$$P(S) = 0.3$$
$$P(A) = 0.6$$

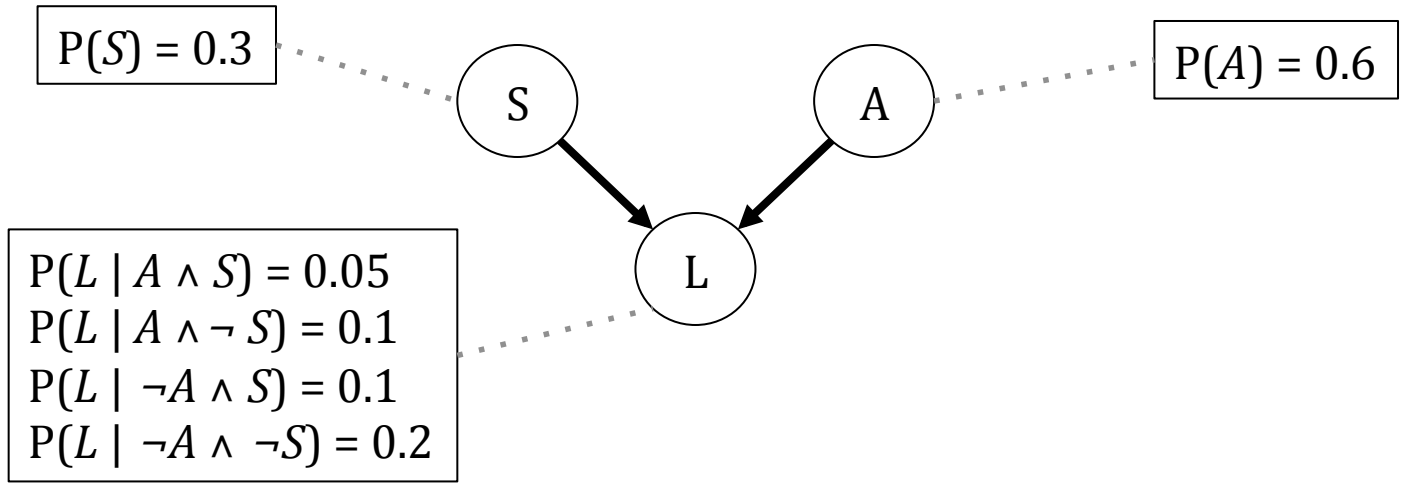
$$P(L \mid A \wedge S) = 0.05$$
$$P(L \mid A \wedge \neg S) = 0.1$$
$$P(L \mid \neg A \wedge S) = 0.1$$
$$P(L \mid \neg A \wedge \neg S) = 0.2$$

- Now we can **derive** a full joint distribution with a "mere" six numbers instead of seven
  - NOTE: Savings are larger for larger numbers of variables

Graphical representation of independence

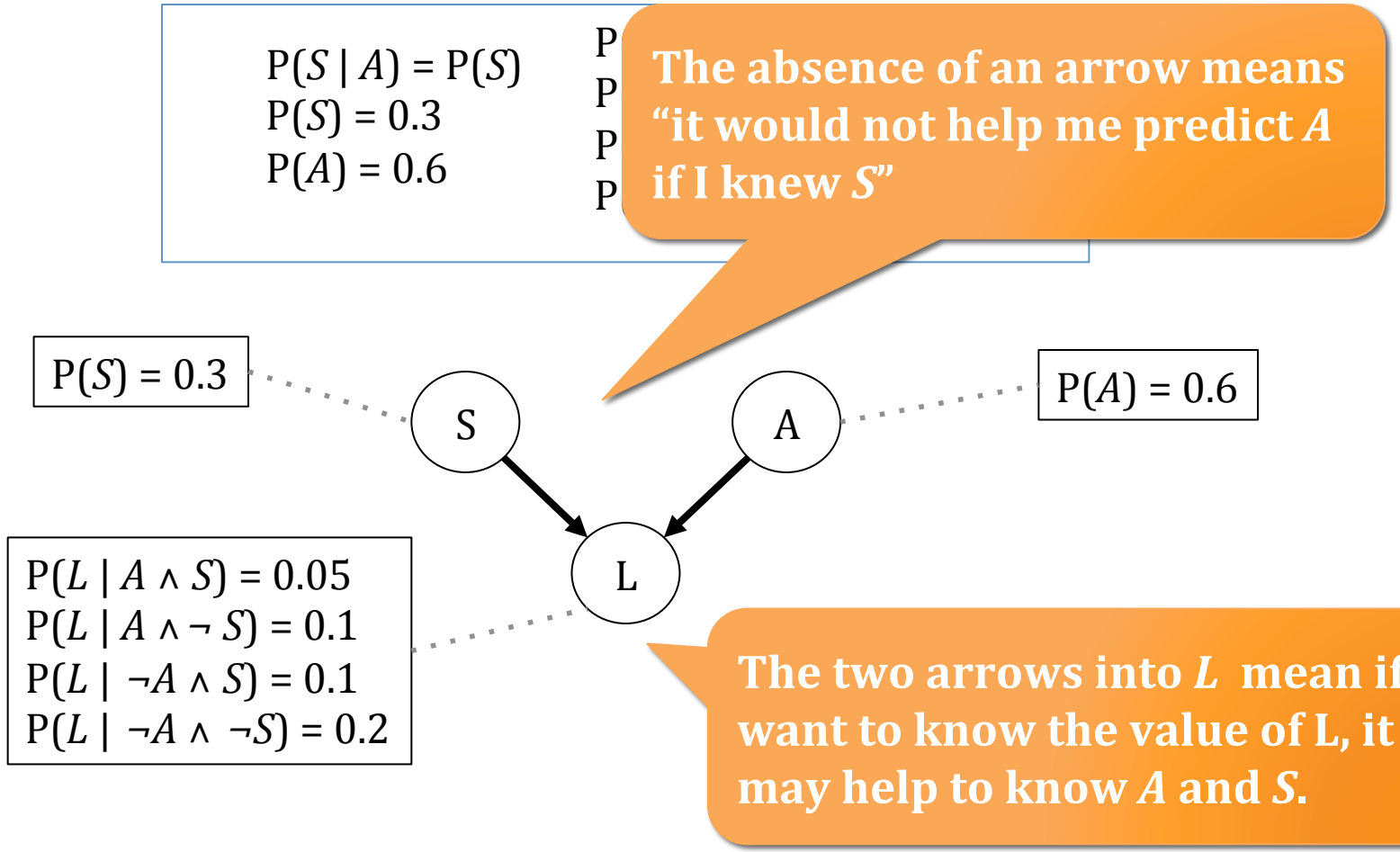
$P(S \mid A) = P(S)$
$P(S) = 0.3$
$P(A) = 0.6$

$P(L \mid A \wedge S) = 0.05$
$P(L \mid A \wedge \neg S) = 0.1$
$P(L \mid \neg A \wedge S) = 0.1$
$P(L \mid \neg A \wedge \neg S) = 0.2$



- This is our second Bayesian inference net!

Graphical representation of independence



- This is our second Bayesian inference net!

Example of conditional independence

- A: Amy teaches the class
- L: The lecturer arrives *slightly* late
- P: The lecture concerns probability

- Assume:
  - Oliver has a higher chance of being late than Amy
  - Amy has a higher chance of giving probability lectures.

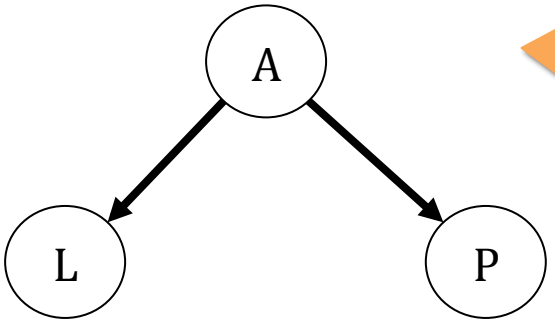
- **Now what kind of independence can we find?**
- How about…
  - $P(L \mid A) = P(L)$?
  - $P(P \mid A) = P(P)$?
  - $P(L \mid P) = P(L)$?

Example of conditional independence

- Once you know who the lecturer is, then whether they arrive late does not affect whether the lecture concerns probability:

$$P(P \mid A,L) = P(P \mid A)$$
$$P(P \mid \neg A,L) = P(P \mid \neg A)$$

- P and L are **conditionally independent** given A

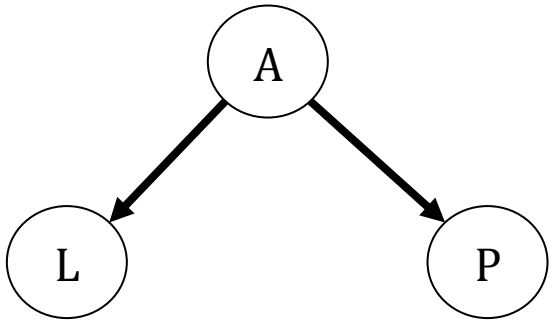


Given knowledge of *A*, knowing anything else in the diagram won't help us with *L*

Conditional independence reduces complexity

- We can write down P(A)—already know this
- Since we know L is only directly influenced by A, we can write down values of P(L | A) and P(L | ¬A) to fully specify L's behavior.
- Same goes for P.

P(A) = 0.6
P(L | A) = 0.085
P(L | ¬A) = 0.17
P(P | A) = 0.6
P(P | ¬A) = 0.3



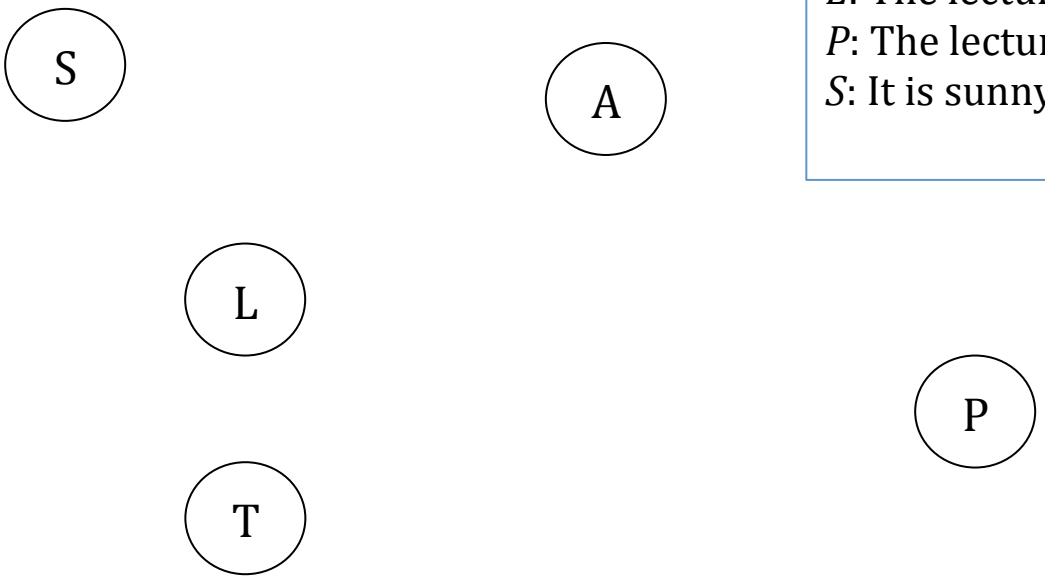
- Again, we can obtain any member of the **full joint distribution**!

How to construct a Bayes net

- Expand the example to include 5 variables:
  - *T*: The lecture started by 6:05
  - *A*: Amy teaches the class
  - *L*: The lecturer arrives *slightly* late
  - *P*: The lecture concerns probability
  - *S*: It is sunny

- **_T_ only directly influenced by _L_ (conditionally independent of _A_, _P_, and _S_ given _L_)**
- **_L_ only directly influenced by _A_ and _S_ (conditionally independent of _P_ given _A_ and _S_)**
- **_P_ only directly influenced by _A_ (conditionally independent of _T_, _L_, and _S_ given _A_)**
- **_A_ and _S_ are independent**

Making a Bayes net

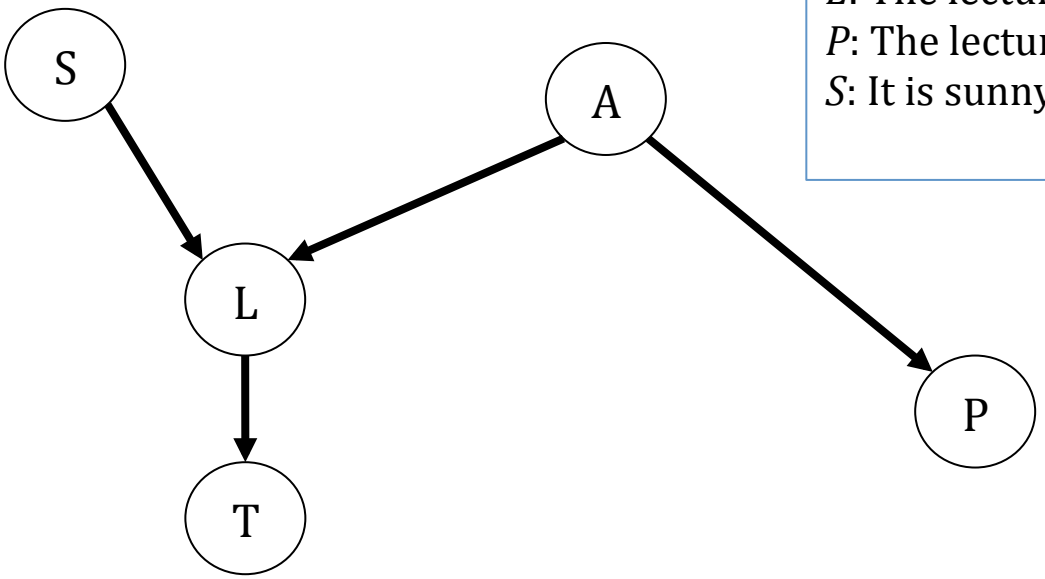


T: The lecture started by 6:05
A: Amy teaches the class
L: The lecturer arrives *slightly* late
P: The lecture concerns probability
S: It is sunny

- **Step 1: add variables**
  - Choose the variables you want to include in the net

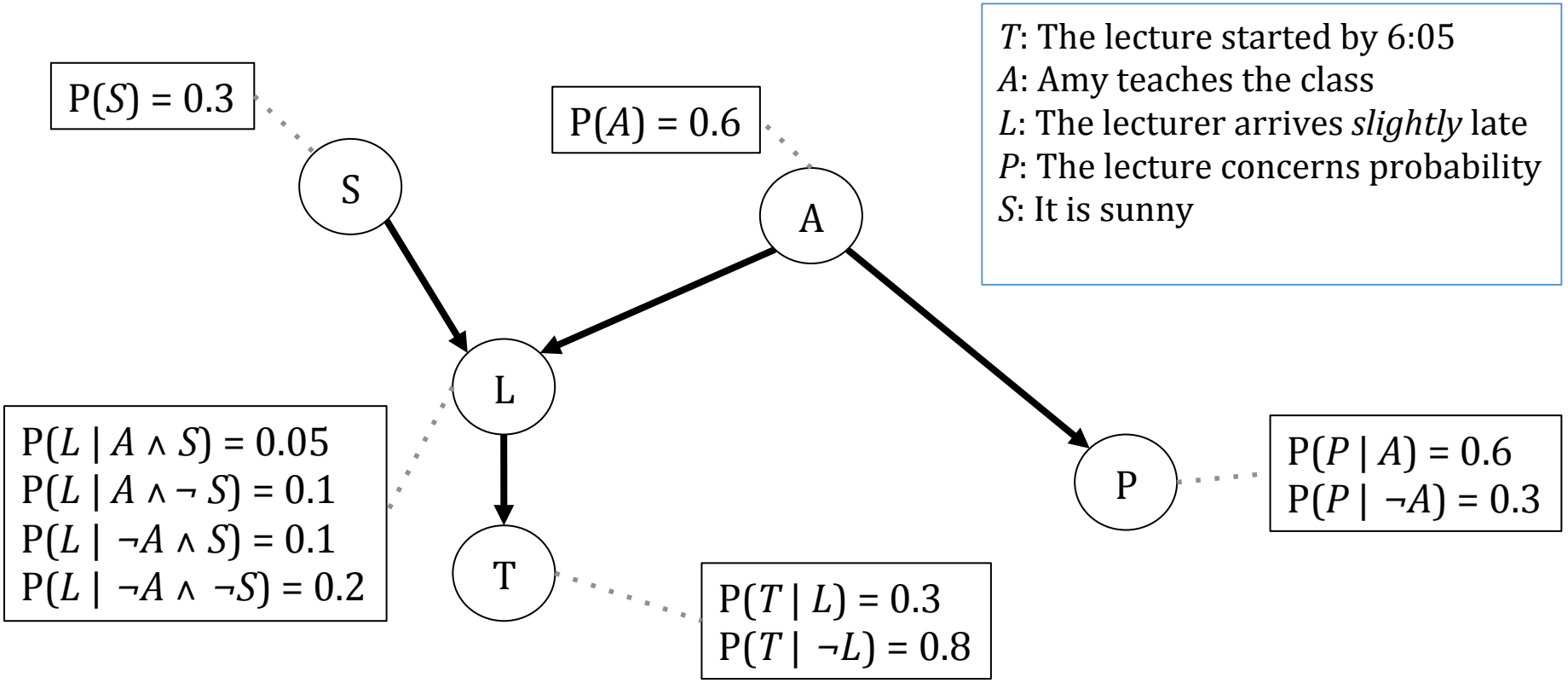
Making a Bayes net

T: The lecture started by 6:05
A: Amy teaches the class
L: The lecturer arrives *slightly* late
P: The lecture concerns probability
S: It is sunny



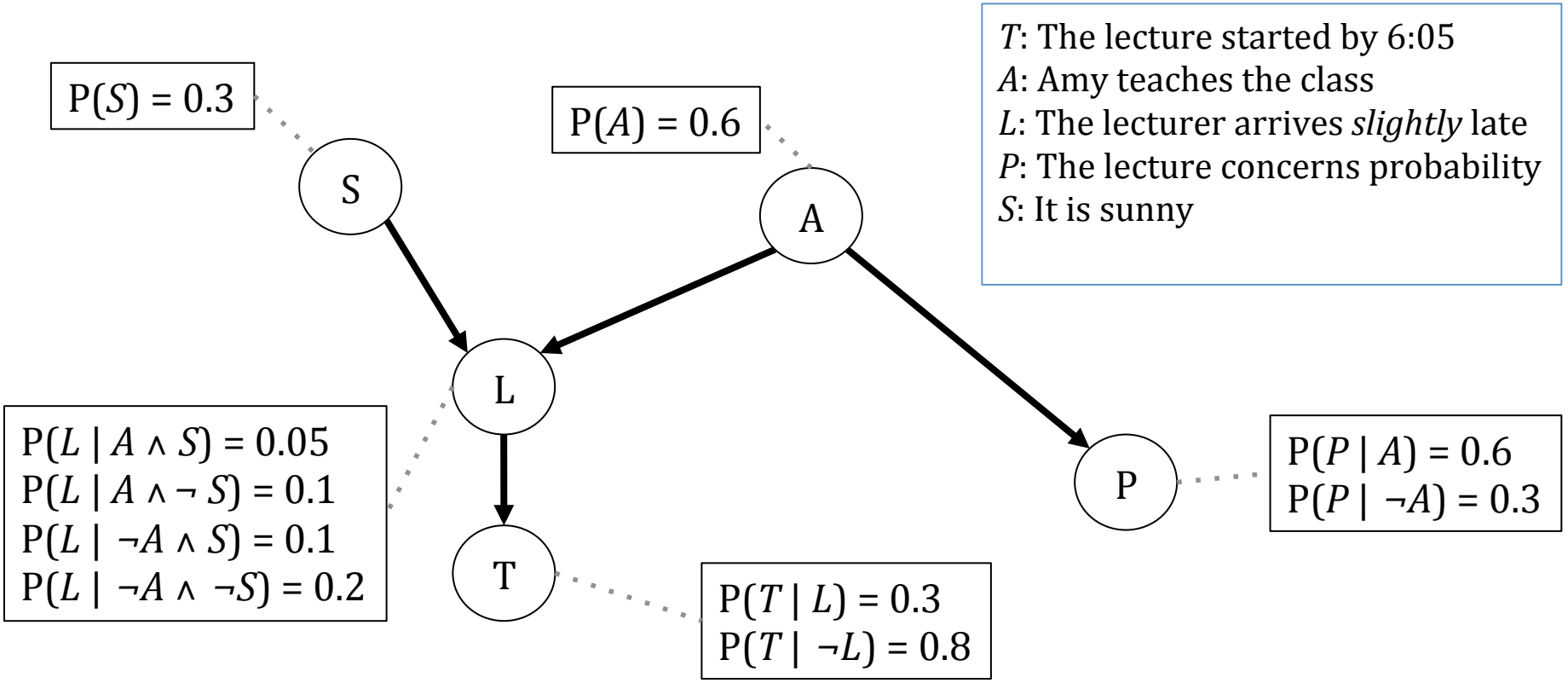
- **Step 2: add links**
  - Link structure must be acyclic
  - If node X is given parents $\{Q_1, Q_2, \ldots, Q_n\}$, then any non-descendant of X is conditionally independent of X given $\{Q_1, Q_2, \ldots, Q_n\}$

Making a Bayes net



- **Step 3: add a probability table for each node**
  - The table for node *X* must list P(*X* | Parent values) for each possible combination of parent values

Making a Bayes net



- **Two unconnected variables may still be correlated**
- Each node is conditionally independent of all non-descendants in the tree **given its parents**
- We can deduce many other conditional independence relations from a Bayes net

Bayes nets formalized

- Bayes net (or **belief network**)—augmented directed, acyclic graph represented by the pair V, E where:
  - V is a set of vertices
  - E is a set of directed edges (no loops of any length are allowed!)

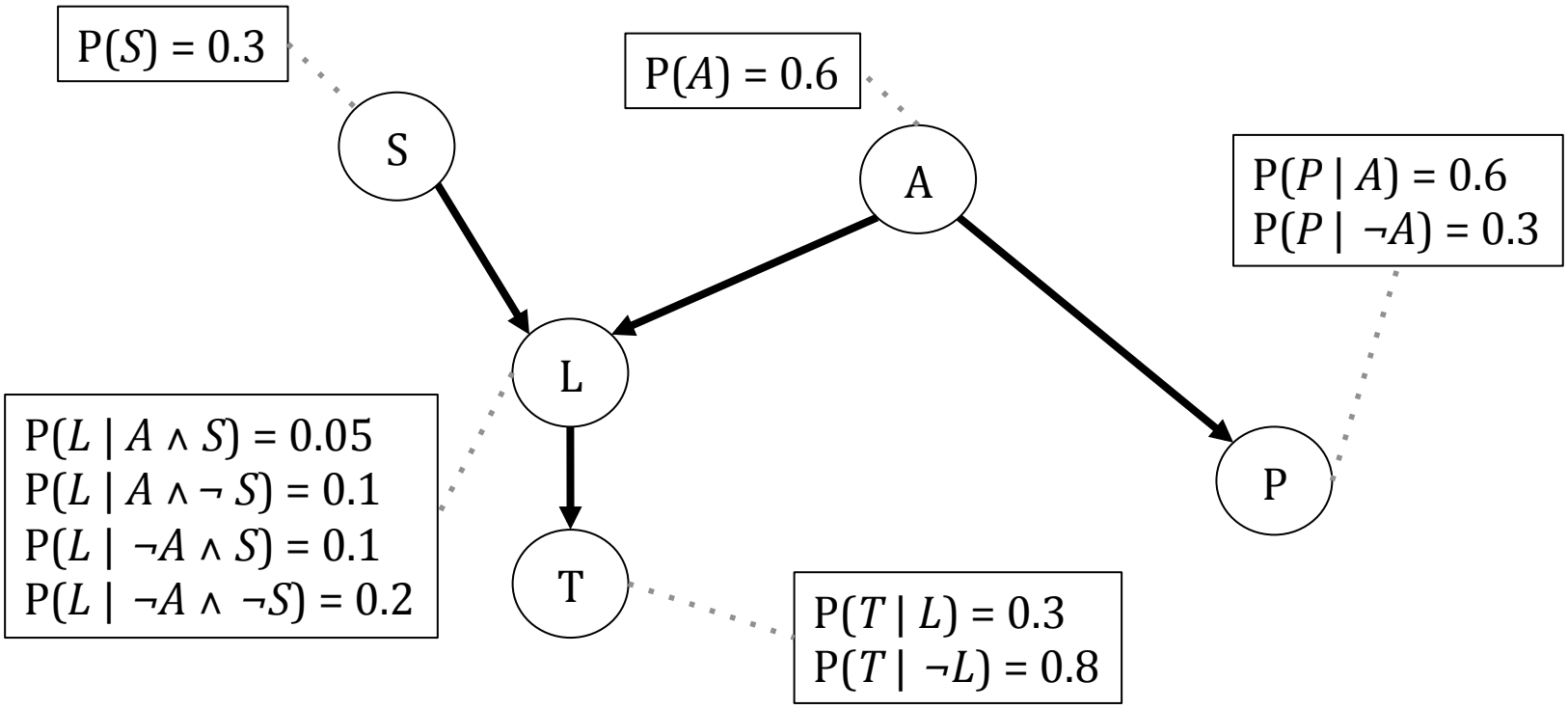
- Each vertex contains the following information
  - Name of a random variable
  - **Probability distribution table** indicating how the probability of this variable's values depends on all possible combinations of parental values

Formal procedure for building a Bayes net

1. Choose a set of relevant variables
2. Choose an ordering for them (often requires domain knowledge)
   - Assume the variables are X_1, …, X_m where X_1 is the first in the ordering, X_2 is the second, etc.
3. For i = 1 to m:
   a) Add the node X_i to the network
   b) Set *Parents*(X_i) to be the **minimal subset** {X_1, …, X_{i-1}} s.t. we have conditional independence of X_i and all other members of {X_1, …, X_{i-1}} given *Parents*(X_i)
   c) Define the conditional probability table of P(X_i | Assignments of *Parents*(X_i))

Calculating using Bayes nets

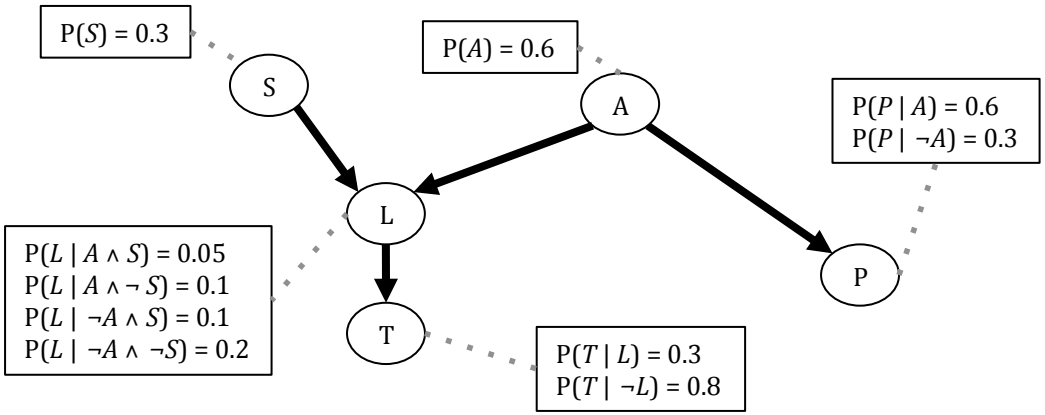
- How to compute an entry in the joint distribution?
  E.g., what is P($S \wedge \neg A \wedge L \wedge \neg P \wedge T$)?



Calculating using Bayes nets

- Use the chain rule and conditional independence!

$$
\begin{aligned}
&\mathrm{P}(S \wedge \neg A \wedge L \wedge \neg P \wedge T) \\
&\quad = \mathrm{P}(T \mid \neg P \wedge L \wedge \neg A \wedge S)\, \mathrm{P}(\neg P \wedge L \wedge \neg A \wedge S) \\
&\quad = \mathrm{P}(\mathrm{T} \mid \mathrm{L})\, \mathrm{P}(\neg P \wedge L \wedge \neg A \wedge S) \\
&\quad = \mathrm{P}(\mathrm{T} \mid \mathrm{L})\, \mathrm{P}(\neg P \mid L \wedge \neg A \wedge S)\, \mathrm{P}(L \wedge \neg A \wedge S) \\
&\quad = \mathrm{P}(\mathrm{T} \mid \mathrm{L})\, \mathrm{P}(\neg P \mid \neg A)\, \mathrm{P}(L \wedge \neg A \wedge S) \\
&\quad = \mathrm{P}(\mathrm{T} \mid \mathrm{L})\, \mathrm{P}(\neg P \mid \neg A)\, \mathrm{P}(L \mid \neg A \wedge S)\, \mathrm{P}(\neg A \wedge S) \\
&\quad = \mathrm{P}(\mathrm{T} \mid \mathrm{L})\, \mathrm{P}(\neg P \mid \neg A)\, \mathrm{P}(L \mid \neg A \wedge S)\, \mathrm{P}(\neg A \mid S)\, \mathrm{P}(S) \\
&\quad = \mathrm{P}(\mathrm{T} \mid \mathrm{L})\, \mathrm{P}(\neg P \mid \neg A)\, \mathrm{P}(L \mid \neg A \wedge S)\, \mathrm{P}(\neg A)\, \mathrm{P}(S)
\end{aligned}
$$



The general case

- $\mathbf{P}(X_1,\ldots,X_n) = \mathbf{P}(X_1,\ldots,X_{n-1})\mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$
  $= \mathbf{P}(X_1,\ldots,X_{n-2})\mathbf{P}(X_{n-1} \mid X_1,\ldots,X_{n-2})\ \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$
  $= \ldots$
  $= \mathbf{P}(X_1)\mathrm{P}(X_2 \mid X_1)\mathbf{P}(X_3 \mid X_1, X_2)\ \ldots\ \mathbf{P}(X_n \mid X_1,\ldots,X_{n-1})$

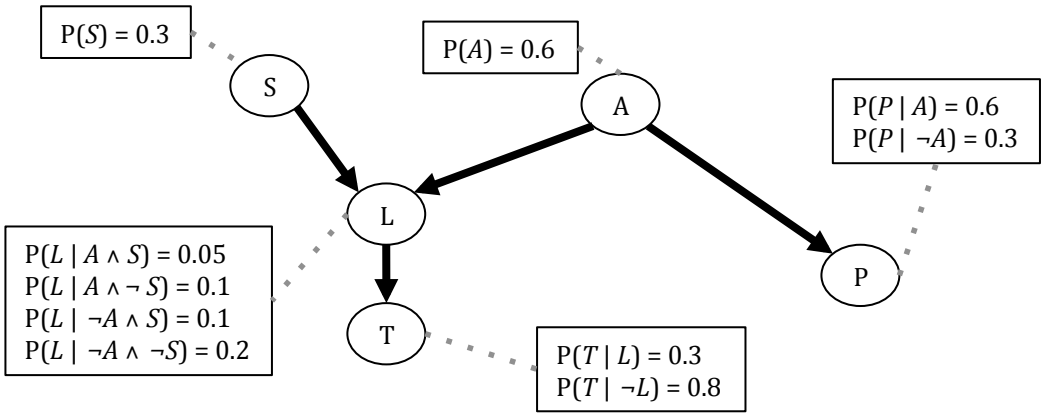
  $= \Pi_{i = 1 \text{ to } n}\, \mathbf{P}(X_i \mid X_1,\ldots,X_{i-1})$

  $= \Pi_{i = 1 \text{ to } n}\, \mathbf{P}(X_i \mid \textit{Assignment of Parents}(X_i))$

- Any entry in the joint distribution can be computed and **any conditional probability** can be computed

Where are we now?

- Have a methodology for building Bayes nets
- Reduced storage necessary for probability tables
  - No longer exponential (like for joint distribution), but only exponential in the maximum number of parents of any node
- Can compute probabilities of any assignment of values to variables
  - Linear time w.r.t. number of nodes
- Can use Bayes net to determine answer to any question!



- **What could we do to compute P($P \mid T, \neg S$)**

Example Bayes net query

- **P($P \mid T, \neg S$)**

- Step 1: compute P($P \wedge T \wedge \neg S$)
  - Sum of all rows in the joint distribution that match $P \wedge T \wedge \neg S$ (4 joint computes)
  - Use Bayes net computation method

P($P \wedge T \wedge \neg S$) = P($P \wedge \boldsymbol{A} \wedge T \wedge \boldsymbol{L} \wedge \neg S$)

= (P($P \mid A$) P(A) P($T \mid L$) P($L \mid A \wedge \neg S$) P($\neg S$)) +
(P($P \mid \neg A$) P($\neg A$) P($T \mid L$) P($L \mid \neg A \wedge \neg S$) P($\neg S$)) +
(P($P \mid A$) P(A) P($T \mid \neg L$) P($\neg L \mid A \wedge \neg S$) P($\neg S$)) +
(P($P \mid \neg A$) P($\neg A$) P($T \mid \neg L$) P($\neg L \mid \neg A \wedge \neg S$) P($\neg S$))

= 0.00756 + 0.00504 + 0.18144 + 0.05376

= 0.2478

Example Bayes net query

- **$\mathrm{P}(P \mid T, \neg S)$**

- Step 2: compute $\mathrm{P}(\neg P \wedge T \wedge \neg S)$
  - Sum of all rows in the joint distribution that match $\neg P \wedge T \wedge \neg S$ (4 joint computes)
  - Use Bayes net computation method

$\mathrm{P}(\neg P \wedge T \wedge \neg S) = (\neg P \wedge \boldsymbol{A} \wedge T \wedge \boldsymbol{L} \wedge \neg S)$

$= (\mathrm{P}(\neg P \mid A)\,\mathrm{P}(A)\,\mathrm{P}(T \mid L)\,\mathrm{P}(L \mid A \wedge \neg S)\,\mathrm{P}(\neg S)) +$
$(\mathrm{P}(\neg P \mid \neg A)\,\mathrm{P}(\neg A)\,\mathrm{P}(T \mid L)\,\mathrm{P}(L \mid \neg A \wedge \neg S)\,\mathrm{P}(\neg S)) +$
$(\mathrm{P}(\neg P \mid A)\,\mathrm{P}(A)\,\mathrm{P}(T \mid \neg L)\,\mathrm{P}(\neg L \mid A \wedge \neg S)\,\mathrm{P}(\neg S)) +$
$(\mathrm{P}(\neg P \mid \neg A)\,\mathrm{P}(\neg A)\,\mathrm{P}(T \mid \neg L)\,\mathrm{P}(\neg L \mid \neg A \wedge \neg S)\,\mathrm{P}(\neg S))$

$= 0.0054 + 0.01176 + 0.12096 + 0.12544$

$= 0.26356$

Example Bayes net query

- **$\mathbf{P}(P \mid T, \neg S)$**

- Step 3: Return $\mathrm{P}(P \mid T, \neg S) = \dfrac{\mathrm{P}(P \wedge T \wedge \neg S)}{\mathrm{P}(P \wedge T \wedge \neg S) + \mathrm{P}(\neg P \wedge T \wedge \neg S)}$

$$= \frac{0.2478}{0.2478 + 0.26356}$$

$$= 0.4846$$

- **Inference by enumeration** algorithm

The good news

- We can do inference!
  - We can compute any conditional probability
    P(Some variable | Some other variable values)

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E2)} = \frac{\sum_{\text{joint entries matching E1 and } E2} P(\textit{joint entry})}{\sum_{\text{joint entries matching } E2} P(\textit{joint entry})}$$

The good news

- We can do inference!
  - We can compute any conditional probability
    P(Some variable | Some other variable values)

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E2)} = \frac{\sum_{\text{joint entries matching E1 and } E2} P(\textit{joint entry})}{\sum_{\text{joint entries matching } E2} P(\textit{joint entry})}$$

- Suppose you have m binary-valued variables in your Bayes net and expression E_2 mentions k variables.
  - **How much work is the above computation?**

The sad, bad news

- Conditional probabilities by enumerating all matching entries in the joint are expensive
  - **Exponential in the number of variables**

- Perhaps there are faster ways of querying Bayes nets?
  - In fact, if you ever manually do a Bayes net inference, you'll find there are often many tricks to save you time
  - All we have to do is program our computer to do those tricks too, right?

- Sadder and worse news:
  - **General querying of Bayes nets is NP-complete**

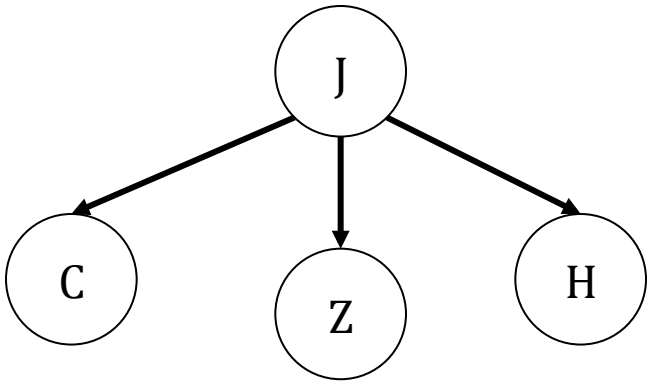
Naïve Bayes classifiers

- Our first example of machine learning!
- A supervised learning method
- Make (naïve) **independence assumption**
  - Can explore a simple subset of Bayesian nets s.t. it is easy to estimate the CPTs from sample data

- **Maximum likelihood estimation**
  - Given a set of correctly classified representative examples
  - **Q:** What estimates of conditional probabilities maximize the likelihood of the data that was observed?
  - **A:** The estimates that reflect the sample proportions

A simple Bayes net

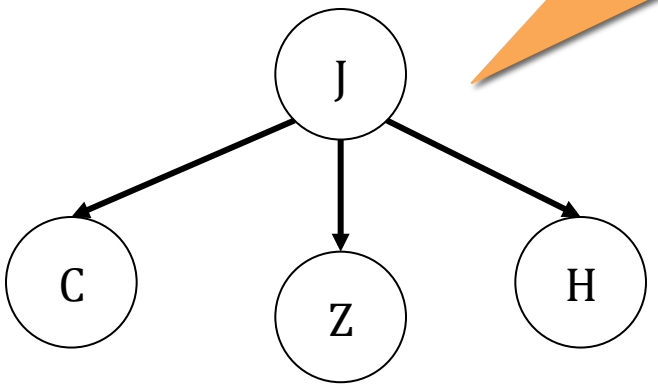
J: Person is a junior
C: Brought coat to classroom
Z: Lives in zip code 12345
H: Saw *Harry Potter* more than once



A simple Bayes net

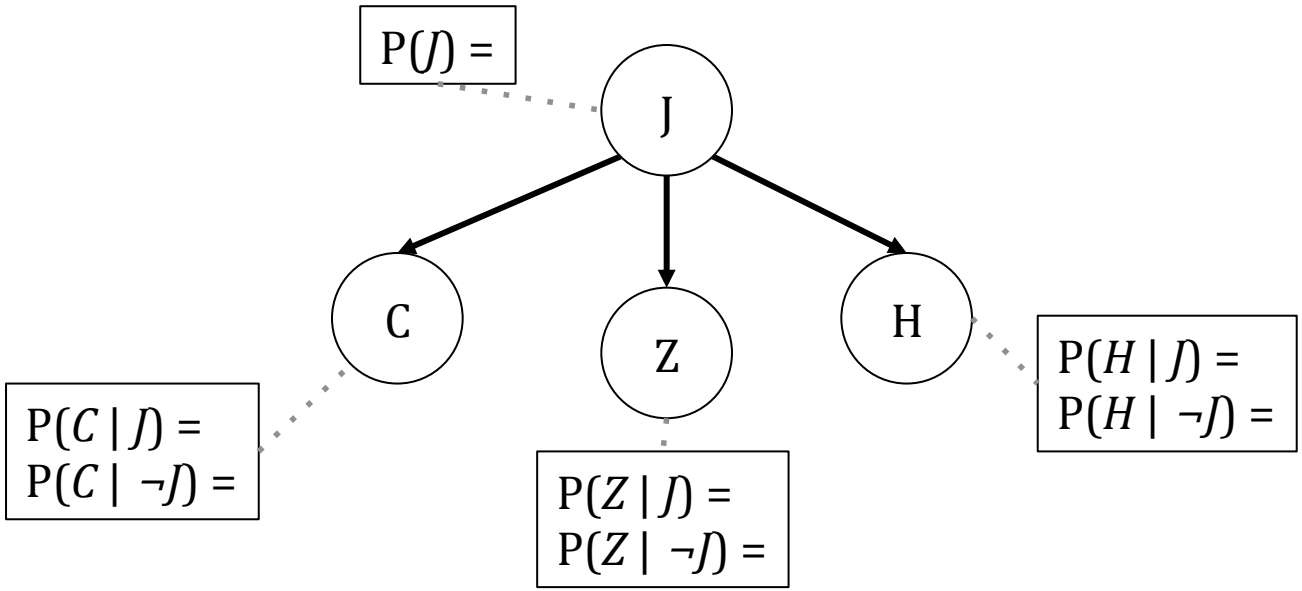
J: Person is a junior
C: Brought coat to cl
Z: Lives in zip code
H: Saw *Harry Potter*

What parameters are stored in the CPTs of this Bayes net?



A simple Bayes net

J: Person is a junior
C: Brought coat to classroom
Z: Lives in zip code 12345
H: Saw *Harry Potter* more than once



A simple Bayes net

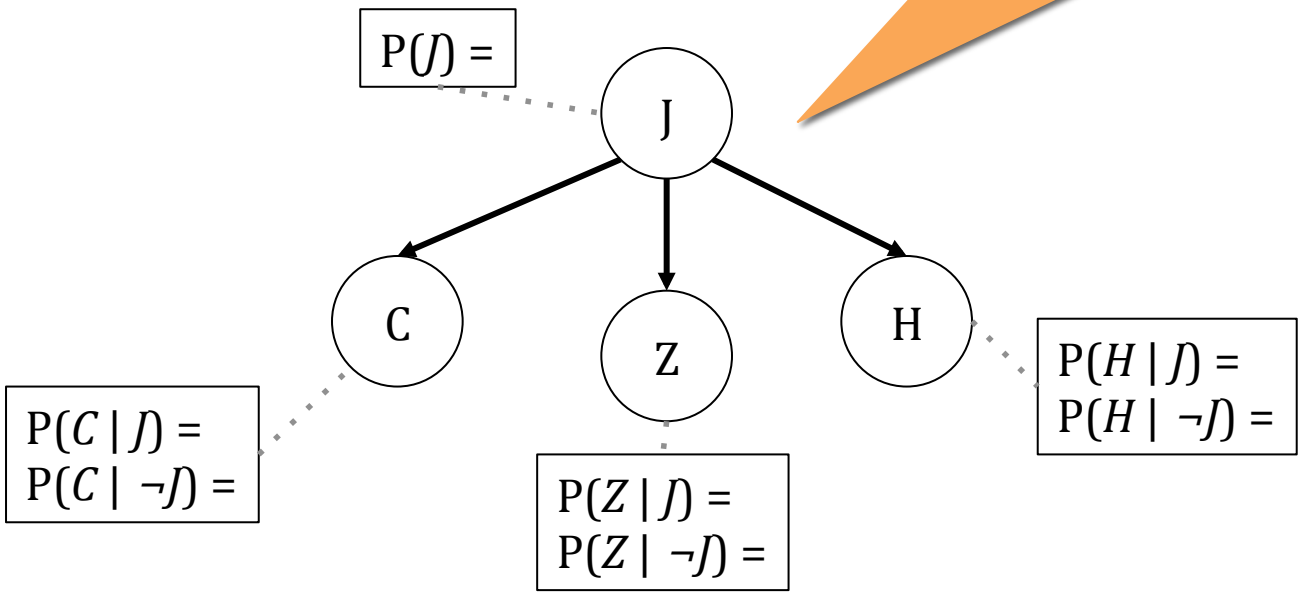
J: Person is a junior

C: Brought coat to cl

Z: Lives in zip code

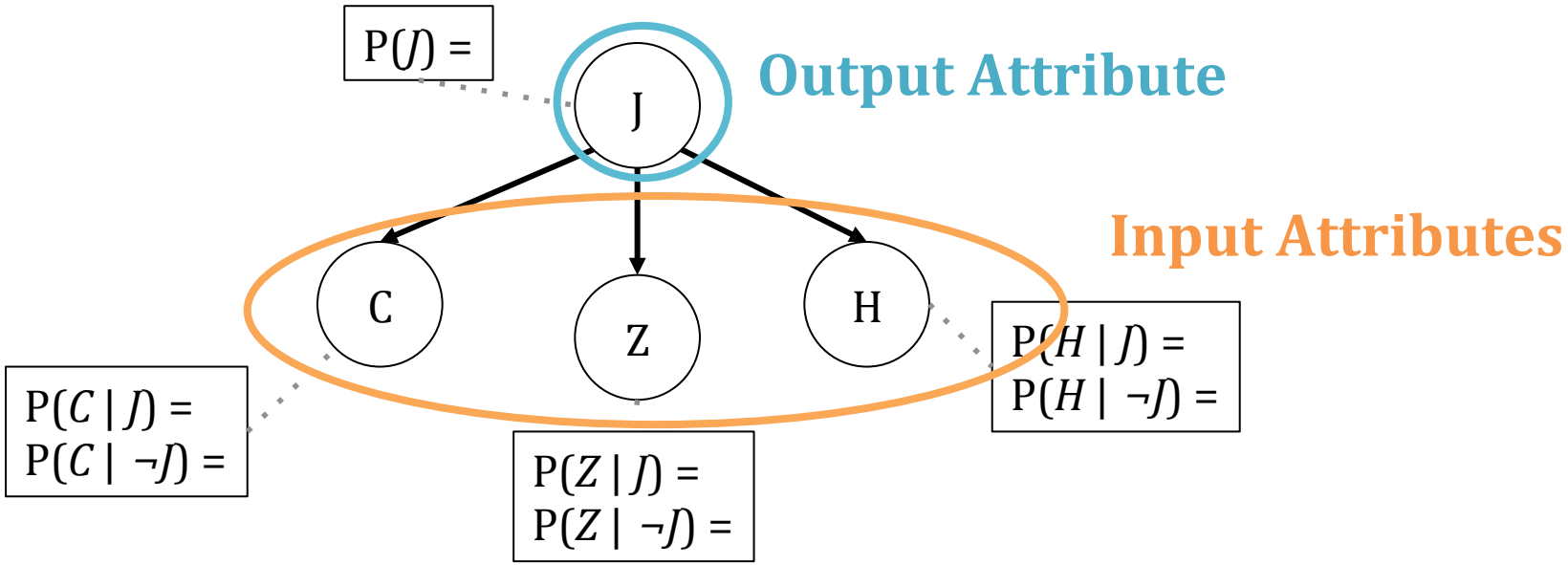
H: Saw *Harry Potter*

Suppose we had a database from 20 people. How could we use that to estimate the values in this CPT?



Naïve Bayes classifier

J: Person is a junior
C: Brought coat to classroom
Z: Lives in zip code 12345
H: Saw *Harry Potter* more than once

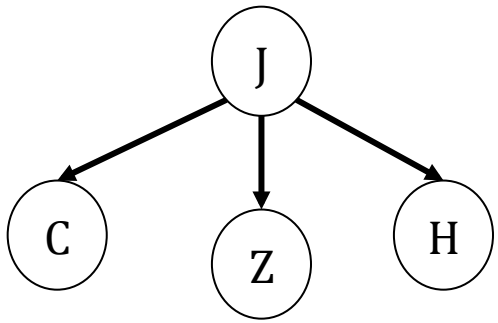


Inference with a Naïve Bayes classifier

- $P(J \mid C \wedge \neg Z \wedge H) = \dfrac{P(J \wedge C \wedge \neg Z \wedge H)}{P(C \wedge \neg Z \wedge H)}$

$$= \frac{P(J \wedge C \wedge \neg Z \wedge H)}{P(J \wedge C \wedge \neg Z \wedge H) + P(\neg J \wedge C \wedge \neg Z \wedge H)}$$

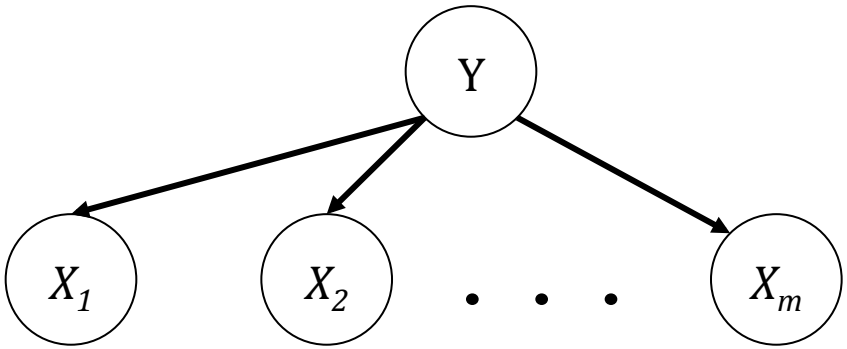
$$= \frac{P(C \mid J)\, P(\neg Z \mid J)\, P(H \mid J)\, P(J)}{\left(\begin{array}{l} P(C \mid J)\, P(\neg Z \mid J)\, P(H \mid J)\, P(J) + \\ P(C \mid \neg J)\, P(\neg Z \mid \neg J)\, P(H \mid \neg J)\, P(\neg J) \end{array}\right)}$$



Naïve Bayes—the general case

- Estimate $P(Y = v)$ as a fraction of records with $Y = v$
- Estimate $P(X_i = u_i \mid Y = v)$ as fraction of $Y = v$ records that also have $X_i = u_i$
- To **predict** the Y value given observations of all the X_i values, compute:

$$Y^{predict} = \underset{v}{argmax}\, P(Y = v \mid X_1 = u_1, \ldots, X_m = u_m)$$



- Assume features $X_1,\ldots,X_m$ are independent of each other given the class Y

Computing Naïve Bayes classification

$$Y^{predict} = \underset{v}{argmax}\, \mathrm{P}(Y = v \mid X_1 = u_1, \ldots, X_m = u_m)$$

- With **Bayes rule** we have

$$Y^{predict} = \underset{v}{argmax}\, \frac{\mathrm{P}(X_1 = u_1, \ldots, X_m = u_m \mid Y = v)\, \mathrm{P}(Y = v)}{\mathrm{P}(X_1 = u_1, \ldots, X_m = u_m)}$$

- We really only need the numerator to make a classification

$$= \underset{v}{argmax}\, \mathrm{P}(X_1 = u_1, \ldots, X_m = u_m \mid Y = v)\, \mathrm{P}(Y = v)$$

$$= \underset{v}{argmax}\, \mathrm{P}(X_1 = u_1 \mid Y = v) \ldots \mathrm{P}(X_m = u_m \mid Y = v)\, \mathrm{P}(Y = v)$$

$$= \underset{v}{argmax}\, \mathrm{P}(Y = v) \prod_i \mathrm{P}(X_i = u_i \mid Y = v)$$