Bayes Nets

Russell and Norvig: Chapter 12, 13
CSE 240: Winter 2023
Lecture 13

Announcements

- Quiz 2 grading in progress
- Assignment 3 is due tonight at 5pm
- Working on regrades

Agenda and Topics

- Probability
  - Independence and conditional independence
- Naives Bayes
- Bayesian Networks

Conditional Independence

- Unfortunately, random variables of interest are rarely independent of each other
- A more suitable notion is that of **conditional independence**
- Two variables X and Y are **conditionally independent** given Z if
  - P(x| y,z) = P(x|z) for all values x,y,z
  - That is, learning the values of Y does not change prediction of X once we know the value of Z
  - Equivalently, P(x, y|z) = P(x|z)P(y|z) for all values x,y,z
- notation: I(X ; Y | Z), $X \perp\!\!\!\perp Y | Z$

The Chain Rule

$$P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots$$

- Trivial decomposition:

$$P(\mathsf{Traffic}, \mathsf{Rain}, \mathsf{Umbrella}) = P(\mathsf{Rain})P(\mathsf{Traffic}|\mathsf{Rain})P(\mathsf{Umbrella}|\mathsf{Rain}, \mathsf{Traffic})$$

- With assumption of conditional independence:

$$P(\mathsf{Traffic}, \mathsf{Rain}, \mathsf{Umbrella}) = P(\mathsf{Rain})P(\mathsf{Traffic}|\mathsf{Rain})P(\mathsf{Umbrella}|\mathsf{Rain})$$

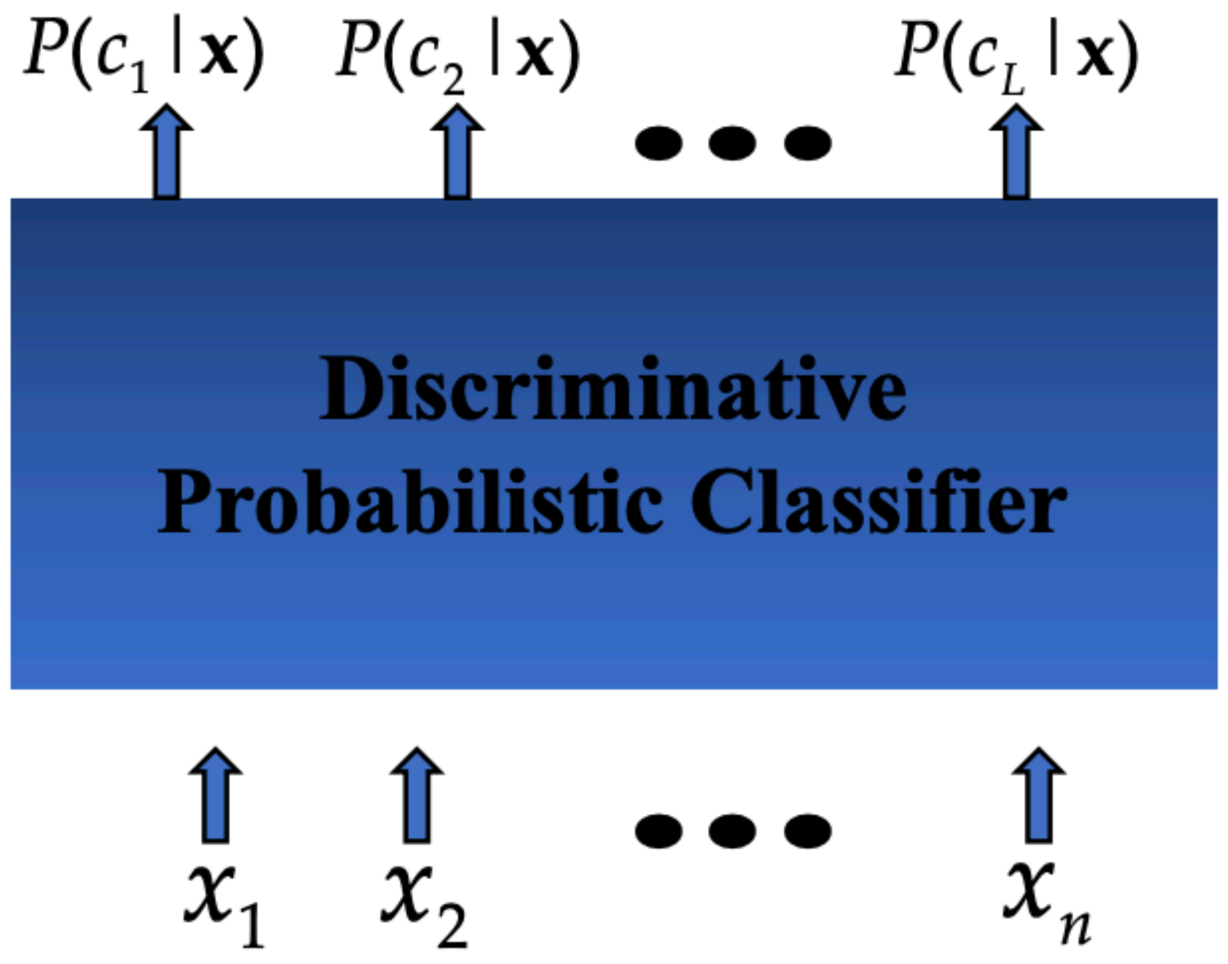
- Bayes' nets / graphical models help us express conditional independence assumptions

Naïve Bayes

Probabilistic Classification Principle

Establishing a probabilistic model for classification **(Discriminative model)**

$$\boxed{P(c\,|\,\mathbf{x}) \quad c = c_1, \cdots, c_L, \mathbf{x} = (x_1, \cdots, x_n)}$$



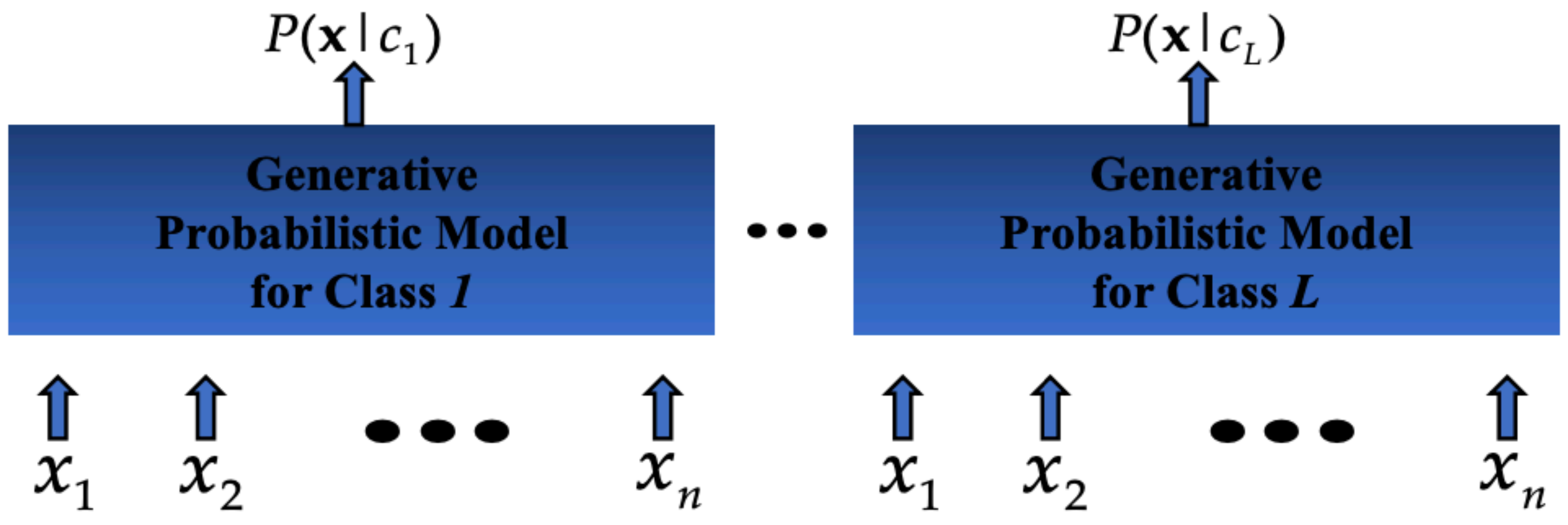
$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

Output L probabilities for L class labels in a probabilistic classifier

Probabilistic Classification Principle

Establishing a probabilistic model for classification **(Generative model)**

$$\boxed{P(\mathbf{x}\,|\,c) \quad c = c_1, \cdots, c_L,\ \mathbf{x} = (x_1, \cdots, x_n)}$$



$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

- *L* probabilistic models have to be trained independently
- Each is trained on only the examples of the same label
- Output *L* probabilities for a given input with *L* models

Probabilistic Classification Principle

- Maximum A Posterior (MAP) classification rule
  - For an input x, find the largest one from L probabilities output by a discriminative probabilistic classifier $P(c_1 | x), \ldots, P(c_L | x)$.
  - Assign x to label c^* if $P(c^* | x)$ is the largest
- Generative classification with MAP rule
  - Apply Bayesian rule to convert them into posterior probabilities
  - $P(c_i | x) = \dfrac{P(x | c_i)P(c_i)}{P(x)} \propto P(x | c_i)p(c_i)$
    - (for $i = 1,2,...L$)
  - Then apply the MAP rule to assign a label

Naïve Bayes

- Bayes classification
  - $P(c \,|\, x) \propto P(x \,|\, c)P(c) = P(x_1, \ldots x_n \,|\, c)P(c)$
    - (for $c = c_1, \ldots c_L$)
  - Difficulty: leaning the joint probability $P(x_1, \ldots x_n \,|\, c)$ is often infeasible!
- Naïve Bayes classification

  Assume all input features are conditionally independent!

$$
\begin{aligned}
P(x_1, \ldots x_n \,|\, c) &= P(x_1 \,|\, x_2, \ldots, x_n, c)P(x_2, \ldots, x_n \,|\, c) \\
&= P(x_1 \,|\, c)P(x_2, \ldots, x_n \,|\, c) \\
&= P(x_1 \,|\, c)P(x_2 \,|\, c) \ldots P(x_n \,|\, c)
\end{aligned}
$$

$$\left[P(a_1 \,|\, c^*) \ldots P(a_n \,|\, c^*)\right] P(c^*) > \left[P(a_1 \,|\, c) \ldots P(a_n \,|\, c)\right] P(c),\ c \neq c^*, c = c_1, \ldots, c_L$$

The Naïve Bayes Model

- The *Naïve Bayes Assumption*:
  - Assume that all features are independent **given the class label Y**.
- Equationally speaking:
  - $P(X_1, \ldots, x_n \mid Y) = \prod_{i=1}^{n} P(X_i \mid Y)$

Why is This Useful?

- # of parameters for modeling $P(X_1, \ldots, X_n \mid Y)$:
  - $2(2^n - 1)$
- # of parameters for modeling $P(X_1 \mid Y), \ldots, P(X_n \mid Y)$:
  - $2n$

Naïve Bayes Training

- Training in Naïve Bayes is **easy:**
  - Estimate $P(Y = v)$ as the fraction of records with $Y = v$

$$P(Y = v) = \frac{Count(Y = v)}{\#records}$$

  - Estimate $P(X_i = u \mid Y = v)$ as the fraction of records with $Y = v$ for which $X_i = u$.

$$P(X_i \mid Y = v) = \frac{Count(X_i = u \wedge Y = v)}{Count(Y = v)}$$

Naïve Bayes Training

- In practice, some of these counts can be zero
- Fix this by adding “virtual” counts

$$P(X_i \mid Y = v) = \frac{Count(X_i = u \wedge Y = v) + 1}{Count(Y = v) + 2}$$

  - This is called Smoothing

Example

The weather data, with counts and probabilities													
outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5	9/14	5/14
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5		
rainy	3/9	2/5	cool	3/9	1/5								

A new day				
outlook	temperature	humidity	windy	play
sunny	cool	high	true	?

Example

- Leaning phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=No
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=Yes) = 9/14$

$P(\text{Play}=No) = 5/14$

Solution

- Test Phase
  - Given a new instance, predict its label

    $\mathbf{x}'$=(Outlook=*Sunny,* Temperature=*Cool,* Humidity=*High,* Wind=*Strong*)
  - Look up tables achieved in the learning phrase

P(Outlook=*Sunny*|Play=*Yes*) = 2/9

P(Temperature=*Cool*|Play=*Yes*) = 3/9

P(Huminity=*High*|Play=*Yes*) = 3/9

P(Wind=*Strong*|Play=*Yes*) = 3/9

P(Play=*Yes*) = 9/14

P(Outlook=*Sunny*|Play=*No*) = 3/5

P(Temperature=*Cool*|Play==*No*) = 1/5

P(Huminity=*High*|Play=*No*) = 4/5

P(Wind=*Strong*|Play=*No*) = 3/5

P(Play=*No*) = 5/14

$$P(Yes|\mathbf{x}') \approx [P(Sunny|Yes)P(Cool|Yes)P(High|Yes)P(Strong|Yes)]P(Play=Yes) = 0.0053$$

$$P(No|\mathbf{x}') \approx [P(Sunny|No)\,P(Cool|No)P(High|No)P(Strong|No)]P(Play=No) = 0.0206$$

Given the fact $P(Yes|\mathbf{x}') < P(No|\mathbf{x}')$, we label $\mathbf{x}'$ to be "*No*".

Decision making with the MAP rule

CE 13: Other Applications

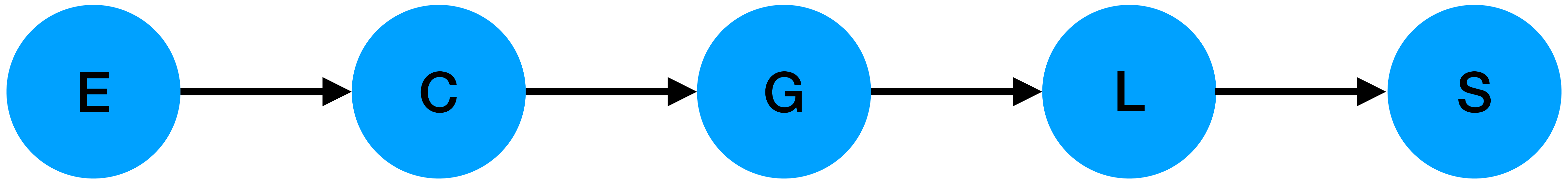
- What are some other applications where we would use Naïve Bayes? Why?

Bayesian Network

Exploiting Conditional Independence

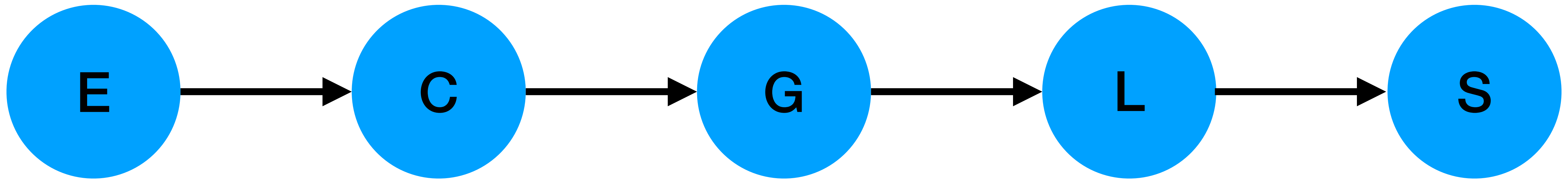
Consider a story:

If Leilani woke up too early E, Leilani probably needs a coffee C; if Leilani needs coffee, she's likely grumpy G. If she is grumpy then it's possible that the lecture won't go smoothly L. If the lecture does not go smoothly then the students will likely be sad S.



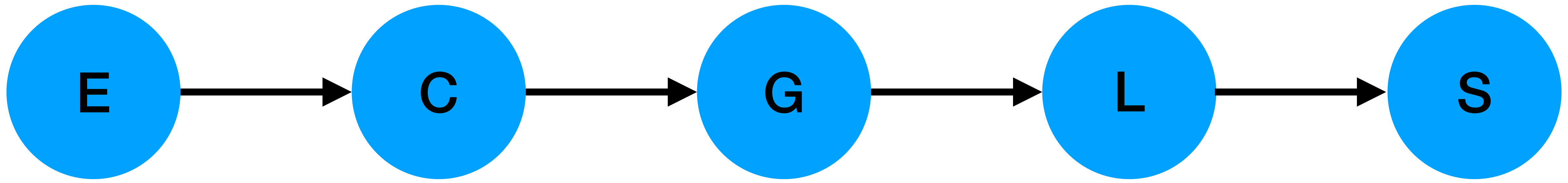
Conditional Independence

- If you learned any of E, C, G, or L, would your assessment of $p(S)$ change?
- If any of these are not seen to true, you would increase $p(S)$ and decrease $p(\neg S)$. So S is not independent of E, or C, or G, or L.
- If you knew the value of L (true or false), would leaning the value of E, C, or G influence $p(S)$?
- Influence that these factors have on S is mediated by their influence on L.
- Student's aren't sad because Leilani woke up early, they are sad because of the lecture. So S is independent of E, C, and G, given L.



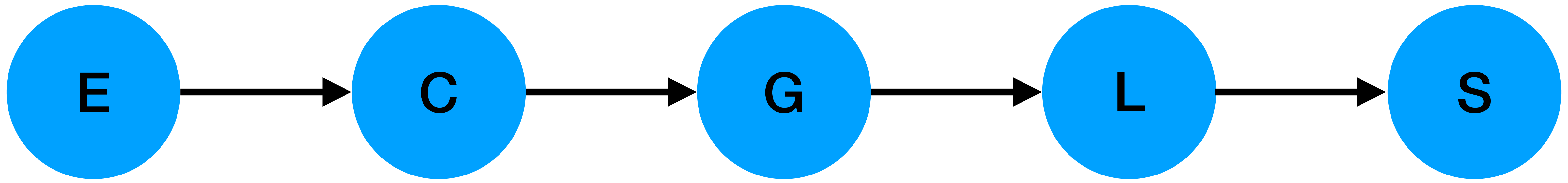
Conditional Independence

- So S is independent of E, C, and G, given L.
- Similarly
  - L is independent of E, and C, given G.
  - G is independent of E, given C
- This means that:
  - $p(S \mid L, \{G, C, E\}) = p(S \mid L)$
  - $p(L \mid G, \{C, E\}) = p(L \mid G)$
  - $p(G \mid C, \{E\}) = p(G \mid C)$
  - $p(C \mid E)$ and $p(E)$ does not simplify further.



Conditional Independence

- By the chain rule (for any instantiation of S…E):
  - $P(S, L, G, C, E) = p(S \mid L, G, C, E)p(L \mid G, C, E)p(G \mid C, E)p(C \mid E)p(E)$
- By our independence assumptions:
  - $P(S, L, G, C, E) = p(S \mid L)p(L \mid G)p(G \mid C)p(C \mid E)p(E)$
- We can specify the full joint probability by specifying five local conditional probabilities
  - $p(S \mid L)$
  - $p(L \mid G)$
  - $p(G \mid C)$
  - $p(C \mid E)$ and $p(E)$



Example Quantification

Specifying the joint requires only 9 parameters

What is $p(g)$

$p(c|e) = 0.9$
$p(\sim c|e) = 0.1$
$p(c|\sim e) = 0.5$
$p(\sim c|\sim e) = 0.5$

$p(l|g) = 0.2$
$p(\sim l|g) = 0.8$
$p(l|\sim g) = 0.1$
$p(\sim l|\sim g) = 0.9$

$$E \rightarrow C \rightarrow G \rightarrow L \rightarrow S$$

$p(e) = 0.8$
$p(\sim e) = 0.2$

$p(g|c) = 0.3$
$p(\sim g|c) = 0.7$
$p(g|\sim c) = 1$
$p(\sim g|\sim c) = 0$

$p(s|l) = 0.9$
$p(\sim s|l) = 0.1$
$p(s|\sim l) = 0.1$
$p(\sim s|\sim l) = 0.9$

Bayesian Networks

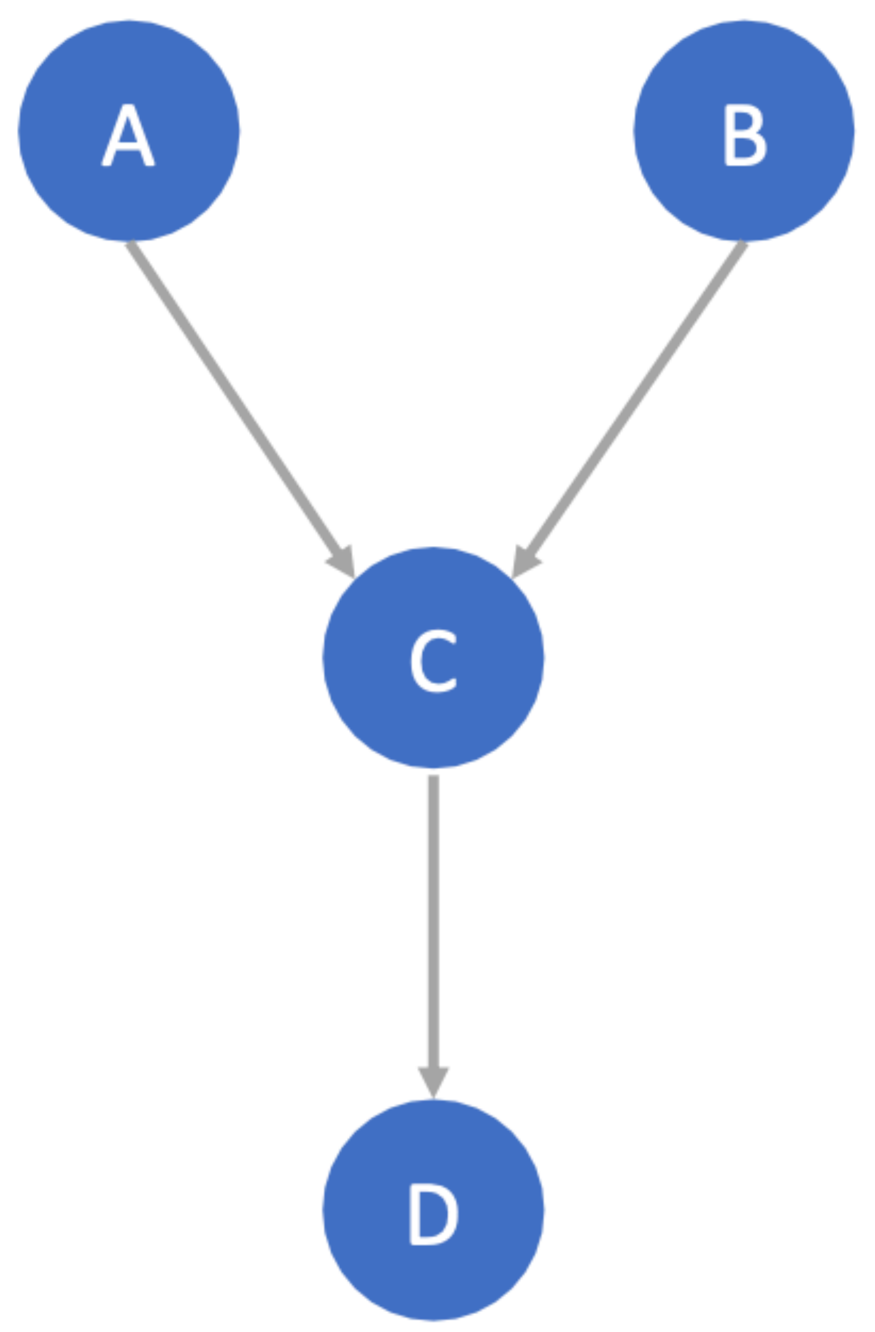
- The structure we just mentioned is a Bayesian network.
- Graphical representation of the direct dependencies over a set of variables + a set of conditional probability tables (CPTs) quantifying the strength of those influences.
- Bayesian Networks generalize the above ideas in very interesting ways, leading to effective means of representation and inference under uncertainty.

Bayesian Networks

- A simple, graphical notation for conditional independence assertions resulting in a compact representation for the full joint distribution
- Syntax:
  - a set of nodes, one per random variable
  - a directed, acyclic graph (link = ‘direct influences’)
  - a conditional distribution (CPT) for each node given its parents:
    $P(X_i|Parents(X_i))$

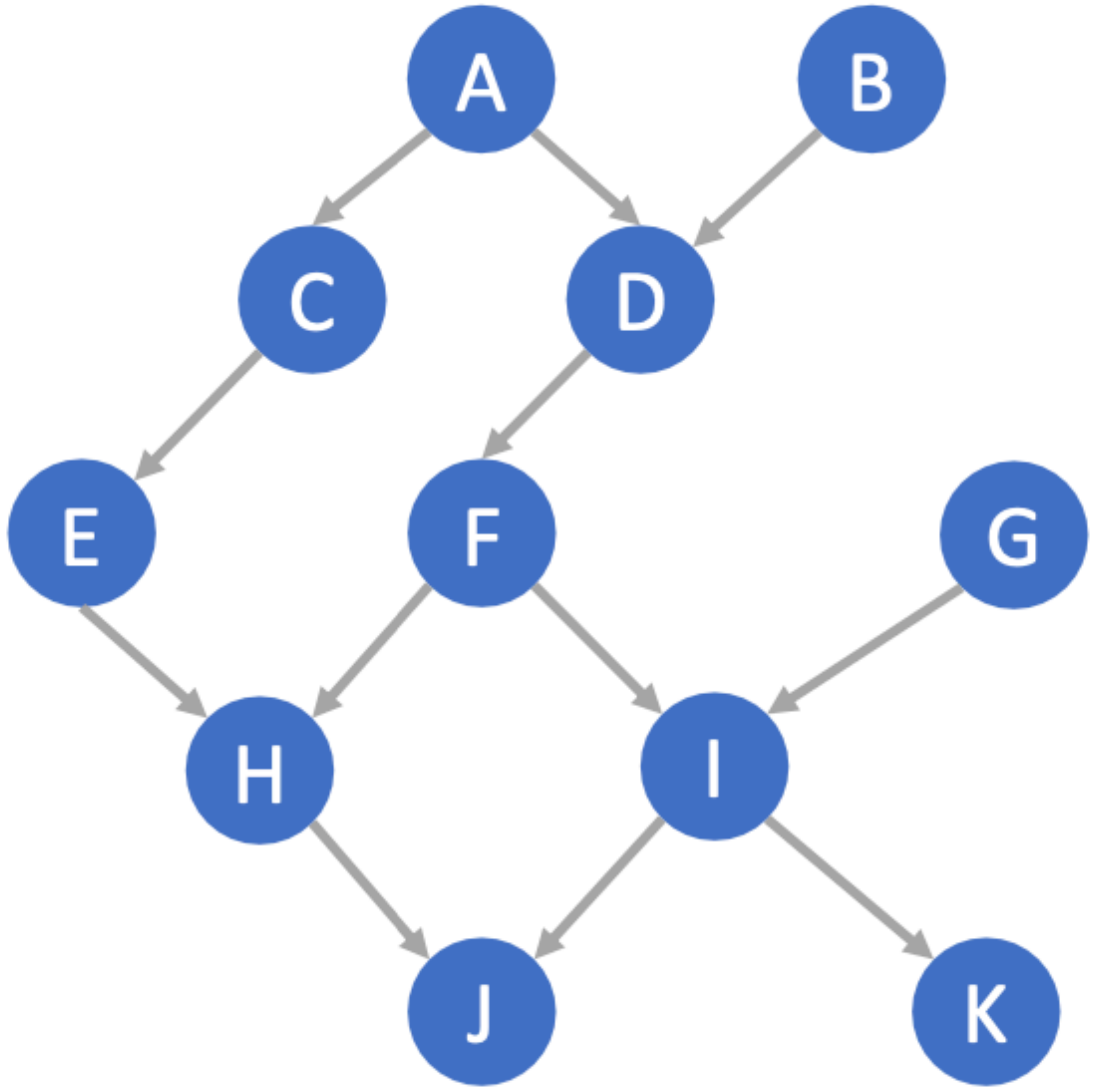
Key Notions

- Some definitions:
  - parents of a node → $par(C) = \{A, B\}$
  - children of node → $children(A) = \{C\}$
  - descendants of a node → $descendants(B) = \{C, D\}$
  - ancestors of a node → $ancestors(D) = \{A, B, C\}$
  - family: set of nodes consisting of x_i and its parents → $family(C) = \{C, A, B\}$
- CPTs are defined over families in the BN



An Example of a Bayes Net

- How many parameters do we need for the following BN?



Semantics of a Bayesian Network

- The structure of the BN means: every x_i is conditionally independent of all of its non-descendants given its parents:
  - $p(x_i | X \cup par(x_i)) = p(par(x_i))$
  - For any subset $S \subseteq non - descendants(x_i)$

How to build a Bayesian Network

1. Define a total order over the random variables: $(x_1, \ldots, x_n)$
2. Apply the chain rule:
$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | x_1, \ldots, x_{i-1})$$
3. For each x_i, select the smallest set of predecessors $par(x_i)$ such that:
$$p(x_i | x_1, \ldots, x_{i-1}) = p(x_i | \, par(x_i))$$
4. Then we can rewrite
$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i | par(x_i))$$

How to build a Bayesian Network (2)

5. This is a compact representation of the initial JPD. Factorization of the JPD based on existing conditional independencies among the variables
6. Construct the Bayesian Net (BN)
   - ✓ Nodes are the random variables
   - ✓ Draw a directed edge from each variable in $par(x_i)$ to x_i
   - ✓ Define a conditional probability table (CPT) for each variable x_i:

$$p(x_i \mid par(x_i))$$

Example for BN Construction

- You want to diagnose whether there is a fire in a building
- You can receive reports (possibly noisy) about whether everyone is leaving the building
- If everyone is leaving, this may have been caused by a firealarm
- If there is a fire alarm, it may have been caused by a fire or by tampering
- If there is a fire, there may be smoke

Fire Diagnosis: Step 1

- Start by choosing the random variables for this domain:
  - Tampering (T) is true when the alarm has been tampered with
  - Fire (F) is true when there is a fire
  - Alarm (A) is true when there is an alarm
  - Smoke (S) is true when there is smoke
  - Leaving (L) is true if there are lots of people leaving the building
  - Report (R) is true if the sensor reports that lots of people are leaving the building

Fire Diagnosis: Step 2

- Define total ordering of variables.
- Let’s choose an order that follows the causal sequence of events:
    - Fire(F), Tampering (T), Alarm (A), Smoke (S), Leaving (L), Report (R)

Fire Diagnosis: Step 3

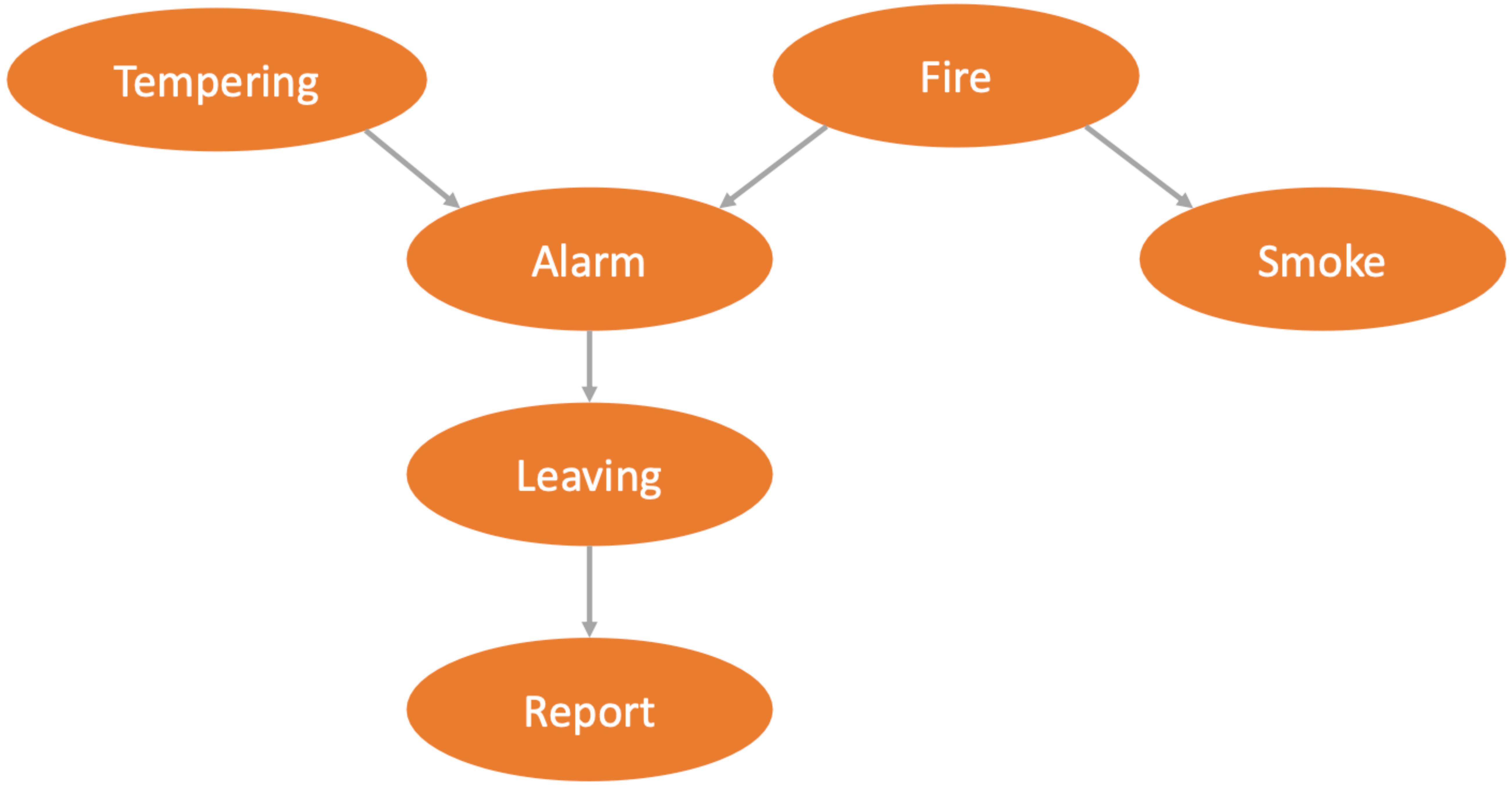
- Ordering:
  - Fire(F), Tampering (T), Alarm (A), Smoke (S), Leaving (L), Report (R)
- Apply the chain rule:
- $p(F, T, A, S, L, R) = p(F)p(T \mid F)p(A \mid F, T)p(S \mid F, T, A)p(L \mid F, T, A, S)p(R \mid F, T, A, S, L)$

Fire Diagnosis: Step 4 & 5

- $p(F, T, A, S, L, R) = p(F)p(T \mid F)p(A \mid F, T)p(S \mid F, T, A)p(L \mid F, T, A, S)p(R \mid F, T, A, S, L)$
- For each variable, x_i choose parents $par(x_i)$ and re-write the joint probability distribution:
  - $p(F, T, A, S, L, R) = p(F)p(T)p(A \mid F, T)p(S \mid F)p(L \mid A)p(R \mid L)$
- Now we need to build the BN based on the above JPD.

Fire Diagnosis: Drawing BN

- $p(F, T, A, S, L, R) = p(F)p(T)p(A \mid F, T)p(S \mid F)p(L \mid A)p(R \mid L)$



Summary and Next Time

- Today:
  - Naïve Bayes
  - Bayes Networks
- Thursday
  - Continue Bayes Networks
  - Markov Decision Processes