

CSE 592 - Discussion 4

Probability & Bayesian Networks

Zhengjie (Jerry) Xu / Fall 26

Logistics

Project proposal is due **September 30** at 11:59 p.m

Homework 2 is out: Due **Oct 7th**.

Anonymous feedback form: +0.5% bonus (instructions on Ed).

Jerry's office hours: Tuesdays 4–5 p.m., 1637 BBB.

Today: probability, Naïve Bayes, and Bayesian networks.

We will practice Bayes' rule, d-separation, and variable elimination.

Recap: outcomes, events, and probability

An outcome is one result of a random experiment. Ω : all outcomes.

An event is a set of outcomes, e.g. "the sum of two dice is 7."

Axioms: $P(A) \geq 0$, $P(\Omega) = 1$, and disjoint events add.

$$P(\neg A) = 1 - P(A)$$

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

A random variable is a function of the outcome.

It is deterministic. The randomness comes from which outcome occurs.

Counting the overlap once

R: it rains today. W: it is windy today.

$$P(R) = 0.3, \quad P(W) = 0.4, \quad P(R \wedge W) = 0.1$$

Q: What is the probability of rain or wind? Of neither?

A: $P(R \vee W) = 0.3 + 0.4 - 0.1 = 0.6$. So $P(\text{neither}) = 0.4$.

Adding $0.3 + 0.4$ counts the rainy, windy days twice. If R and W were disjoint, $P(R \wedge W) = 0$ and the sum 0.7 would be right.

Recap: joint, marginal, and conditional

A joint distribution gives a probability to every full combination of values.

Marginalize: sum out a variable. $P(R) = P(R, U) + P(R, \neg U)$

Condition: keep only the cases where B holds, then renormalize.

$P(A \mid B) = P(A \wedge B) / P(B)$, when $P(B) > 0$

Product rule: $P(A \wedge B) = P(A \mid B) P(B)$

The joint answers every query, but it has one entry per combination.

Reading a joint table

R: it rains. U: a person carries an umbrella.

	U	$\neg U$	Total
R	0.24	0.06	0.30
$\neg R$	0.14	0.56	0.70
Total	0.38	0.62	1.00

Q: What are $P(R \mid U)$ and $P(U \mid R)$?

A: $P(R \mid U) = 0.24 / 0.38 \approx 0.63$. $P(U \mid R) = 0.24 / 0.30 = 0.80$.

Recap: Bayes' rule

$$P(H \mid E) = P(E \mid H) P(H) / P(E)$$

Prior $P(H)$: belief before seeing the evidence.

Likelihood $P(E \mid H)$: how expected the evidence is if H holds.

Posterior $P(H \mid E)$: belief after seeing the evidence.

Total probability: $P(E) = P(E \mid H) P(H) + P(E \mid \neg H) P(\neg H)$

$P(E)$ is the same for every H . It only normalizes.

Base rates: a smoke detector

1% of homes have a fire. F: fire. A: the detector beeps.

$P(A | F) = 0.99$. $P(A | \neg F) = 0.02$.

Q: The detector beeps. How likely is a fire?

A: $P(F | A) = 0.0099 / (0.0099 + 0.0198) = 1/3$.

Expected utility: take an umbrella?

$P(\text{rain}) = 0.3$. Utilities of each action in each outcome:

Action	Rain	No rain	Expected utility
Take umbrella	0	-1	-0.7
Leave it	-10	0	-3.0

Q: Below what rain probability should we leave the umbrella?

A: Solve $-(1 - p) = -10p$. The two actions tie at $p = 1/11 \approx 0.09$.

Inference gives p . Utilities decide where the threshold sits.

Recap: learning parameters

Likelihood $P(D \mid \theta)$: probability of the data, as a function of θ .

MLE: choose θ that maximizes $P(D \mid \theta)$.

MAP: choose θ that maximizes $P(D \mid \theta) P(\theta)$.

Bayesian learning keeps the whole posterior $P(\theta \mid D)$.

For i.i.d. data, maximize the log-likelihood: a sum instead of a product.

Categorical MLE is counting: $\hat{\theta} = (\text{count of the value}) / (\text{total count})$.

MAP vs. MLE: a trick coin

A jar holds 90% fair coins and 10% two-headed coins. We flip one: HH.

Coin	Prior	Likelihood	Prior $\times$ lik.	Posterior
Fair ($\theta = 0.5$)	0.9	1/4	0.225	≈ 0.69
Two-headed ($\theta = 1$)	0.1	1	0.100	≈ 0.31

Q: Which coin does MAP choose? Which does MLE choose?

A: MAP picks fair ($0.225 > 0.1$). MLE picks two-headed ($1 > 1/4$).

Q: Under which condition does MAP and MLE agree with each other?

A: The prior are uniform.

Counting, zeros, and smoothing

A die is rolled 10 times. Counts for faces 1–6: 2, 1, 3, 0, 2, 2.

MLE: $\hat{P}(3) = 3/10$. $\hat{P}(4) = 0/10$.

Q: Should face 4 be impossible? What does add-one smoothing give?

A: No, we just did not see it. $\hat{P}(4) = (0 + 1) / (10 + 6) = 1/16$.

The denominator adds 6: one extra count for each face. $\hat{P}(3)$ becomes $4/16$. The six estimates still sum to 1.

Recap: independence

$X \perp Y$: $P(X, Y) = P(X) P(Y)$ for all values

$X \perp Y \mid Z$: $P(X, Y \mid Z) = P(X \mid Z) P(Y \mid Z)$ for all values

Full joint over n Boolean variables: $2^n - 1$ free parameters. (Why -1?)

If all n are mutually independent: only n parameters.

$n = 30$: about 1.07 billion vs. 30.

Independence is often an assumption. It is useful when it is roughly true.

Conditional independence $\neq$ independence

Two thermometers T_1 and T_2 measure the same patient. F : fever.

Given F , each thermometer reads high with its own noise.

Q: Are T_1 and T_2 independent? Are they independent given F ?

A: Not independent: a high T_1 makes F likely, so T_2 likely reads high.

A: Given F , T_1 tells us nothing more about T_2 : $T_1 \perp T_2 \mid F$.

Same pattern as the coin flips in lecture: dependent until we know the coin.

Recap: Naïve Bayes

Story: pick class c from $P(c)$. Draw each word independently from $P(w \mid c)$.

$$\text{score}(c) = P(c) \cdot \prod_w P(w \mid c)^{n_w}$$

Train by counting: $\hat{P}(c) = N_c / N$, $\hat{\theta}_{wc} = C_{wc} / T_c$

Bag of words: only counts are kept; word order is ignored.

Predict the class with the larger score. Normalize to get probabilities.

Add log-probabilities instead of multiplying to avoid underflow.

Laplace smoothing: $\hat{\theta}_{wc} = (C_{wc} + 1) / (T_c + V)$

Naïve Bayes by hand

6 positive and 4 negative reviews. Each class has 20 tokens. $V = 5$ words.

Class	great	boring	plot	fun	awful
Pos	6	1	5	8	0
Neg	1	7	5	1	6

Review "great plot boring":

Pos: $0.6 \cdot (6/20)(5/20)(1/20) = 0.00225$

Neg: $0.4 \cdot (1/20)(5/20)(7/20) = 0.00175$

Normalize: $P(\text{Pos} \mid \text{review}) = 0.00225 / 0.004 \approx 0.56$.

Q: "plot" has the same rate in both classes. Does it matter?

A: No. It multiplies both scores by $1/4$, which cancels after normalizing.

One zero vetoes everything

Review: "fun fun awful". "awful" never appeared in a positive review.

Unsmoothed Pos score: $0.6 \cdot (8/20)^2 \cdot (0/20) = 0$. Neg wins by default.

Q: What changes with add-one smoothing ($V = 5$)?

A: Denominators become $20 + 5 = 25$.

Pos: $0.6 \cdot (9/25)^2 \cdot (1/25) \approx 0.00311$

Neg: $0.4 \cdot (2/25)^2 \cdot (7/25) \approx 0.00072$ $P(\text{Pos} \mid \text{review}) \approx 0.81$

The two "fun"s were strong evidence. The zero had silenced them.

Class	great	boring	plot	fun	awful
Pos	6	1	5	8	0
Neg	1	7	5	1	6

What the independence assumption ignores

Q: How does Naïve Bayes see “not good” vs. “good, not”?

A: Identically. The bag of words drops order, so negation is lost.

Q: What happens to “New York” in a review?

A: “New” and “York” count as two separate pieces of evidence.

Correlated words are double-counted, so posteriors are often overconfident.

The predicted class is still often right. That is why NB works in practice.

Recap: Bayesian networks

A DAG over the variables, plus one CPT $P(X \mid \text{Parents}(X))$ per node.

$$P(x_1, \dots, x_n) = \prod_i P(x_i \mid \text{parents}(x_i))$$

A Boolean node with k Boolean parents needs 2^k free parameters.

Arrows encode a factorization, not necessarily causation.

Local Markov property: given its parents,
a node is independent of its nondescendants.

Today's network

$$P(c) = 0.5$$

$$P(s \mid c) = 0.1 \quad P(s \mid \neg c) = 0.5$$

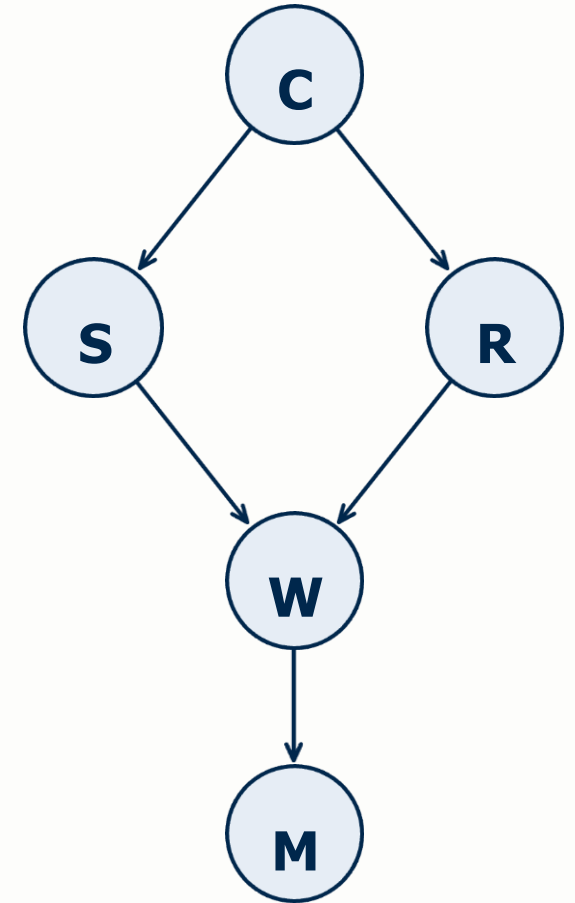
$$P(r \mid c) = 0.8 \quad P(r \mid \neg c) = 0.2$$

$$P(w \mid s, r) = 0.99 \quad P(w \mid s, \neg r) = 0.90$$

$$P(w \mid \neg s, r) = 0.90 \quad P(w \mid \neg s, \neg r) = 0.01$$

$$P(m \mid w) = 0.8 \quad P(m \mid \neg w) = 0.1$$

Fork: $S \leftarrow C \rightarrow R$. Collider: $S \rightarrow W \leftarrow R$. Chain: $R \rightarrow W \rightarrow M$.



C cloudy · S sprinkler · R rain
W wet grass · M muddy shoes

Probability of one world

$$\begin{aligned}P(c, \neg s, r, w, m) &= P(c) P(\neg s \mid c) P(r \mid c) P(w \mid \neg s, r) P(m \mid w) \\&= 0.5 \cdot 0.9 \cdot 0.8 \cdot 0.9 \cdot 0.8 \approx 0.259\end{aligned}$$

Q: Why is there no term like $P(m \mid c, s, r, w)$?

A: Given its parent W , M is independent of C , S , and R .

Each world takes one entry from each CPT: five numbers.

$$P(c) = 0.5$$

$$P(s \mid c) = 0.1 \quad P(s \mid \neg c) = 0.5$$

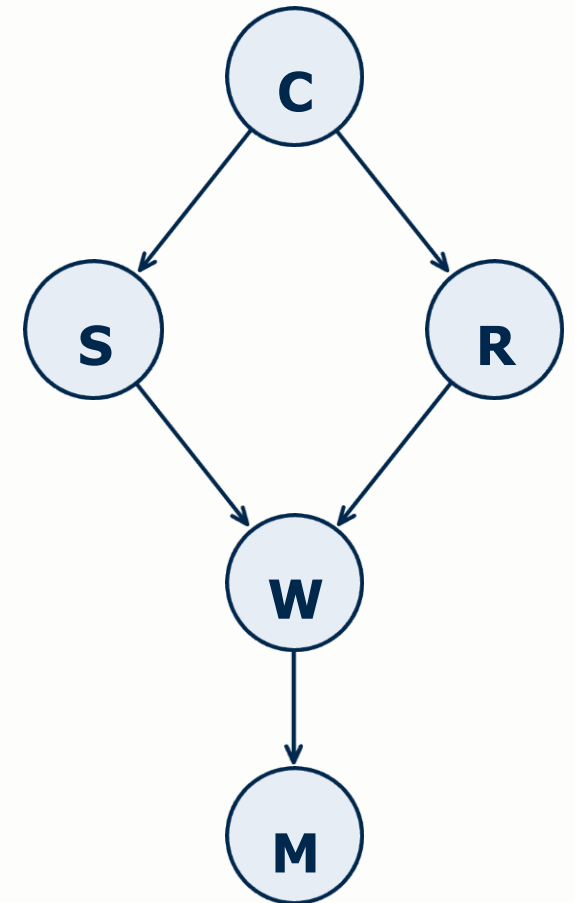
$$P(r \mid c) = 0.8 \quad P(r \mid \neg c) = 0.2$$

$$P(w \mid s, r) = 0.99 \quad P(w \mid s, \neg r) = 0.90$$

$$P(w \mid \neg s, r) = 0.90 \quad P(w \mid \neg s, \neg r) = 0.01$$

$$P(m \mid w) = 0.8 \quad P(m \mid \neg w) = 0.1$$

Fork: $S \leftarrow C \rightarrow R$. Collider: $S \rightarrow W \leftarrow R$. Chain: $R \rightarrow W \rightarrow M$.



C cloudy · S sprinkler · R rain
W wet grass · M muddy shoes

How many parameters?

Full joint over 5 Boolean variables: $2^5 - 1 = 31$.

Network: $C \ 1 + S \ 2 + R \ 2 + W \ 4 + M \ 2 = 11$.

Q: We add an arrow $C \rightarrow W$. How many parameters now?

A: W has 3 parents: $2^3 = 8$ rows. Total: $1 + 2 + 2 + 8 + 2 = 15$.

The graph is still a DAG.

Each extra parent doubles that node's CPT.

$$P(c) = 0.5$$

$$P(s \mid c) = 0.1 \quad P(s \mid \neg c) = 0.5$$

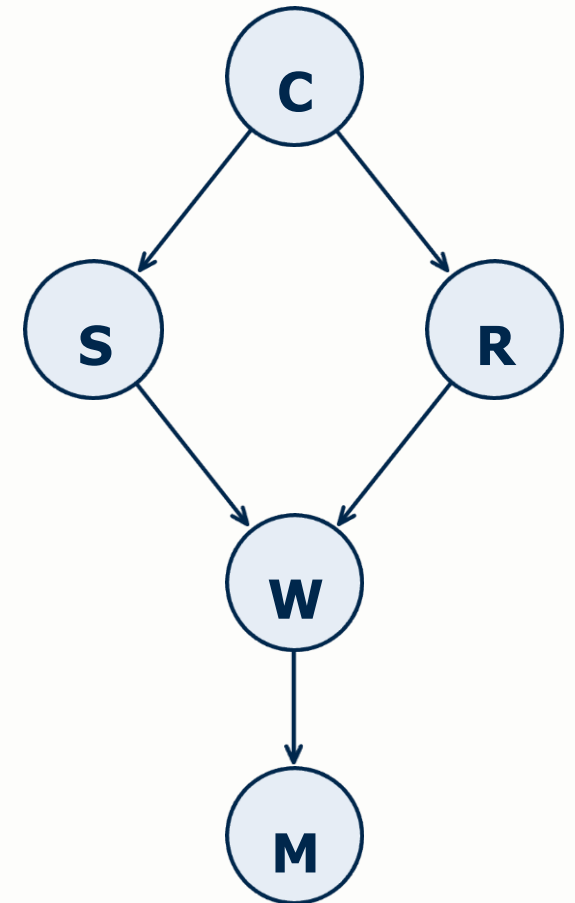
$$P(r \mid c) = 0.8 \quad P(r \mid \neg c) = 0.2$$

$$P(w \mid s, r) = 0.99 \quad P(w \mid s, \neg r) = 0.90$$

$$P(w \mid \neg s, r) = 0.90 \quad P(w \mid \neg s, \neg r) = 0.01$$

$$P(m \mid w) = 0.8 \quad P(m \mid \neg w) = 0.1$$

Fork: $S \leftarrow C \rightarrow R$. Collider: $S \rightarrow W \leftarrow R$. Chain: $R \rightarrow W \rightarrow M$.



C cloudy · S sprinkler · R rain
W wet grass · M muddy shoes

Recap: d-separation

A trail is a path that ignores arrow directions.

Chain $X \rightarrow Y \rightarrow Z$ and fork $X \leftarrow Y \rightarrow Z$: blocked if Y is observed.

Collider $X \rightarrow Y \leftarrow Z$: blocked unless Y or a descendant of Y is observed.

A trail is active only if every middle node lets it through.

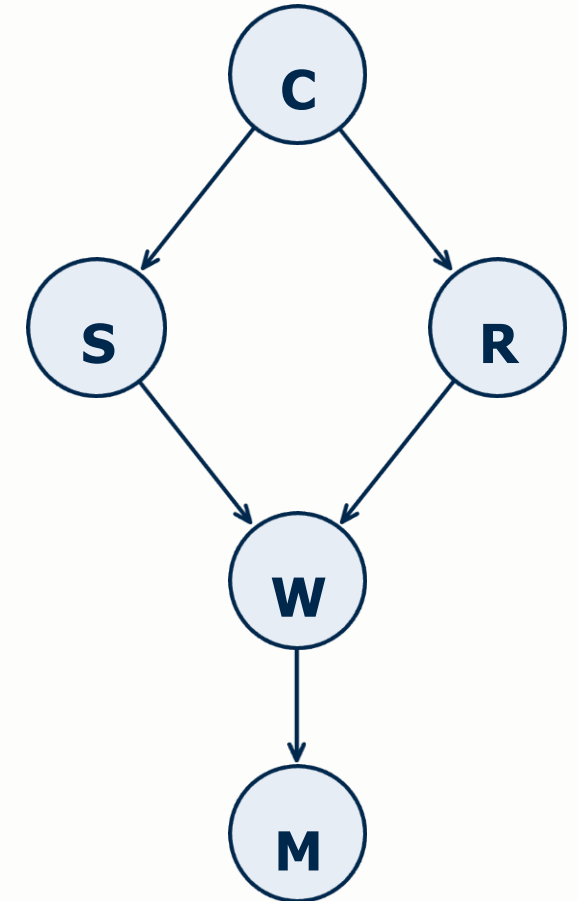
If every trail between X and Y is blocked by E , then $X \perp Y \mid E$
holds for any CPTs on this graph.

Which independences does the graph guarantee?

Q: For each claim: guaranteed or not?

Claim	Holds?	Why
$S \perp R$	No	Fork $S \leftarrow C \rightarrow R$ is open
$S \perp R \mid C$	Yes	Fork blocked; W unobserved
$S \perp R \mid C, W$	No	Observed collider W
$S \perp R \mid C, M$	No	M observed, a descendant of W
$C \perp M \mid W$	Yes	Every trail passes W as a chain

A: Observing W or M makes S and R dependent again.



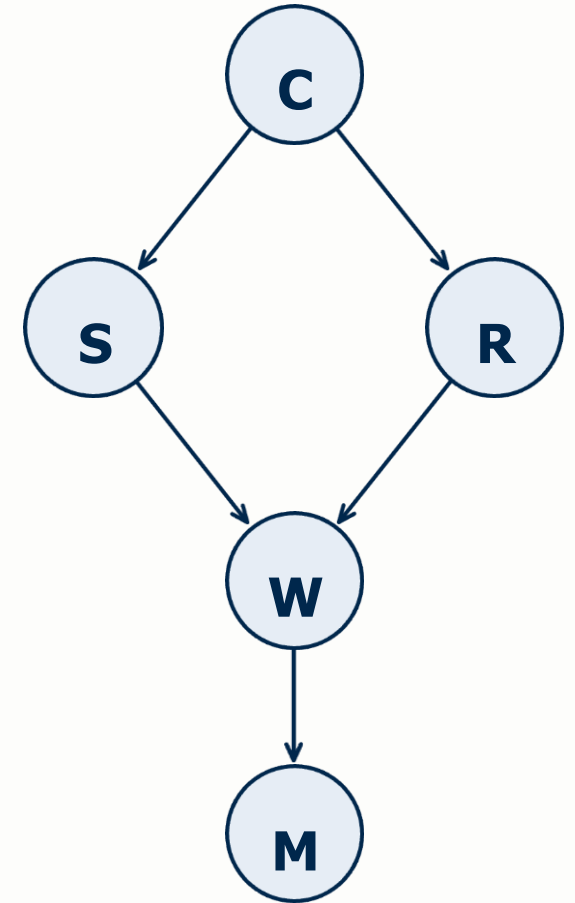
C cloudy · S sprinkler · R rain
W wet grass · M muddy shoes

The Markov blanket

Markov blanket: parents, children, and children's other parents.
Given its blanket, a node is independent of every other node.

Q: What is the Markov blanket of S? Of W?

A: $MB(S) = \{C, W, R\}$. $MB(W) = \{S, R, M\}$.



C cloudy · S sprinkler · R rain
W wet grass · M muddy shoes

Recap: variable elimination

Query X , evidence $E = e$, hidden variables Y .

$$P(X \mid e) \propto \sum_y \prod_i P(x_i \mid \text{parents}(x_i))$$

1. Restrict every factor to the observed evidence values.
2. For each hidden variable: multiply the factors containing it, then sum it out.
3. Multiply the remaining factors and normalize.

Prune first: drop nodes that are not query or evidence nodes or their ancestors.

Exact, but the elimination order controls the cost.

Variable elimination: setting up $P(R \mid w)$

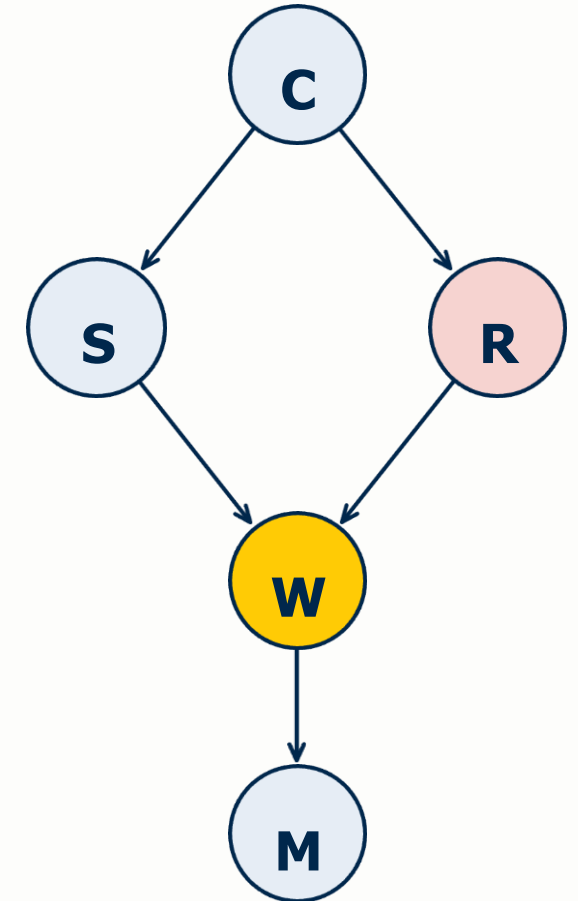
Query: R . Evidence: $W = \text{true}$. Hidden: C, S, M .

Q: Which hidden variable can we drop without any arithmetic?

A: M . It is not an ancestor of R or W , and $\sum_m P(m \mid w) = 1$.

Remaining factors: $P(C), P(S \mid C), P(R \mid C), P(w \mid S, R)$.

Plan: eliminate C , then S .



C cloudy · S sprinkler · R rain
W wet grass · M muddy shoes
maize = observed · pink = query

Eliminate C

$$f_1(S, R) = \sum_c P(c) P(S \mid c) P(R \mid c)$$

S	R	Computation	$f_1(S, R)$
T	T	$0.5 \cdot 0.1 \cdot 0.8 + 0.5 \cdot 0.5 \cdot 0.2$	0.09
T	F	$0.5 \cdot 0.1 \cdot 0.2 + 0.5 \cdot 0.5 \cdot 0.8$	0.21
F	T	$0.5 \cdot 0.9 \cdot 0.8 + 0.5 \cdot 0.5 \cdot 0.2$	0.41
F	F	$0.5 \cdot 0.9 \cdot 0.2 + 0.5 \cdot 0.5 \cdot 0.8$	0.29

Q: The four entries sum to 1. Why?

A: $f_1(S, R)$ is the joint $P(S, R)$: we summed C out of $P(C, S, R)$.

$$P(c) = 0.5$$

$$P(s \mid c) = 0.1 \qquad P(s \mid \neg c) = 0.5$$

$$P(r \mid c) = 0.8 \qquad P(r \mid \neg c) = 0.2$$

$$P(w \mid s, r) = 0.99 \qquad P(w \mid s, \neg r) = 0.90$$

$$P(w \mid \neg s, r) = 0.90 \qquad P(w \mid \neg s, \neg r) = 0.01$$

$$P(m \mid w) = 0.8 \qquad P(m \mid \neg w) = 0.1$$

Fork: $S \leftarrow C \rightarrow R$. Collider: $S \rightarrow W \leftarrow R$. Chain: $R \rightarrow W \rightarrow M$.

Eliminate S and normalize

$$\text{score}(R) = \sum_s f_1(s, R) P(w \mid s, R)$$

$$\text{score}(r) = 0.09 \cdot 0.99 + 0.41 \cdot 0.90 = 0.4581$$

$$\text{score}(\neg r) = 0.21 \cdot 0.90 + 0.29 \cdot 0.01 = 0.1919$$

$$P(r \mid w) = 0.4581 / 0.65 \approx 0.70. \quad \text{The prior was } P(r) = 0.5.$$

Q: What does the normalizer 0.65 represent?

A: It is $P(w)$, the probability of the evidence.

$$P(c) = 0.5$$

$$P(s \mid c) = 0.1 \quad P(s \mid \neg c) = 0.5$$

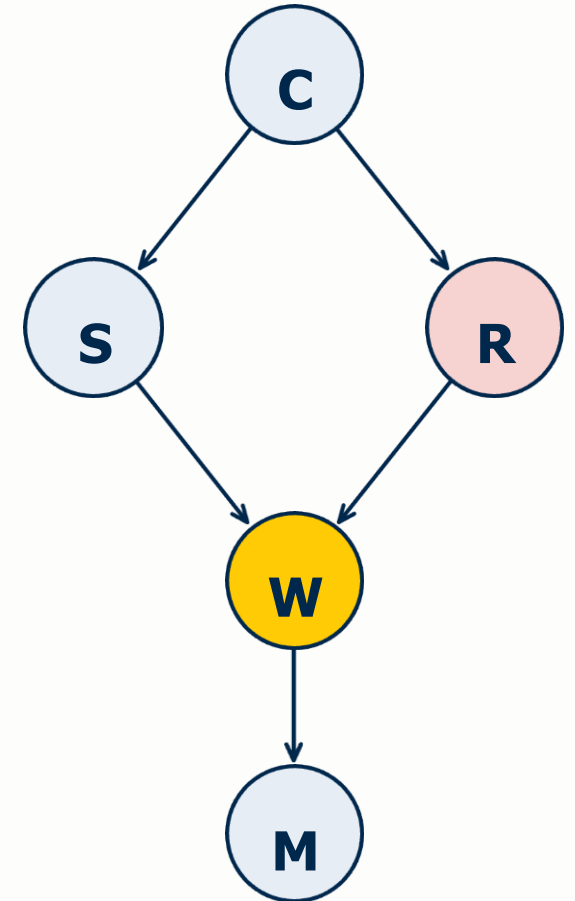
$$P(r \mid c) = 0.8 \quad P(r \mid \neg c) = 0.2$$

$$P(w \mid s, r) = 0.99 \quad P(w \mid s, \neg r) = 0.90$$

$$P(w \mid \neg s, r) = 0.90 \quad P(w \mid \neg s, \neg r) = 0.01$$

$$P(m \mid w) = 0.8 \quad P(m \mid \neg w) = 0.1$$

Fork: $S \leftarrow C \rightarrow R$. Collider: $S \rightarrow W \leftarrow R$. Chain: $R \rightarrow W \rightarrow M$.



C cloudy · S sprinkler · R rain
W wet grass · M muddy shoes
maize = observed · pink = query

Explaining away

Now we also learn the sprinkler was on: evidence w and s .

$$\text{score}(r) = f_1(s, r) \cdot 0.99 = 0.0891$$

$$\text{score}(\neg r) = f_1(s, \neg r) \cdot 0.90 = 0.189$$

$$P(r \mid w, s) = 0.0891 / 0.2781 \approx 0.32$$

Q: Why did rain drop from 0.70 to 0.32?

A: The sprinkler already explains the wet grass.

S and R share the observed child W . Learning one cause lowers the other.

$$P(c) = 0.5$$

$$P(s \mid c) = 0.1$$

$$P(s \mid \neg c) = 0.5$$

$$P(r \mid c) = 0.8$$

$$P(r \mid \neg c) = 0.2$$

$$P(w \mid s, r) = 0.99$$

$$P(w \mid s, \neg r) = 0.90$$

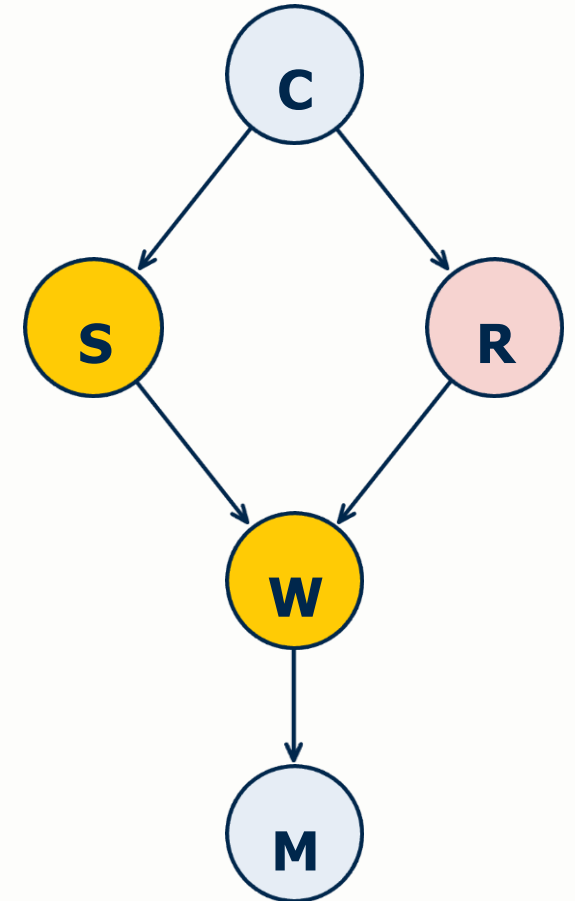
$$P(w \mid \neg s, r) = 0.90$$

$$P(w \mid \neg s, \neg r) = 0.01$$

$$P(m \mid w) = 0.8$$

$$P(m \mid \neg w) = 0.1$$

Fork: $S \leftarrow C \rightarrow R$. Collider: $S \rightarrow W \leftarrow R$. Chain: $R \rightarrow W \rightarrow M$.



C cloudy · S sprinkler · R rain
W wet grass · M muddy shoes
maize = observed · pink = query

Recap: approximate inference by sampling

Forward sampling: visit nodes in topological order and sample each from its CPT.

Estimate $P(x)$ by the fraction of samples with $X = x$.

Rejection sampling: forward sample, then discard samples that contradict the evidence.

Gibbs sampling: fix the evidence and resample one variable at a time from $P(X \mid \text{its Markov blanket})$.

Discard the early burn-in samples before counting.

Forward sampling by hand

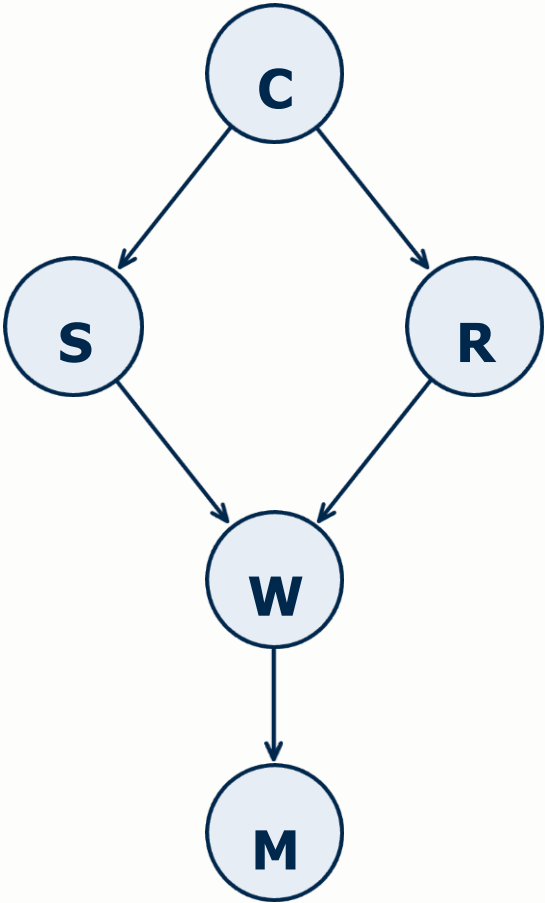
Rule: set $X = \text{true}$ if the random number $u < P(x \mid \text{parents})$.

Node	$P(\text{true} \mid \text{parents})$	u	Value
C	$P(c) = 0.5$	0.30	true
S	$P(s \mid c) = 0.1$	0.60	false
R	$P(r \mid c) = 0.8$	0.10	true
W	$P(w \mid \neg s, r) = 0.9$	0.50	true
M	$P(m \mid w) = 0.8$	0.95	false

Q: Why must we follow a topological order?

A: Each CPT row is chosen by the parents' values, so parents go first.

$P(c) = 0.5$
 $P(s \mid c) = 0.1$ $P(s \mid \neg c) = 0.5$
 $P(r \mid c) = 0.8$ $P(r \mid \neg c) = 0.2$
 $P(w \mid s, r) = 0.99$ $P(w \mid s, \neg r) = 0.90$
 $P(w \mid \neg s, r) = 0.90$ $P(w \mid \neg s, \neg r) = 0.01$
 $P(m \mid w) = 0.8$ $P(m \mid \neg w) = 0.1$
Fork: $S \leftarrow C \rightarrow R$. Collider: $S \rightarrow W \leftarrow R$. Chain: $R \rightarrow W \rightarrow M$.



C cloudy · S sprinkler · R rain
W wet grass · M muddy shoes

Rejection sampling and rare evidence

To estimate $P(r \mid w)$, keep only the samples with $W = \text{true}$.

$P(w) = 0.65$, so about 65 of every 100 samples survive.

Q: If the evidence had probability 0.001, how many samples give 100 accepted?

A: About 100,000. Almost all of the work is thrown away.

Gibbs sampling fixes the evidence, so it never rejects a sample.

One Gibbs update

Evidence $W = \text{true}$. Current state: $C = \text{true}$, $R = \text{true}$, $M = \text{true}$. Resample S .

$$P(S \mid c, r, w) \propto P(S \mid c) \cdot P(w \mid S, r)$$

$$s: 0.1 \cdot 0.99 = 0.099 \qquad \neg s: 0.9 \cdot 0.90 = 0.81$$

$$P(s \mid c, r, w) = 0.099 / 0.909 \approx 0.11$$

Q: Why do we not need M or $P(r \mid c)$?

A: M is outside S 's blanket.

A: $P(r \mid c)$ does not involve S , so it cancels when we normalize.

$$P(c) = 0.5$$

$$P(s \mid c) = 0.1 \qquad P(s \mid \neg c) = 0.5$$

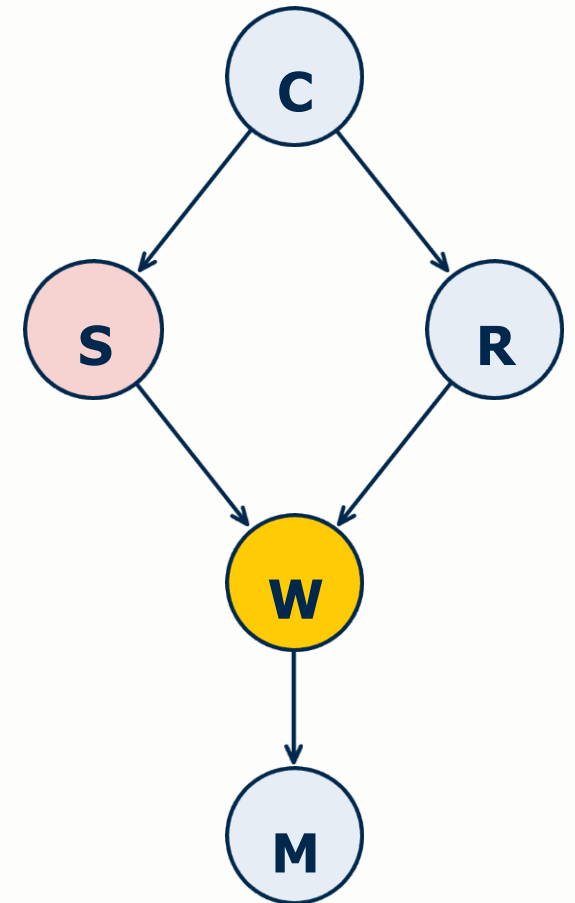
$$P(r \mid c) = 0.8 \qquad P(r \mid \neg c) = 0.2$$

$$P(w \mid s, r) = 0.99 \qquad P(w \mid s, \neg r) = 0.90$$

$$P(w \mid \neg s, r) = 0.90 \qquad P(w \mid \neg s, \neg r) = 0.01$$

$$P(m \mid w) = 0.8 \qquad P(m \mid \neg w) = 0.1$$

Fork: $S \leftarrow C \rightarrow R$. Collider: $S \rightarrow W \leftarrow R$. Chain: $R \rightarrow W \rightarrow M$.



C cloudy · S sprinkler · R rain
W wet grass · M muddy shoes
maize = observed · pink = query