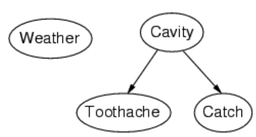
Bayesian Networks

Motivation

- The conditional independence assumption made by naïve Bayes classifiers may seem too rigid, especially for classification problems in which the attributes are somewhat correlated.
- We talk today about a more flexible approach for modeling the conditional probabilities.

Bayesian networks

- A simple, graphical notation for conditional independence assertions.



Syntax:

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

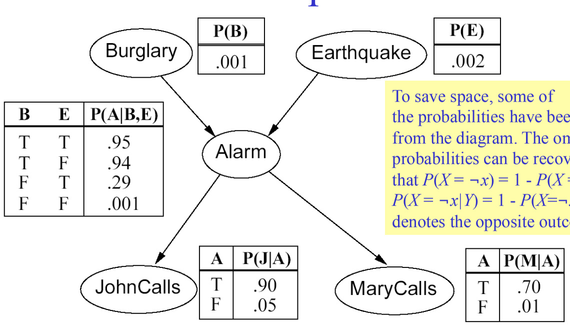
  P (X_i | Parents (X_i))

- The conditional distribution is represented as a conditional probability table (CPT) giving the distribution over X_i for each combination of parent values.

Example (Perls' 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?
- John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm.
- Mary likes rather loud music and sometimes misses the alarm.
- 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 cont'd



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

To save space, some of the probabilities have been omitted from the diagram. The omitted probabilities can be recovered by noting that $P(X = \neg x) = 1 - P(X = x)$ and $P(X = \neg x|Y) = 1 - P(X=\neg x|Y)$, where $\neg x$ denotes the opposite outcome of x.

The topology shows that burglary and earthquakes directly affect the probability of alarm, but whether Mary or John call depends only on the alarm.

Thus our assumptions are that they don't perceive any burglaries directly, and they don't confer before calling.

Inference in Bayesian Networks

- The basic task for a probabilistic inference system is to compute the conditional probability for a **query variable (class attribute)**, given some observed **events**
  - that is, some assignment of values to a set of **evidence variables (some of the other attributes)**.
- Notation:
  - X denotes query variable
  - **E** denotes the *set* of evidence variables $E_1,\ldots,E_m$, and **e** is a particular event, i.e. an assignment to the variables in **E**.
  - **Y** will denote the set of the remaining variables (hidden variables).
- A typical query asks for the posterior probability $P(x|e_1,\ldots,e_m)$
- E.g. We could ask: What's the probability of a burglary if both Mary and John call, *P*(*burglary* | *johhcalls*, *marycalls*)?

Classification

- Suppose, we are given for the evidence variables $E_1,\ldots,E_m$, their values $e_1,\ldots,e_m$, and we want to predict whether the query variable X has the value x or not.
- For this we compute and compare the following:

$$P(x \mid e_1,\ldots,e_m) = \frac{P(x,e_1,\ldots,e_m)}{P(e_1,\ldots,e_m)} = \alpha P(x,e_1,\ldots,e_m)$$

$$P(\neg x \mid e_1,\ldots,e_m) = \frac{P(\neg x,e_1,\ldots,e_m)}{P(e_1,\ldots,e_m)} = \alpha P(\neg x,e_1,\ldots,e_m)$$

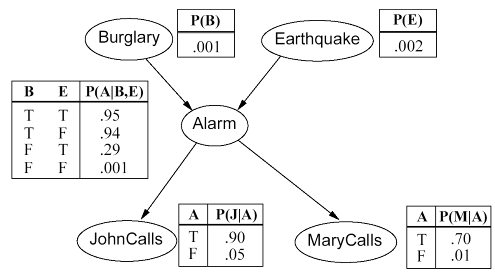
- However, how do we compute:

$$\alpha P(x,e_1,\ldots,e_m)$$
and
$$\alpha P(\neg x,e_1,\ldots,e_m)?$$

What about the hidden variables $Y_1,\ldots,Y_k$?

Conditional Independence

- In a Bayes Net
  - Each variable is **conditionally independent** of all its non-descendants in the graph given the value of all its parents.
  - P(John calls | Alarm, Burglary) = P(John calls | Alarm)



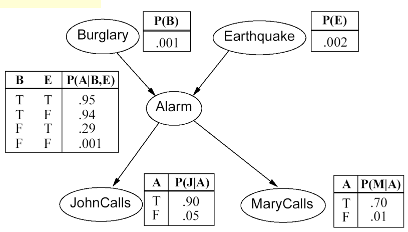
Semantics

Suppose we have the variables (attr.) $X_1,\ldots,X_n$, sorted in a topological order.
The probability for them to have the values $x_1,\ldots,x_n$ respectively is $P(x_n,\ldots,x_1)$:

$= P(x_n,\ldots,x_1)$
$= P(x_n \mid x_{n-1},\ldots,x_1)P(x_{n-1},\ldots,x_1)$
$= P(x_n \mid x_{n-1},\ldots,x_1)P(x_{n-1} \mid x_{n-2},\ldots,x_1)P(x_{n-2},\ldots,x_1)$
$= \ldots$
$= \prod_{i=1}^{n} P(x_i \mid x_{i-1},\ldots,x_1) = \prod_{i=1}^{n} P(x_i \mid parents(x_i))$

$P(x_n,\ldots,x_1)$:
is short for
$P(X_n{=}x_n,\ldots, X_n{=}\, x_1)$:

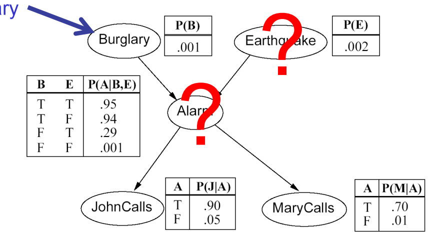
e.g.,
$P(j, m, a, \neg b, \neg e)$
$=P(j| m, a, \neg b, \neg e)$
$\quad *P(m| a, \neg b, \neg e)$
$\quad *P(a| \neg b, \neg e) *P(\neg b| \neg e)$
$\quad *P(\neg e)$
=P(j| a)*P(m| a)*$P(a| \neg b, \neg e)$
$\quad *P(\neg b) *P(\neg e)$



Hidden Vars

- But we don't get to observe if there was an alarm or an earthquake, we just know if John or Mary calls.

Trying to predict burglary



Inference by enumeration

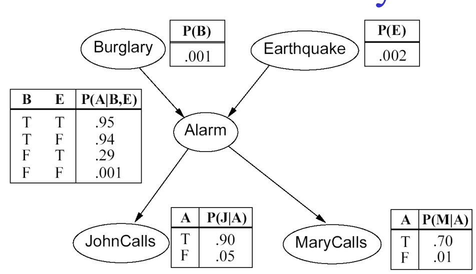
Example: *P(burglary | johhcalls, marycalls)*? (Abbrev. $P(b|j,m)$)

$P(b \mid j,m)$
$= \alpha P(b,j,m)$
$= \alpha \sum_a \sum_e P(b,j,m,a,e)$
$= \alpha\big(P(b,j,m,a,e) + P(b,j,m,\neg a,e) + P(b,j,m,a,\neg e) + P(b,j,m,\neg a,\neg e)\big)$

In general with hidden vars:

$$P(x \mid e_1,\ldots,e_m) = \alpha P(x,e_1,\ldots,e_m) = \alpha \sum_{y_1} \cdots \sum_{y_k} P(x,e_1,\ldots,e_m,y_1,\ldots,y_k)$$
and
$$P(\neg x \mid e_1,\ldots,e_m) = \alpha P(\neg x,e_1,\ldots,e_m) = \alpha \sum_{y_1} \cdots \sum_{y_k} P(\neg x,e_1,\ldots,e_m,y_1,\ldots,y_k)$$

Numerically…



$P(b \mid j,m) = \alpha\, P(b) \sum_a P(j|a)P(m|a)\sum_e P(a|b,e)P(e) = \ldots = \alpha * 0.00059$
$P(\neg b \mid j,m) = \alpha\, P(\neg b) \sum_a P(j|a)P(m|a)\sum_e P(a| \neg b,e)P(e) = \ldots = \alpha * 0.0015$

P(B | j,m) = α <0.00059, 0.0015> = **<0.28, 0.72>**.

Details of $P(b \mid j,m)$

$P(b \mid j,m) = \alpha\, P(b) \sum_a P(j|a)P(m|a)\sum_e P(a|b,e)P(e)$

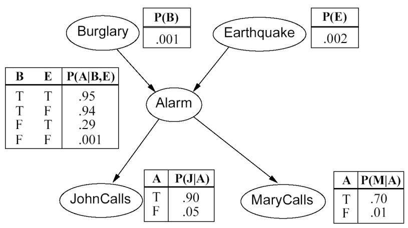
$= \alpha\, P(b) \sum_a P(j|a)P(m|a)(P(a|b,e)P(e) + P(a|b,\neg e)P(\neg e))$

$= \alpha\, P(b)(\ P(j|a)P(m|a)(\ P(a|b,e)P(e) + P(a|b,\neg e)P(\neg e)\ )$

$\quad + P(j|\neg a)P(m|\neg a)(\ P(\neg a|b,e)P(e) + P(\neg a|b,\neg e)P(\neg e)\ ))$

$= \alpha$ * .001*(.9*.7*(.95*.002 + .94*.998) +.05*.01*(.05*.002 + .71*.998))

$= \alpha$ * .00059



Details of $P(\neg b \mid j,m)$

$P(\neg b \mid j,m) = \alpha\, P(\neg b) \sum_a P(j|a)P(m|a)\sum_e P(a|\neg b,e)P(e)$

$= \alpha\, P(\neg b) \sum_a P(j|a)P(m|a)(P(a|\neg b,e)P(e) + P(a|\neg b,\neg e)P(\neg e))$

$= \alpha\, P(\neg b)(\ P(j|a)P(m|a)(\ P(a|\neg b,e)P(e) + P(a|\neg b,\neg e)P(\neg e)\ )$

$\quad + P(j|\neg a)P(m|\neg a)(\ P(\neg a|\neg b,e)P(e) + P(\neg a|\neg b,\neg e)P(\neg e)\ ))$

$= \alpha$ * .999*(.9*.7*(.29*.002 + .001*.998) +.05*.01*(.71*.002 + .999*.998))

$= \alpha$ * .0015

α = 1/(.00059 + .0015)

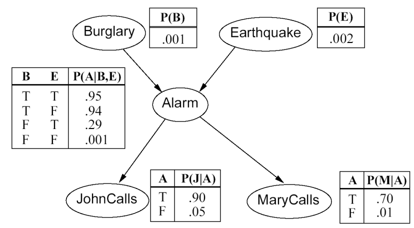
= 478.5

$P(b \mid j,m)$ = 478.5 * .00059

=.28

$P(\neg b \mid j,m)$ = 478.5 * .0015

=.72



Constructing Bayesian networks

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

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

This choice of parents guarantees:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1}) \quad \text{(chain rule)}$$
$$= \prod_{i=1} \mathbf{P}(X_i \mid Parents(X_i)) \quad \text{(by construction)}$$

- Choosing the parents from $X_1, \ldots, X_{i-1}$ is done by domain human experts.

Bayes Net vs Naïve Bayes

- Bayes Net
  - Pros:
    - Model could be a better fit to the data
    - Fewer params than modeling the full joint distribution
  - Cons
    - A person needs to make the graph
    - Code needs to either be custom for the graph/problem, or able to read in graph structure
- Naïve Bayes
  - $P(\neg b \mid j,m)$ Let's say alarm and earthquake are missing variables
    $= \alpha P(j \mid \neg b) * P(m \mid \neg b) * P(\neg b)$
  - No hand made graph
  - But data may not respect the independence assumption
  - Draw the graph for Naïve Bayes