Bayes' Rule *

Bayes' rule:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- *P(h)*——prior probability of hypothesis *h*
- *P(D)*——prior probability of training data *D*
- *P(D|h)*——probability of *D* given *h, also called* **likelihood of *D* given *h***
- *P(h|D)*——probability of *h* given *D*

- Useful for assessing diagnostic probability from causal probability:
  - P(Cause|Effect) = P(Effect|Cause) P(Cause) / P(Effect)

Review

Choosing hypotheses *

- ***Maximum a posteriori hypothesis h_{MAP}***

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D)$$

$$= \arg\max_{h \in H} \frac{P(D \mid h) P(h)}{P(D)}$$

$$= \arg\max_{h \in H} P(D \mid h) P(h)$$

- If assume $P(h_i)=P(h_j)$ for h_{MAP}, then can further simplify and choose the ***Maximum Likelihood (ML) hypothesis***

$$h_{ML} = \arg\max_{h \in H} P(D \mid h)$$

Review

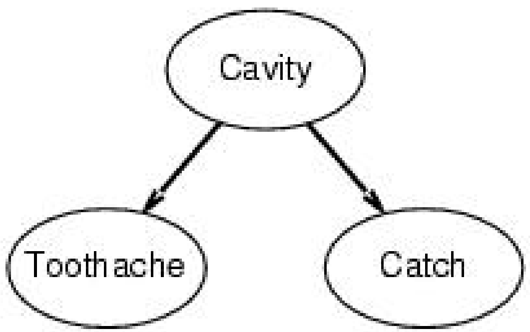
Bayes' Rule and conditional independence

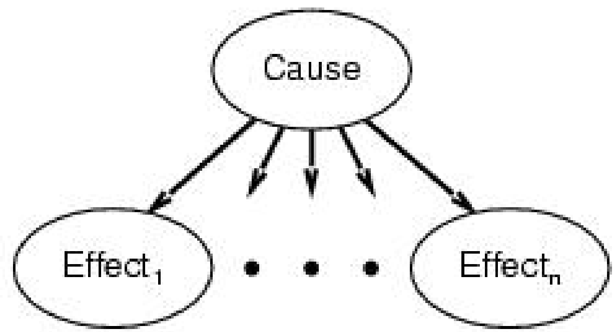
- a naïve Bayes model:

$$\mathbf{P}(\text{Cause},\text{Effect}_1, \ldots ,\text{Effect}_n) = \mathbf{P}(\text{Cause})\ \pi_i \mathbf{P}(\text{Effect}_i|\text{Cause})$$

where Effect_i given Cause are conditional independence

- Total number of parameters is linear in *n*





Example 1b: a medical diagnosis*

Additional Exercise for assignment

- two alternative hypotheses: the patient has a particular form of cancer (cancer), and the patient does not (┐ cancer).
- The available data is from a particular laboratory test with two possible outcomes: positive (⊕) and negative (⊙)
- prior knowledges:
  - over the entire population of people only 0.8% people have this disease
  - The test returns a correct positive result in only 98% of the cases in which the disease is actually present
  - a correct negative result in only 97% of the cases in which the disease is not present
- Question: Consider again the example application. Suppose the doctor decides to order a second laboratory test for the same patient, and suppose the second test returns a positive result as well. Should we diagnose the patient as having cancer or not following these two tests? Assume that the two tests are independent.
- h_{MAP}? Based on P(canner|++) or P(┐ canner|++)

volunteer for example 1b

Bayesian networks

Section 1 – 2, Chapter 14

LI Xiaoan (Dustin) 李孝安, Dr. & Associate Prof.
MARS-lab
School of Computer Science and Engineering
Northwestern Polytechnical University

E-mail: dustinli@nwpu.edu.cn
WeChat group: AI-2021-Li
Cell phone: 18629662731

Outline

- Bayesian Networks
- Bayesian-based Inference
- Bayesian Networks: Classification
- Approaches to Uncertain Reasoning: an overview

Bayesian networks

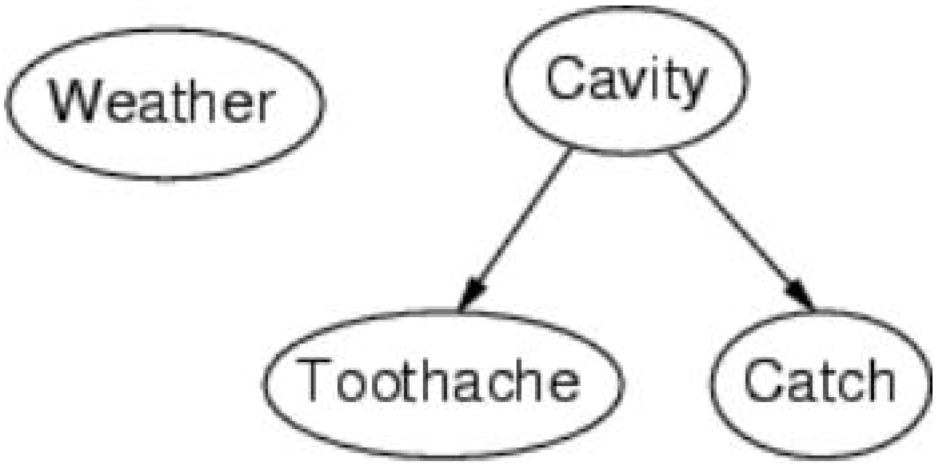
- A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

- Syntax:
  - a set of nodes, one per variable
  - a directed, acyclic graph (link ≈ "directly influences")
  - a conditional distribution for each node given its parents:

$$\mathbf{P}(X_i \mid \text{Parents}(X_i))$$

- In the simplest case, conditional distribution represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values

Example



- Topology of network encodes conditional independence assertions
- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

Example

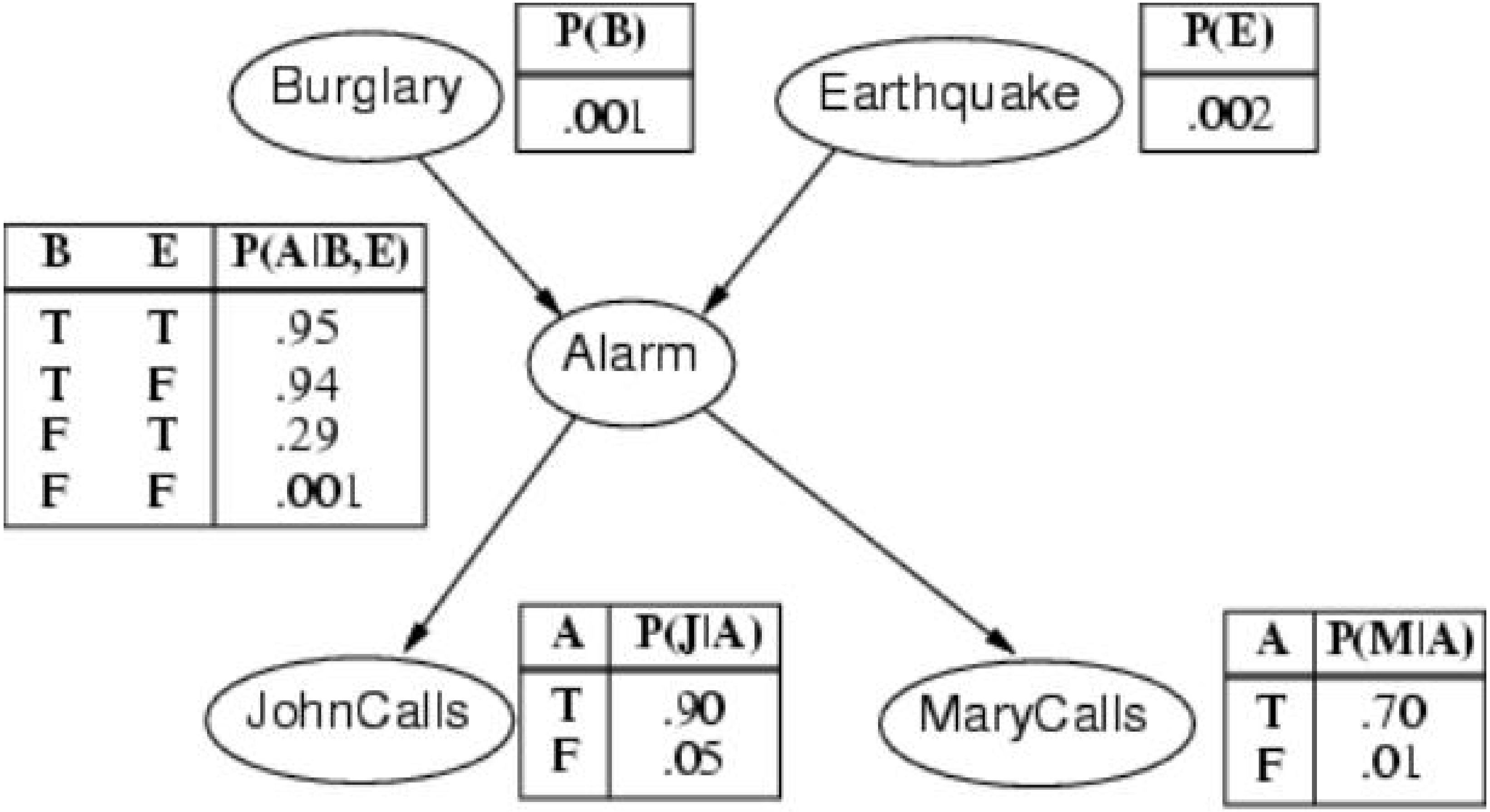
- I'm at work, neighbor John calls to say my alarm is ringing, but neighbor Mary doesn't call. Sometimes it's set off by minor earthquakes.

  Is there a burglar?

- Variables: *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

- Network topology reflects "causal" knowledge:
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

Example contd.



P(B)
.001

P(E)
.002

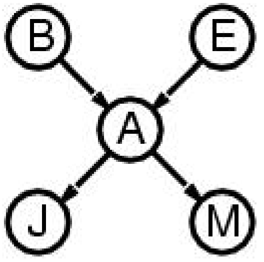
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Compactness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values



- Each row requires one number p for X_i = *true*
(the number for X_i = *false* is just $1-p$)

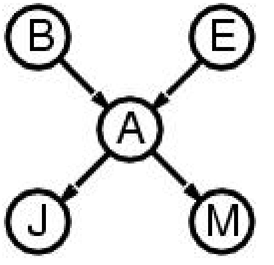
- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers

I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution

- For burglary net, 1 + 1 + 4 + 2 + 2 = 10 numbers (vs. 2^5-1 = 31)

Semantics

The full joint distribution is defined as the product of the local conditional distributions:



$$\boldsymbol{P}\,(X_1,\ \ldots\ ,X_n) = \pi^{n}_{i\,=\,1}\ \boldsymbol{P}\,(X_i \mid Parents(X_i))$$

e.g., $\boldsymbol{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

$$= \boldsymbol{P}\,(j \mid a)\ \boldsymbol{P}\,(m \mid a)\ \boldsymbol{P}\,(a \mid \neg b, \neg e)\ \boldsymbol{P}\,(\neg b)\ \boldsymbol{P}\,(\neg e)$$

Constructing Bayesian networks

- 1. Choose an ordering of variables $X_1, \dots, X_n$
- 2. For i = 1 to n
  - add X_i to the network
  - select parents from $X_1, \dots, X_{i-1}$ such that

$$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \dots X_{i-1})$$

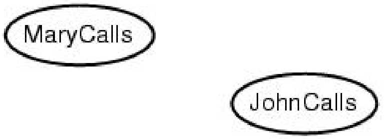
This choice of parents guarantees:

$\mathbf{P}(X_1, \dots, X_n) = \pi_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \dots, X_{i-1})$
(chain rule)

$= \pi_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$
(by construction)

Example

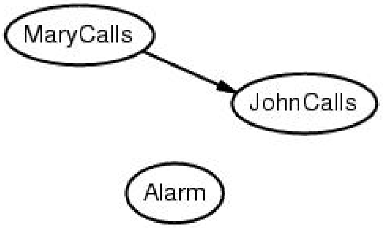
- Suppose we choose the ordering *M, J, A, B, E*



P*(J | M)* = ***P****(J)?*

Example

- Suppose we choose the ordering *M, J, A, B, E*



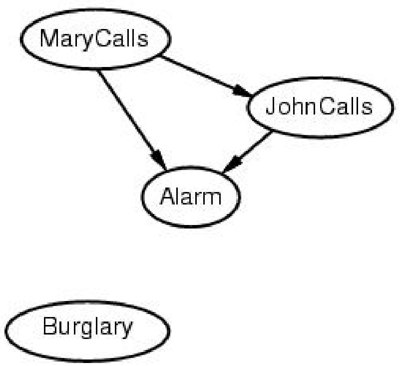
P(J | M) = P(J)?

No

P(A | J, M) = P(A | J)? P(A | J, M) = P(A)?

Example

- Suppose we choose the ordering *M, J, A, B, E*



P(J | M) = P(J)?

No

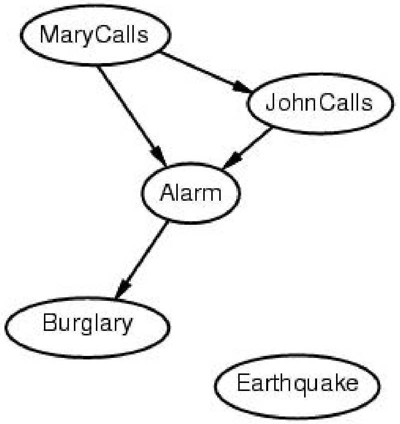
P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? No

P(B | A, J, M) = P(B | A)?

P(B | A, J, M) = P(B)?

Example

- Suppose we choose the ordering M, J, A, B, E



P(J | M) = P(J)?

No

P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? No

P(B | A, J, M) = P(B | A)? Yes

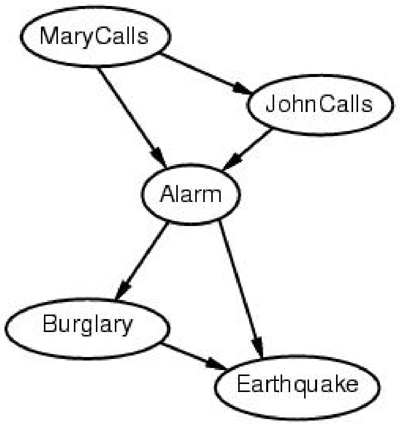
P(B | A, J, M) = P(B)? No

P(E | B, A ,J, M) = P(E | A)?

P(E | B, A, J, M) = P(E | A, B)?

Example

- Suppose we choose the ordering M, J, A, B, E



P(J | M) = P(J)?
No

P(A | J, M) = P(A | J)? P(A | J, M) = P(A)? No

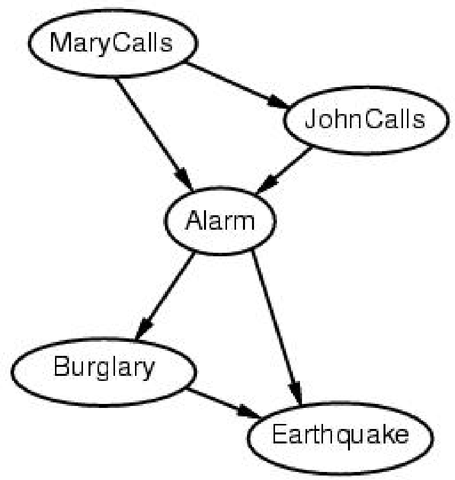
P(B | A, J, M) = P(B | A)? Yes

P(B | A, J, M) = P(B)? No

P(E | B, A ,J, M) = P(E | A)? No

P(E | B, A, J, M) = P(E | A, B)? Yes

Example contd.



- Deciding conditional independence is hard in noncausal directions

(Causal models and conditional independence seem hardwired for humans!)

- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed

Summary

- Bayesian networks provide a natural representation for (causally induced) conditional independence
- Topology + CPTs =

  compact representation of joint distribution
- Generally easy for domain experts to construct
- Comments to Strong AI vs Weak AI

Assignment

- Exercise 14.1
- Additional examples and materials to study in the following pages

Bayesian-based Inference

- Diagnostic Inference
- *Causal inference*

Example 1: Causes and Bayes' Rule

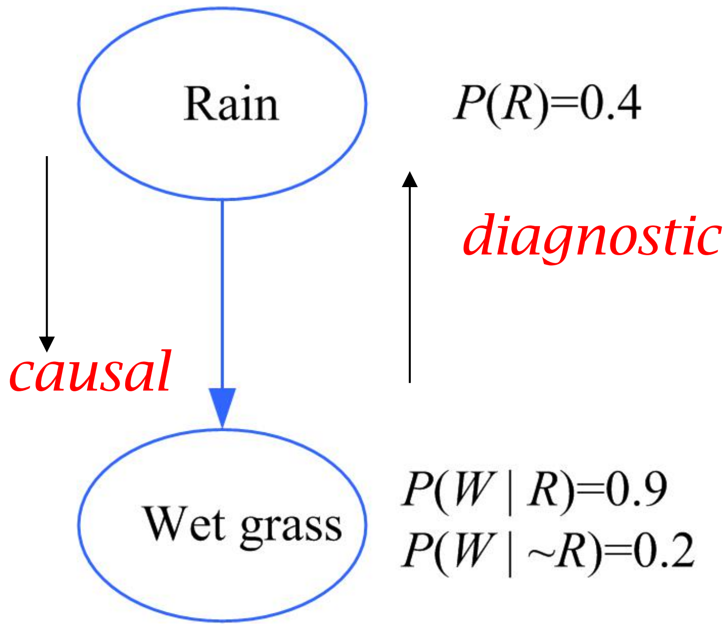


Fig.3-2

Diagnostic inference:
Knowing that the grass is wet, what is the probability that rain is the cause?

$$
\begin{aligned}
P(R \mid W) &= \frac{P(W \mid R)P(R)}{P(W)} \\
&= \frac{P(W \mid R)P(R)}{P(W \mid R)P(R) + P(W \mid \sim R)P(\sim R)} \\
&= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75
\end{aligned}
$$

Example 2: Causal vs Diagnostic Inference

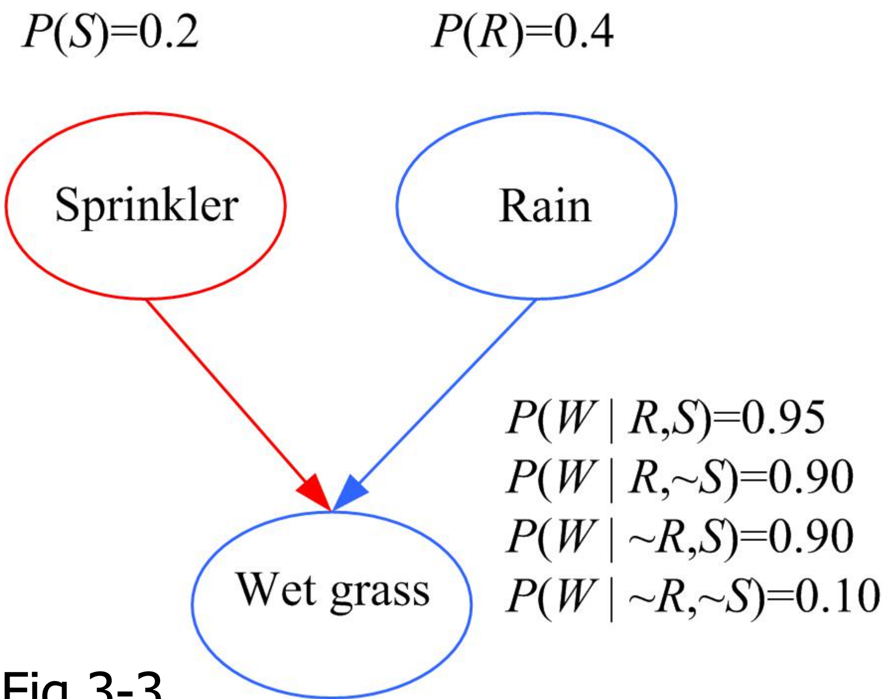


Fig.3-3

Causal inference: *If the sprinkler is on, what is the probability that the grass is wet?*

$$P(W|S) = P(W|R,S)\, P(R|S) + P(W|\sim R,S)\, P(\sim R|S)$$
$$= P(W|R,S)\, P(R) + P(W|\sim R,S)\, P(\sim R)$$
$$= 0.95 \times 0.4 + 0.9 \times 0.6 = 0.92$$

Diagnostic inference: *If the grass is wet, what is the probability that the sprinkler is on, $P(S|W)$? $P(S|R,W)$=?*

Result 1: $P(S|W) = 0.35 > 0.2\ P(S)$

Result2： $P(S|R,W) = 0.21$ ***Explaining away:*** *Knowing that it has Rained decreases the probability that the sprinkler is on.*

Example 3: Bayesian Networks: Causes

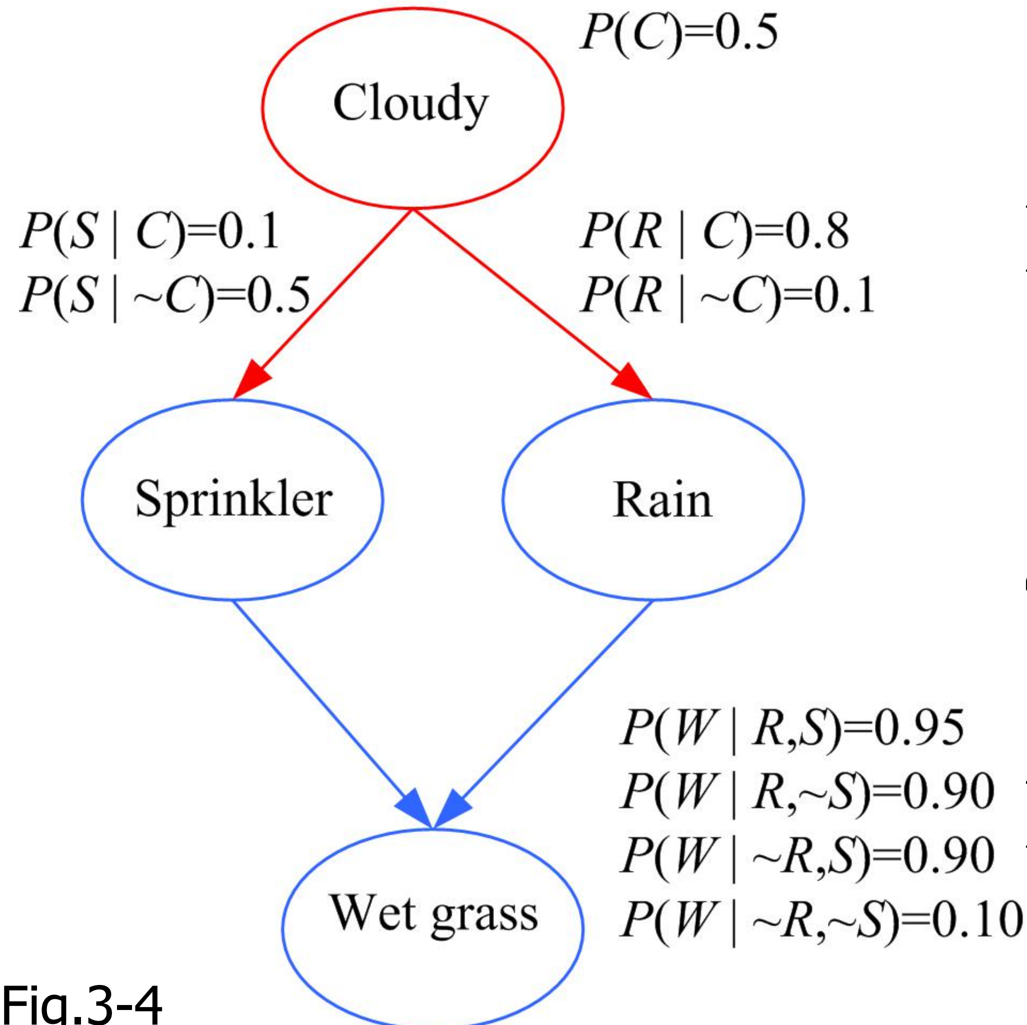


Fig.3-4

Causal inference:

$P(W|C)=?$

$$P(W|C) = P(W|R,S)\,P(R,S|C) + P(W|\sim R,S)\,P(\sim R,S|C) + P(W|R,\sim S)\,P(R,\sim S|C) + P(W|\sim R,\sim S)\,P(\sim R,\sim S|C)$$

and use the fact that

$$P(R,S|C) = P(R|C)\,P(S|C)$$

Diagnostic inference:

$P(C|W) = ?$

Example 4: Bayesian Nets: Local structure

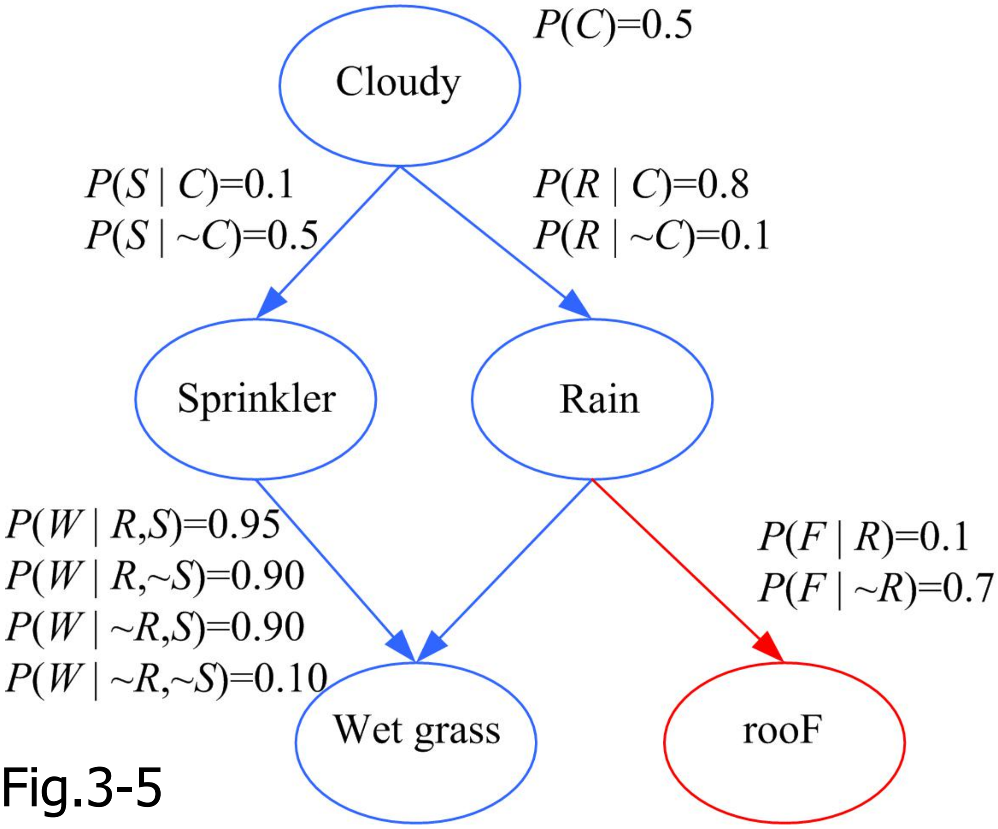


Fig.3-5

$$P(F \mid C) = ?$$

$$P(F \mid S) = ?$$

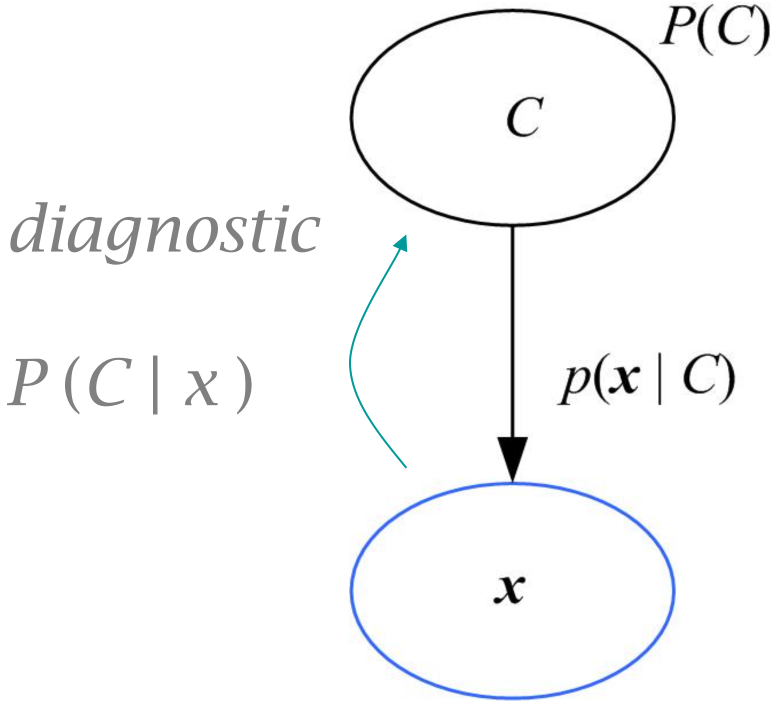
$$P(C,S,R,W,F) = ?$$

$$2^5 - 1 = 31 \, vs \quad 11$$

$$P(C,S,R,W,F) = P(C)P(S \mid C)P(R \mid C)P(W \mid S,R)P(F \mid R)$$

$$P(X_1, \dots X_d) = \prod_{i=1}^{d} P(X_i \mid \text{parents}(X_i))$$

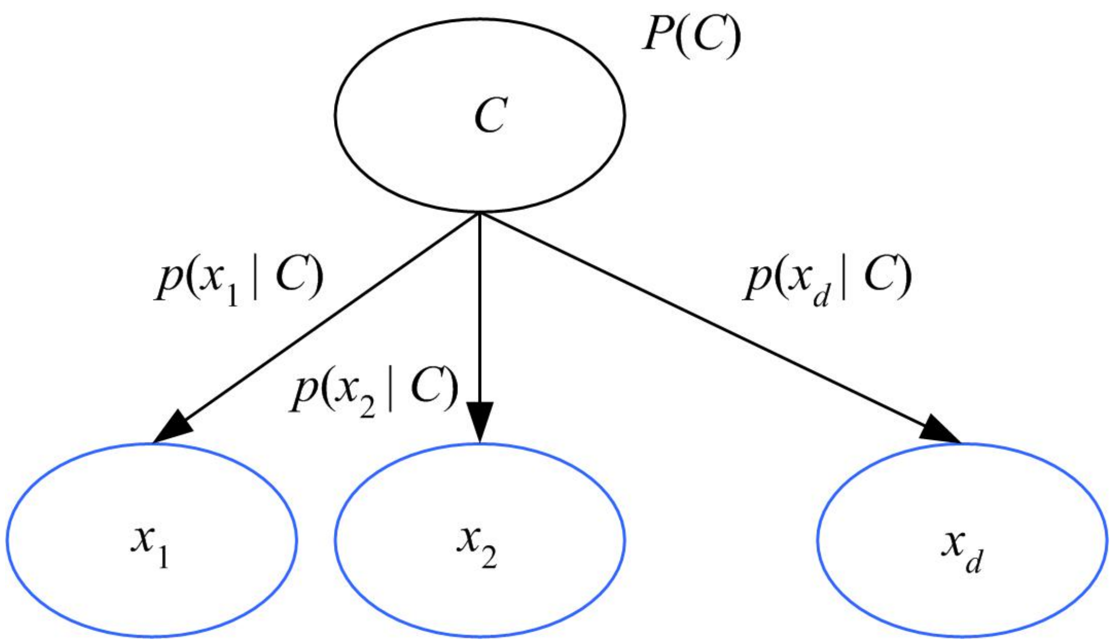
Bayesian Networks: Classification



Bayes' rule inverts the arc:

$$P(C \mid \boldsymbol{x}) = \frac{p(\boldsymbol{x} \mid C)P(C)}{p(\boldsymbol{x})}$$

Naive Bayes' Classifier *



Given C, x_j are independent:

$$p(\boldsymbol{x}|C) = p(x_1|C)\ p(x_2|C)\ \ldots\ p(x_d|C)$$

Approaches to Uncertain Reasoning: an overview

- Bayesian-based Inference
  - *Diagnostic inference*
  - *Causal inference*
- Other approaches
- Fuzzy sets and Fuzzy logic
  - Fuzzy set theory: A means of specifying how well an object satisfies a vague description
  - Fuzzy logic: a method for reasoning with logical expression describing membership in fuzzy sets.
- HMM
- ANN