Tutorial on Bayesian Networks

Jack Breese & Daphne Koller

First given as a AAAI'97 tutorial.

Overview

- Decision-theoretic techniques
  - Explicit management of uncertainty and tradeoffs
  - Probability theory
  - Maximization of expected utility
- Applications to AI problems
  - Diagnosis
  - Expert systems
  - Planning
  - Learning

Science- AAAI-97

- Model Minimization in Markov Decision Processes
- Effective Bayesian Inference for Stochastic Programs
- Learning Bayesian Networks from Incomplete Data
- Summarizing CSP Hardness With Continuous Probability Distributions
- Speeding Safely: Multi-criteria Optimization in Probabilistic Planning
- Structured Solution Methods for Non-Markovian Decision Processes

Applications

@COMPUTERWORLD
The online connection for information technology leaders

Microsoft's cost-cutting helps users

04/21/97

A Microsoft Corp. strategy to cut its support costs by letting users solve their own problems using electronic means is paying off for users.In March, the company began rolling out a series of Troubleshooting Wizards on its World Wide Web site.

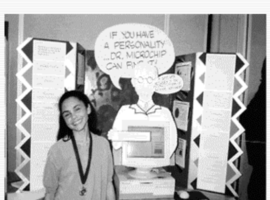
Troubleshooting Wizards save time and money for users who don't have Windows NT specialists on hand at all times, said Paul Soares, vice president and general manager of Alden Buick Pontiac, a General Motors Corp. car dealership in Fairhaven, Mass

Teenage Bayes

Microsoft Researchers Exchange Brainpower with Eighth-grader

Teenager Designs Award-Winning Science Project

.. For her science project, which she called "Dr. Sigmund Microchip," Tovar wanted to create a computer program to diagnose the probability of certain personality types. With only answers from a few questions, the program was able to accurately diagnose the correct personality type 90 percent of the time.



Elena Tovar stands proudly in front of "Dr. Sigmund Microchip," the science project she created using the advanced mathematical formulas that Microsoft Research uses to build artificial intelligence programs.

Course Contents

- » Concepts in Probability
  - Probability
  - Random variables
  - Basic properties (Bayes rule)
- Bayesian Networks
- Inference
- Decision making
- Learning networks from data
- Reasoning over time
- Applications

Probabilities

- Probability distribution $P(X|\xi)$
  - X is a random variable
    - Discrete
    - Continuous
  - ξ is background state of information

7

Discrete Random Variables

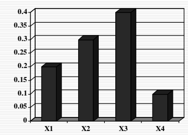
- Finite set of possible outcomes

$$X \in \{x_1, x_2, x_3, \ldots, x_n\}$$

$$P(x_i) \geq 0$$

$$\sum_{i=1}^{n} P(x_i) = 1$$

X binary: $P(x) + P(\overline{x}) = 1$



8

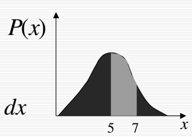
Continuous Random Variable

- Probability distribution (density function) over continuous values

$X \in [0,10]$ $P(x) \geq 0$

$$\int_0^{10} P(x)\,dx = 1$$

$$P(5 \leq x \leq 7) = \int_5^7 P(x)\,dx$$



9

More Probabilities

- Joint

$$P(x, y) \equiv P(X = x \wedge Y = y)$$

  - Probability that both X=x and Y=y
- Conditional

$$P(x \mid y) \equiv P(X = x \mid Y = y)$$

  - Probability that X=x given we know that Y=y

10

Rules of Probability

- Product Rule

$$P(X,Y) = P(X \mid Y)P(Y) = P(Y \mid X)P(X)$$

- Marginalization

$$P(Y) = \sum_{i=1}^{n} P(Y, x_i)$$

X binary: $P(Y) = P(Y,x) + P(Y,\overline{x})$

11

Bayes Rule

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

$$\boxed{P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}}$$

12

Course Contents

- Concepts in Probability
- » Bayesian Networks
  - Basics
  - Additional structure
  - Knowledge acquisition
- Inference
- Decision making
- Learning networks from data
- Reasoning over time
- Applications

13

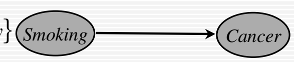
Bayesian networks

- Basics
  - Structured representation
  - Conditional independence
  - Naïve Bayes model
  - Independence facts

14

Bayesian Networks

$S \in \{no, light, heavy\}$



$C \in \{none, benign, malignant\}$

P(S=no)	0.80
P(S=light)	0.15
P(S=heavy)	0.05

Smoking=	*no*	*light*	*heavy*
P(C=none)	0.96	0.88	0.60
P(C=benign)	0.03	0.08	0.25
P(C=malig)	0.01	0.04	0.15

15

Product Rule

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

S⇓ C⇒	*none*	*benign*	*malignant*
no	0.768	0.024	0.008
light	0.132	0.012	0.006
heavy	0.035	0.010	0.005

16

Marginalization

S⇓ C⇒	*none*	*benign*	*malig*	total
no	0.768	0.024	0.008	.80
light	0.132	0.012	0.006	.15
heavy	0.035	0.010	0.005	.05
total	0.935	0.046	0.019	

P(Smoke)

P(Cancer)

17

Bayes Rule Revisited

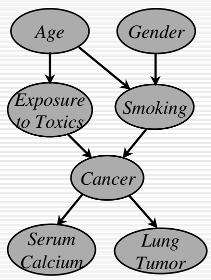
$$P(S|C) = \frac{P(C|S)P(S)}{P(C)} = \frac{P(C,S)}{P(C)}$$

S⇓ C⇒	*none*	*benign*	*malig*
no	0.768/.935	0.024/.046	0.008/.019
light	0.132/.935	0.012/.046	0.006/.019
heavy	0.030/.935	0.015/.046	0.005/.019

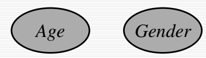
Cancer=	*none*	*benign*	*malignant*
P(S=no)	0.821	0.522	0.421
P(S=light)	0.141	0.261	0.316
P(S=heavy)	0.037	0.217	0.263

18

A Bayesian Network



Independence



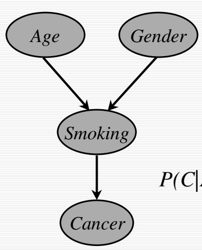
Age and *Gender* are independent.

$P(A,G) = P(G)P(A)$

$P(A|G) = P(A) \quad A \perp G$
$P(G|A) = P(G) \quad G \perp A$

$P(A,G) = P(G|A)\, P(A) = P(G)P(A)$
$P(A,G) = P(A|G)\, P(G) = P(A)P(G)$

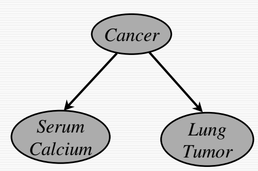
Conditional Independence



Cancer is independent of *Age* and *Gender* given *Smoking*.

$P(C|A,G,S) = P(C|S) \quad C \perp A,G \mid S$

More Conditional Independence: Naïve Bayes

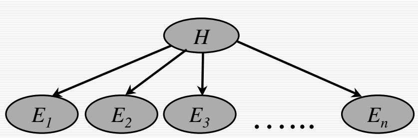


Serum Calcium and *Lung Tumor* are dependent

Serum Calcium is independent of *Lung Tumor*, given *Cancer*

$P(L|SC,C) = P(L|C)$

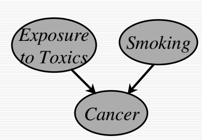
Naïve Bayes in general



2n + 1 parameters: $P(h)$

$P(e_i \mid h), P(e_i \mid \bar{h}), \ i = 1, \ldots, n$

More Conditional Independence: Explaining Away



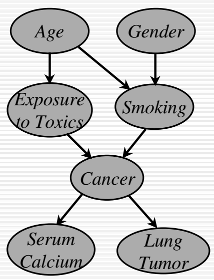
Exposure to Toxics and *Smoking* are independent

$E \perp S$

Exposure to Toxics is **dependent** on *Smoking*, given *Cancer*

$P(E = heavy \mid C = malignant) >$
$P(E = heavy \mid C = malignant, S = heavy)$

Put it all together



$$P(A,G,E,S,C,L,SC) =$$
$$P(A) \cdot P(G) \cdot$$
$$P(E \mid A) \cdot P(S \mid A,G) \cdot$$
$$P(C \mid E,S) \cdot$$
$$P(SC \mid C) \cdot P(L \mid C)$$

25

General Product (Chain) Rule for Bayesian Networks

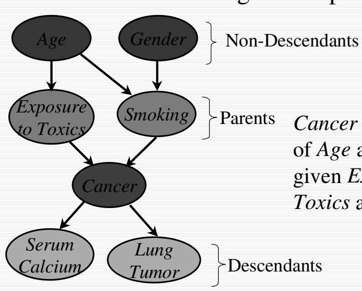
$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \boldsymbol{Pa_i})$$

$\boldsymbol{Pa_i}$=*parents*(X_i)

26

Conditional Independence

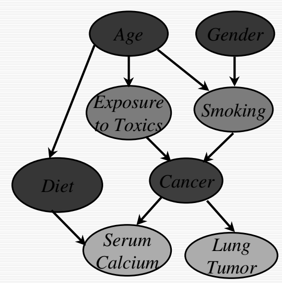
A variable (node) is conditionally independent of its non-descendants given its parents.



Cancer is independent of *Age* and *Gender* given *Exposure to Toxics* and *Smoking*.

27

Another non-descendant



Cancer is independent of *Diet* given *Exposure to Toxics* and *Smoking*.

28

Independence and Graph Separation

- Given a set of observations, is one set of variables dependent on another set?
- Observing effects can induce dependencies.
- d-separation (Pearl 1988) allows us to check conditional independence graphically.

29

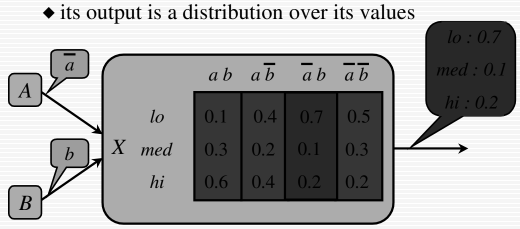
Bayesian networks

- Additional structure
  - Nodes as functions
  - Causal independence
  - Context specific dependencies
  - Continuous variables
  - Hierarchy and model construction

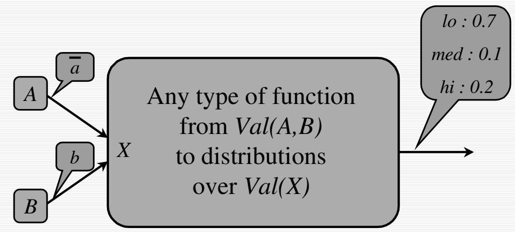
30

Nodes as functions

- A BN node is conditional distribution function
  - its parent values are the inputs
  - its output is a distribution over its values

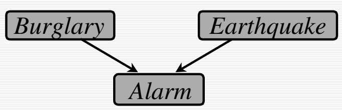


31



32

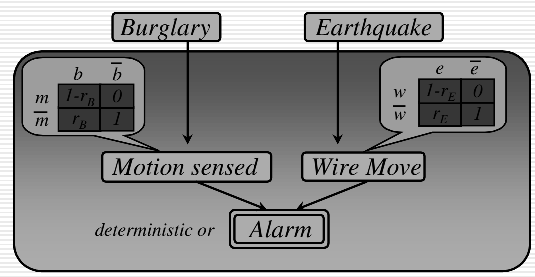
Causal Independence



- *Burglary* causes *Alarm* iff motion sensor clear
- *Earthquake* causes *Alarm* iff wire loose
- Enabling factors are independent of each other

33

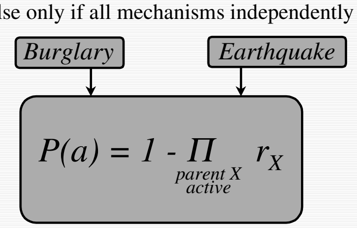
Fine-grained model



34

Noisy-Or model

Alarm false only if all mechanisms independently inhibited

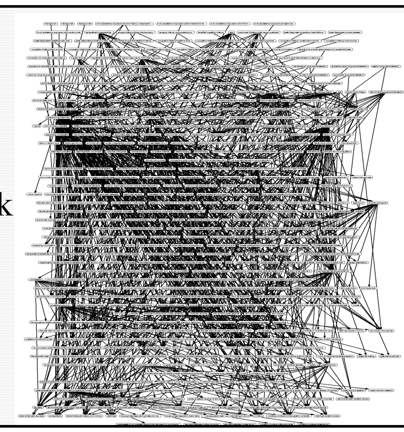


$$P(a) = 1 - \prod_{\substack{\text{parent } X \\ \text{active}}} r_X$$

# of parameters is linear in the # of parents

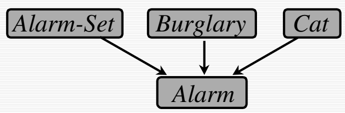
35

CPCS Network



36

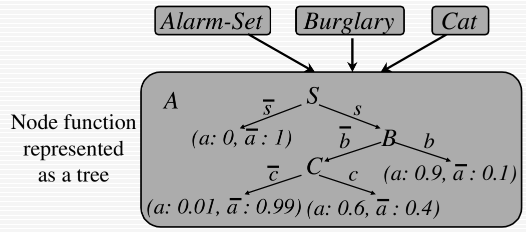
Context-specific Dependencies



- *Alarm* can go off only if it is *Set*
- A burglar and the cat can both set off the alarm
- If a burglar comes in, the cat hides and does not set off the alarm

37

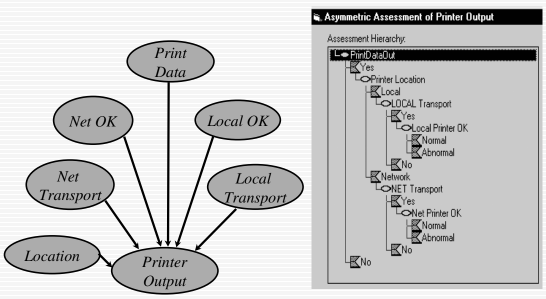
Asymmetric dependencies



- *Alarm* independent of
  - *Burglary*, *Cat* given $\bar{s}$
  - *Cat* given *s* and *b*

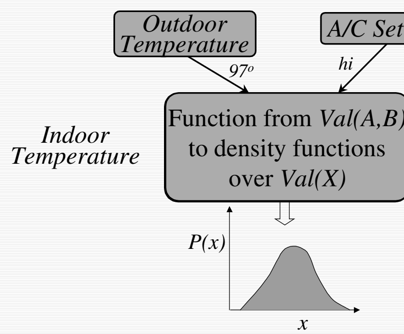
38

Asymmetric Assessment



39

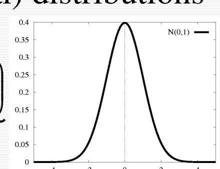
Continuous variables

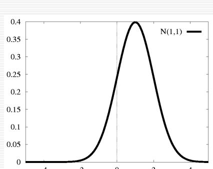


40

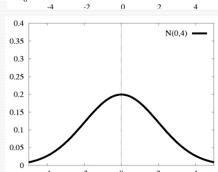
Gaussian (normal) distributions

$$\underbrace{P(x) = \frac{1}{\sqrt{2\pi}\,\sigma}\exp\left(\frac{-(x-\mu)^2}{2\sigma}\right)}_{N(\mu,\ \sigma)}$$





different mean

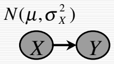


different variance

41

Gaussian networks

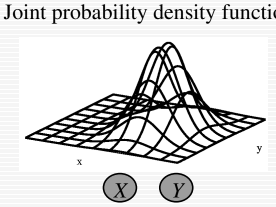
$X \sim N(\mu, \sigma_X^2)$

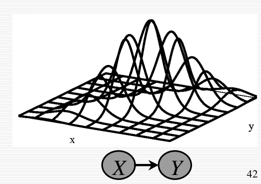


$Y \sim N(ax+b, \sigma_Y^2)$

Each variable is a linear function of its parents, with Gaussian noise

Joint probability density functions:



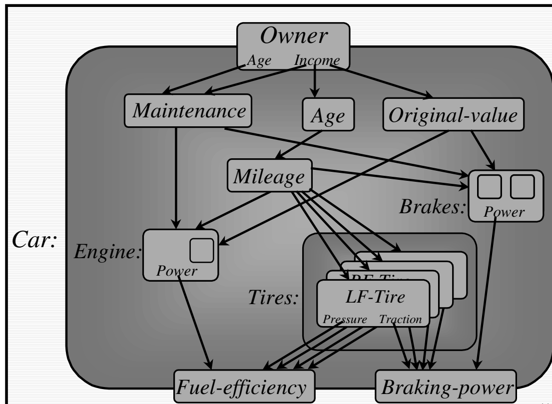


42

Composing functions

- Recall: a BN node is a function
- We can compose functions to get more complex functions.
- The result: A hierarchically structured BN.
- Since functions can be called more than once, we can reuse a BN model fragment in multiple contexts.

43



44

Bayesian Networks

- Knowledge acquisition
  - Variables
  - Structure
  - Numbers

45

What is a variable?

- Collectively exhaustive, mutually exclusive values

$$x_1 \vee x_2 \vee x_3 \vee x_4$$

$$\neg(x_i \wedge x_j) \quad i \neq j$$

~~Error Occured~~

~~No Error~~

- Values versus Probabilities

~~Risk of Smoking~~ Smoking

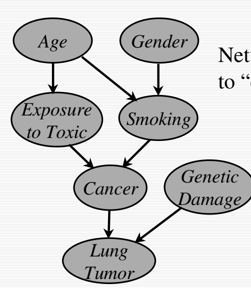
46

Clarity Test: Knowable in Principle

- Weather {Sunny, Cloudy, Rain, Snow}
- Gasoline: Cents per gallon
- Temperature { ≥ 100F , < 100F}
- User needs help on Excel Charting {Yes, No}
- User's personality {dominant, submissive}

47

Structuring



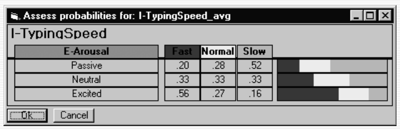
Network structure corresponding to "causality" is usually good.

Extending the conversation.

48

Do the numbers really matter?

- Second decimal usually does not matter
- Relative Probabilities


Assess probabilities for: I-TypingSpeed_avg

I-TypingSpeed

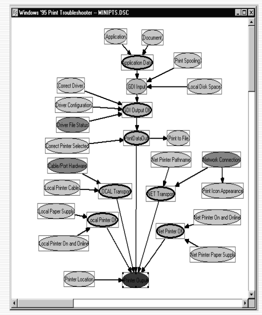
E-Arousal	Fast	Normal	Slow
Passive	.20	.28	.52
Neutral	.33	.33	.33
Excited	.56	.27	.16

Ok Cancel

- Zeros and Ones
- Order of Magnitude : 10^{-9} vs 10^{-6}
- Sensitivity Analysis

49

Local Structure



- Causal independence: from 2^n to $n+1$ parameters
- Asymmetric assessment: similar savings in practice.
- Typical savings (#params):
  - 145 to 55 for a small hardware network;
  - 133,931,430 to 8254 for CPCS !!

50

Course Contents

- Concepts in Probability
- Bayesian Networks
- » Inference
- Decision making
- Learning networks from data
- Reasoning over time
- Applications

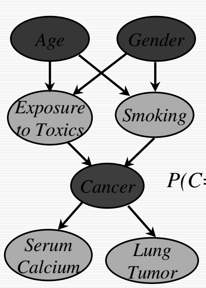
51

Inference

- Patterns of reasoning
- Basic inference
- Exact inference
- Exploiting structure
- Approximate inference

52

Predictive Inference

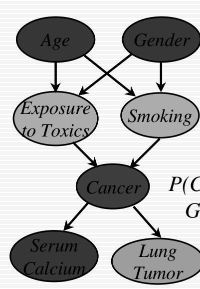


How likely are elderly males to get malignant cancer?

P(C=malignant | Age>60, Gender= male)

53

Combined

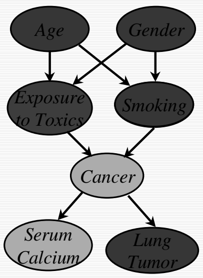


How likely is an elderly male patient with high Serum Calcium to have malignant cancer?

P(C=malignant | Age>60, Gender= male, Serum Calcium = high)

54

Explaining away



- If we see a lung tumor, the probability of heavy smoking and of exposure to toxics both go up.
- If we then observe heavy smoking, the probability of exposure to toxics goes back down.

55

Inference in Belief Networks

- Find $P(Q{=}q|\boldsymbol{E}{=}\,\boldsymbol{e})$
  - Q the query variable
  - $\boldsymbol{E}$ set of evidence variables

$$P(q \mid \boldsymbol{e}) = \frac{P(q, \boldsymbol{e})}{P(\boldsymbol{e})}$$

$X_1, \dots, X_n$ are network variables except Q, $\boldsymbol{E}$

$$P(q, \boldsymbol{e}) = \sum_{x_1, \dots, x_n} P(q, e, x_1, \dots, x_n)$$

56

Basic Inference

$A \rightarrow B$

$P(b) = ?$

57

Product Rule

$S \rightarrow C$

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

S⇓ *C*⇒	*none*	*benign*	*malignant*
no	0.768	0.024	0.008
light	0.132	0.012	0.006
heavy	0.035	0.010	0.005

58

Marginalization

S⇓ *C*⇒	*none*	*benign*	*malig*	total
no	0.768	0.024	0.008	.80
light	0.132	0.012	0.006	.15
heavy	0.035	0.010	0.005	.05
total	0.935	0.046	0.019	

P(Smoke)

P(Cancer)

59

Basic Inference

$A \rightarrow B \rightarrow C$

$$P(b) = \sum_a P(a, b) = \sum_a P(b \mid a)\, P(a)$$

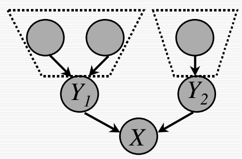
$$P(c) = \sum_b P(c \mid b)\, P(b)$$

$$P(c) = \sum_{b,a} P(a, b, c) = \sum_{b,a} P(c \mid b)\, P(b \mid a)\, P(a)$$

$$= \sum_b P(c \mid b) \underbrace{\sum_a P(b \mid a)\, P(a)}_{P(b)}$$

60

Inference in trees



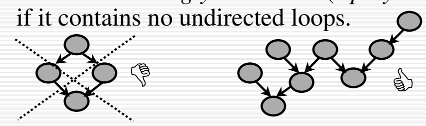
$$P(x) = \sum_{y_1, y_2} P(x \mid y_1, y_2)\, P(y_1, y_2)$$

because of independence of Y_1, Y_2:

$$= \sum_{y_1, y_2} P(x \mid y_1, y_2)\, P(y_1)\, P(y_2)$$

Polytrees

- A network is *singly connected* (a *polytree*) if it contains no undirected loops.

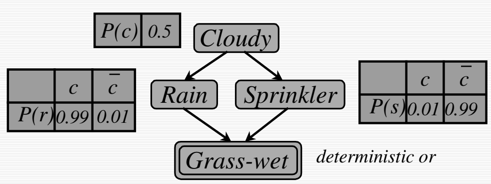


Theorem: Inference in a singly connected network can be done in linear time*.

Main idea: in variable elimination, need only maintain distributions over single nodes.

* in network size including table sizes.

The problem with loops



The grass is dry only if no rain and no sprinklers.

$$P(\bar{g}) = P(\bar{r}, \bar{s}) \sim 0$$

The problem with loops contd.

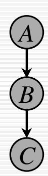
$$P(\bar{g}) = \overbrace{P(\bar{g} \mid r, s)}^{0} P(r, s) + \overbrace{P(\bar{g} \mid r, \bar{s})}^{0} P(r, \bar{s})$$
$$+ \underbrace{P(\bar{g} \mid \bar{r}, s)}_{0} P(\bar{r}, s) + \underbrace{P(\bar{g} \mid \bar{r}, \bar{s})}_{1} P(\bar{r}, \bar{s})$$

$$= P(\bar{r}, \bar{s}) \sim 0$$

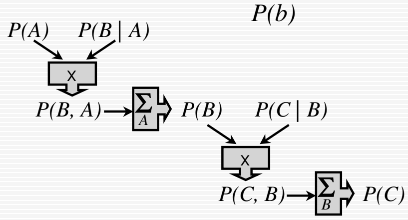
$$\neq P(\bar{r})\, P(\bar{s}) \sim 0.5 \cdot 0.5 = 0.25$$

problem

Variable elimination



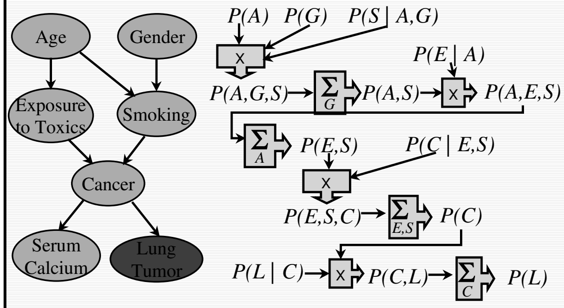
$$P(c) = \sum_b P(c \mid b) \underbrace{\sum_a P(b \mid a)\, P(a)}_{P(b)}$$



Inference as variable elimination

- A **factor** over ***X*** is a function from *val(**X**)* to numbers in [0,1]:
  - A CPT is a factor
  - A joint distribution is also a factor
- BN inference:
  - factors are multiplied to give new ones
  - variables in factors summed out
- A variable can be summed out as soon as all factors mentioning it have been multiplied.

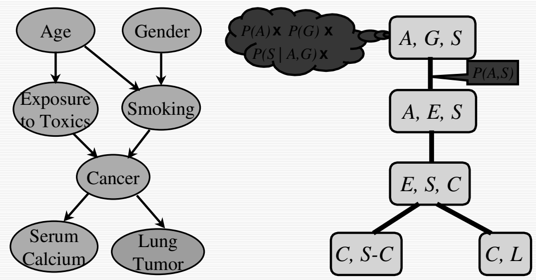
Variable Elimination with loops



Complexity is exponential in the size of the factors

Join trees*

A join tree is a partially precompiled factorization

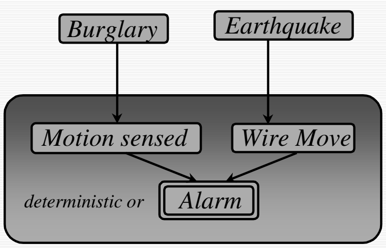


* aka junction trees, Lauritzen-Spiegelhalter, Hugin alg., …

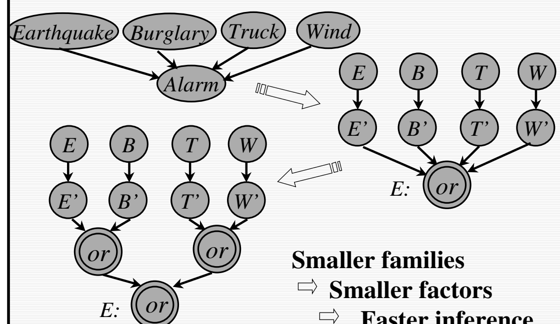
Exploiting Structure

Idea: explicitly decompose nodes

Noisy or:



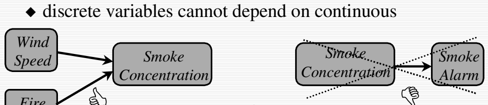
Noisy-or decomposition



Smaller families
⇨ **Smaller factors**
⇨ **Faster inference**

Inference with continuous variables

- Gaussian networks: polynomial time inference regardless of network structure
- Conditional Gaussians:
  - discrete variables cannot depend on continuous



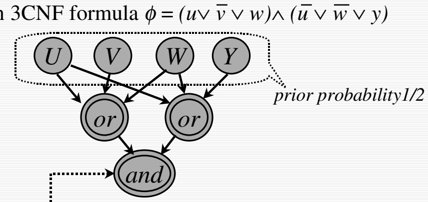
$S \sim N(a_F w + b_F, \sigma_F^2)$

- These techniques do not work for general hybrid networks.

Computational complexity

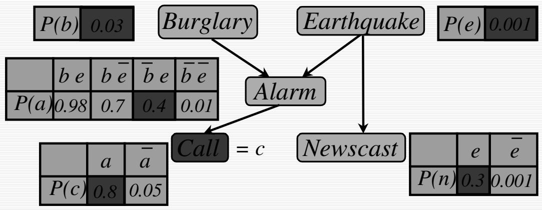
- **Theorem:** Inference in a multi-connected Bayesian network is NP-hard.

Boolean 3CNF formula $\phi = (u \vee \bar{v} \vee w) \wedge (\bar{u} \vee \bar{w} \vee y)$



Probability ( ) = 1/2ⁿ · # satisfying assignments of ϕ

Stochastic simulation



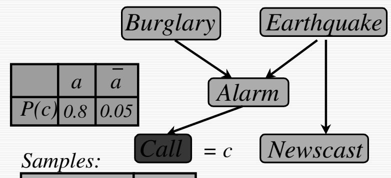
Samples:

B E A C N
$\bar{b}$ e a c $\bar{n}$
~~b $\bar{e}$ a $\bar{c}$ n~~
⋮

$$P(b|c) \sim \frac{\#\ \textit{of live samples with } B=b}{\textit{total \# of live samples}}$$

73

Likelihood weighting



Samples:

B E A C N	*weight*
$\bar{b}$ e a c $\bar{n}$	0.8
b $\bar{e}$ $\bar{a}$ $\bar{c}$ n	0.95
⋮	

$$P(b|c) = \frac{\textit{weight of samples with } B=b}{\textit{total weight of samples}}$$

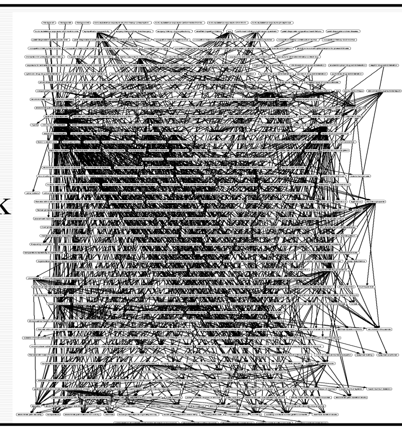
74

Other approaches

- Search based techniques
  - search for high-probability instantiations
  - use instantiations to approximate probabilities
- Structural approximation
  - simplify network
    - eliminate edges, nodes
    - abstract node values
    - simplify CPTs
  - do inference in simplified network

75

CPCS Network



76

Course Contents

- Concepts in Probability
- Bayesian Networks
- Inference
- » Decision making
- Learning networks from data
- Reasoning over time
- Applications

77

Decision making

- Decisions, Preferences, and Utility functions
- Influence diagrams
- Value of information

78

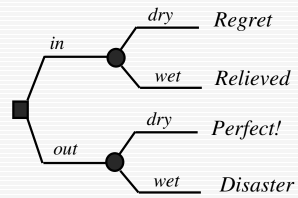
Decision making

- Decision - an irrevocable allocation of domain resources
- Decision should be made so as to maximize expected utility.
- View decision making in terms of
  - Beliefs/Uncertainties
  - Alternatives/Decisions
  - Objectives/Utilities

79

A Decision Problem

Should I have my party inside or outside?



80

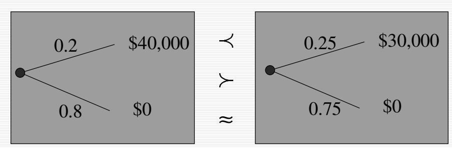
Value Function

- A numerical score over all possible states of the world.

Location?	Weather?	Value
in	dry	\$50
in	wet	\$60
out	dry	\$100
out	wet	\$0

81

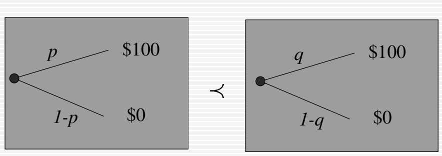
Preference for Lotteries



82

Desired Properties for Preferences over Lotteries

If you prefer \$100 to \$0 and $p < q$ then

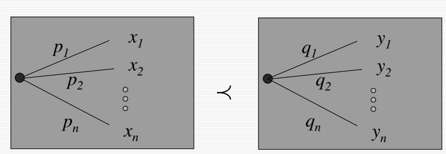


(always)

83

Expected Utility

Properties of preference $\Rightarrow$
existence of function U, that satisfies:

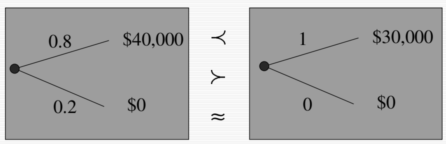


iff

$$\Sigma_i\, p_i\, U(x_i) \;<\; \Sigma_i\, q_i\, U(y_i)$$

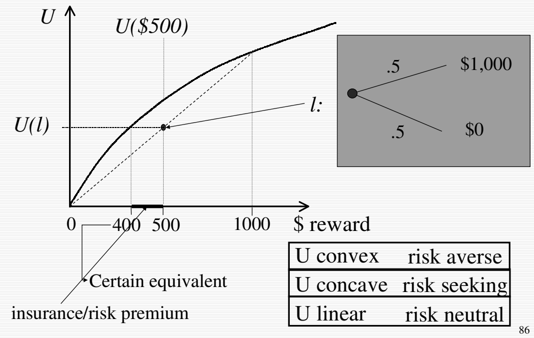
84

Some properties of U



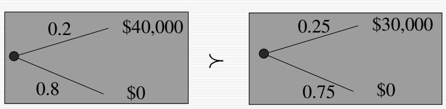
⇒U ≠ monetary payoff

Attitudes towards risk

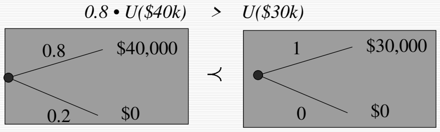


U convex	risk averse
U concave	risk seeking
U linear	risk neutral

Are people rational?

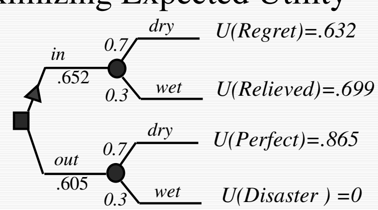


0.2 • U($40k) > *0.25 • U($30k)*
0.8 • U($40k) > *U($30k)*



0.8 • U($40k) < *U($30k)*

Maximizing Expected Utility



choose the action that maximizes expected utility

$EU(in) = 0.7 \cdot .632 + 0.3 \cdot .699 = .652$ ➡ Choose *in*

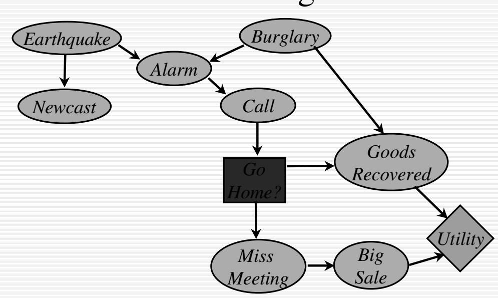
$EU(out) = 0.7 \cdot .865 + 0.3 \cdot 0 = .605$

Multi-attribute utilities

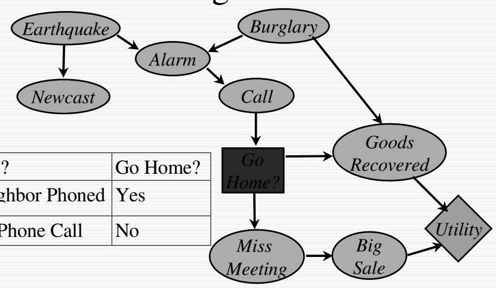
(or: Money isn't everything)

- Many aspects of an outcome combine to determine our preferences.
  - vacation planning: cost, flying time, beach quality, food quality, …
  - medical decision making: risk of death (micromort), quality of life (QALY), cost of treatment, …
- For rational decision making, must combine all relevant factors into single utility function.

Influence Diagrams



Decision Making with Influence Diagrams



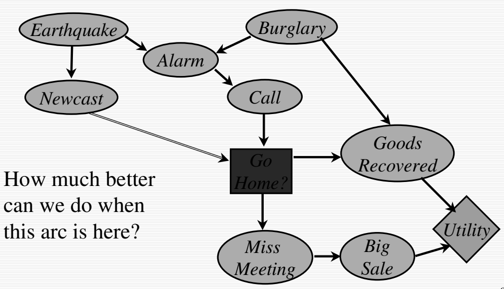
Call?	Go Home?
Neighbor Phoned	Yes
No Phone Call	No

Expected Utility of this policy is 100

Value-of-Information

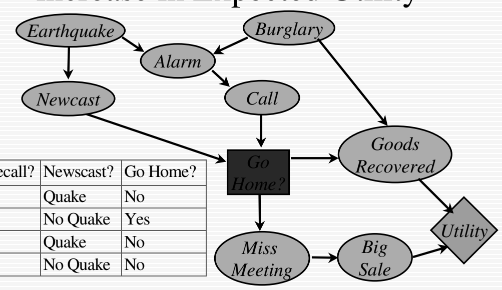
- What is it worth to get another piece of information?
- What is the increase in (maximized) expected utility if I make a decision with an additional piece of information?
- Additional information (if free) cannot make you worse off.
- There is no value-of-information if you will not change your decision.

Value-of-Information in an Influence Diagram



How much better can we do when this arc is here?

Value-of-Information is the increase in Expected Utility



Phonecall?	Newscast?	Go Home?
Yes	Quake	No
Yes	No Quake	Yes
No	Quake	No
No	No Quake	No

Expected Utility of this policy is 112.5

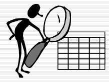
Course Contents

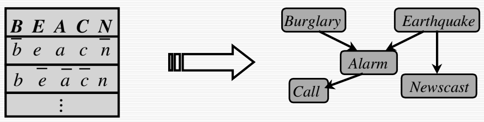
- Concepts in Probability
- Bayesian Networks
- Inference
- Decision making
- » Learning networks from data
- Reasoning over time
- Applications

Learning networks from data

- The learning task
- Parameter learning
  - Fully observable
  - Partially observable
- Structure learning
- Hidden variables

The learning task





Input: training data *Output: BN modeling data*

- Input: fully or partially observable data cases?
- Output: parameters or also structure?

Parameter learning: one variable

- Unfamiliar coin:
  - Let θ = bias of coin (long-run fraction of heads)
- If θ known (given), then
  - $P(X = heads \mid \theta) = \theta$
- Different coin tosses independent given θ

$\Rightarrow \underbrace{P(X_1, \ldots, X_n}_{h \text{ heads, } t \text{ tails}} \mid \theta) = \theta^h (1-\theta)^t$

Maximum likelihood

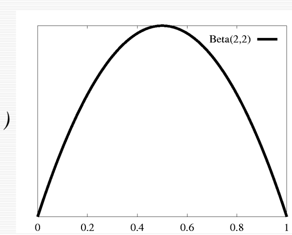
- Input: a set of previous coin tosses
  - $X_1, \ldots, X_n = \{\underbrace{\text{H, T, H, H, H, T, T, H, . . ., H}}_{h \text{ heads, } t \text{ tails}}\}$
- Goal: estimate θ
- The likelihood $P(X_1, \ldots, X_n \mid \theta) = \theta^h (1-\theta)^t$
- The maximum likelihood solution is:

$$\theta^* = \frac{h}{h+t}$$

Bayesian approach

Uncertainty about $\theta \Rightarrow$ distribution over its values

$P(\theta)$



θ

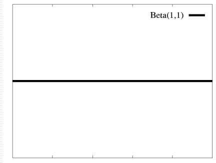
$$P(X = heads) = \int_{-\infty}^{\infty} P(X = heads \mid \theta) P(\theta) d\theta = \int_{-\infty}^{\infty} \theta \, P(\theta)\, d\theta$$

Conditioning on data

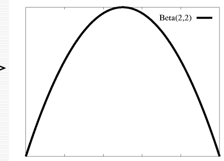
h heads, t tails

D

$P(\theta) \Longrightarrow P(\theta \mid D) \propto P(\theta)\, P(D \mid \theta) = P(\theta)\, \theta^h (1-\theta)^t$

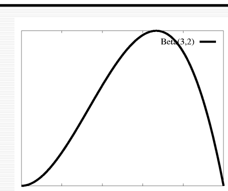


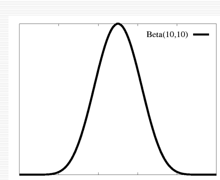
1 head
1 tail

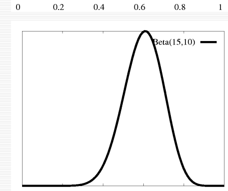


Good parameter distribution:

$$Beta(\alpha_h, \alpha_t) \propto \theta^{\alpha_h - 1} (1-\theta)^{\alpha_t - 1}$$



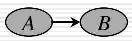




* Dirichlet distribution generalizes Beta to non-binary variables.

General parameter learning

- A multi-variable BN is composed of several independent parameters ("coins").



Three parameters:
$\theta_A,\ \theta_{B|a},\ \theta_{B|\bar{a}}$

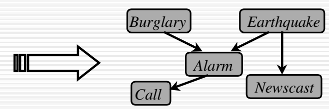
- Can use same techniques as one-variable case to learn each one separately

Max likelihood estimate of $\theta_{B|\bar{a}}$ *would be:*

$$\theta^*_{B|\bar{a}} = \frac{\#\text{data cases with } b, \bar{a}}{\#\text{data cases with } \bar{a}}$$

Partially observable data

B	E	A	C	N
$\bar{b}$	?	a	c	?
b	?	$\bar{a}$	?	n
⋮				



- Fill in missing data with "expected" value
  - expected = distribution over possible values
  - use "best guess" BN to estimate distribution

Intuition

- In fully observable case:

$$\theta^*_{n|e} = \frac{\#\text{data cases with } n, e}{\#\text{data cases with } e} = \frac{\sum_j I(n,e \mid d_j)}{\sum_j I(e \mid d_j)}$$

$$I(e \mid d_j) = \begin{cases} 1 & \text{if } E{=}e \text{ in data case } d_j \\ 0 & \text{otherwise} \end{cases}$$

- In partially observable case I is unknown.

Best estimate for I is: $\hat{I}(n,e|d_j) = P_{\theta^*}(n,e|d_j)$

Problem: θ^* unknown.

Expectation Maximization (EM)

Repeat :

- Expectation (E) step
  - Use current parameters θ to estimate filled in data.

$$\hat{I}(n,e|d_j) = P_{\theta}(n,e|d_j)$$

- Maximization (M) step
  - Use filled in data to do max likelihood estimation

$$\tilde{\theta}_{n|e} = \frac{\sum_j \hat{I}(n,e|d_j)}{\sum_j \hat{I}(e|d_j)}$$

- Set: $\theta := \tilde{\theta}$

until convergence.

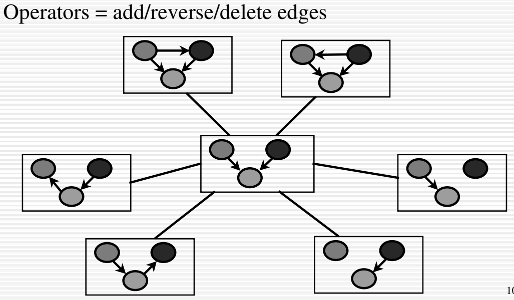
Structure learning

Goal:
find "good" BN structure (relative to data)

Solution:
do heuristic search over space of network structures.

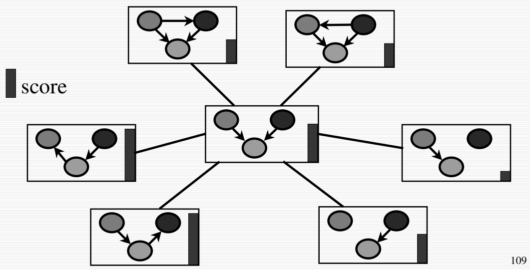
Search space

Space = network structures
Operators = add/reverse/delete edges



Heuristic search

Use scoring function to do heuristic search (any algorithm).
Greedy hill-climbing with randomness works pretty well.



109

Scoring

- Fill in parameters using previous techniques & score completed networks.
- One possibility for score:
  likelihood function: *Score(B) = P(data | B)*

  Example: *X, Y* independent coin tosses
  typical *data = (27 h-h, 22 h-t, 25 t-h, 26 t-t)*

  Maximum likelihood network structure:

$$X \rightarrow Y$$

Max. likelihood network typically fully connected

This is not surprising: maximum likelihood always overfits…

110

Better scoring functions

- MDL formulation: balance fit to data and model complexity (# of parameters)

  Score(B) = P(data | B) - model complexity

- Full Bayesian formulation
  - prior on network structures & parameters
  - more parameters ⇒ higher dimensional space
  - get balance effect as a byproduct*

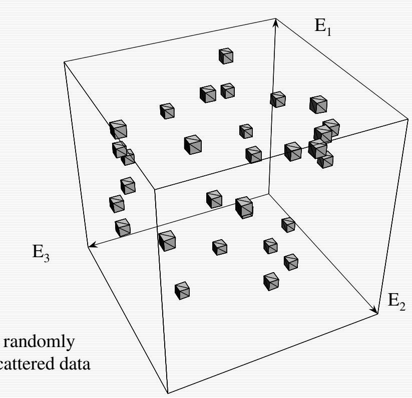
* with Dirichlet parameter prior, MDL is an approximation to full Bayesian score.

111

Hidden variables

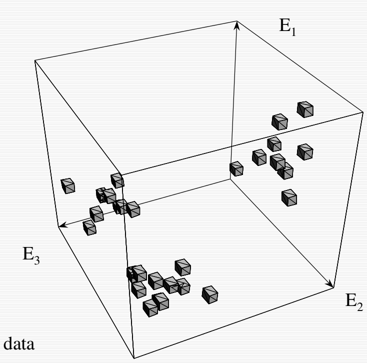
- There may be interesting variables that we never get to observe:
  - topic of a document in information retrieval;
  - user's current task in online help system.
- Our learning algorithm should
  - hypothesize the existence of such variables;
  - learn an appropriate state space for them.

112



randomly scattered data

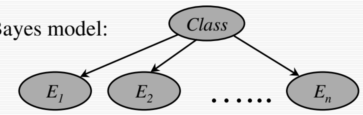
113



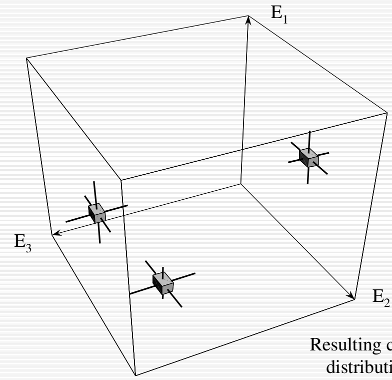
actual data

Bayesian clustering (Autoclass)

naïve Bayes model:



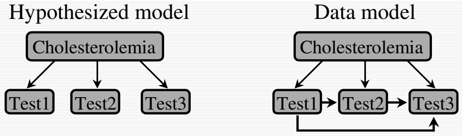
- (hypothetical) class variable never observed
- if we know that there are *k* classes, just run EM
- learned classes = clusters
- Bayesian analysis allows us to choose *k*, trade off fit to data with model complexity



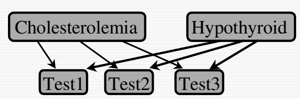
Resulting cluster distributions

Detecting hidden variables

- Unexpected correlations ⇨ hidden variables.



“Correct” model



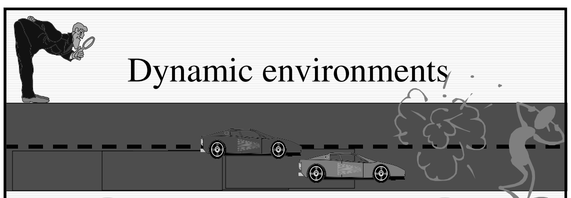
Course Contents

- Concepts in Probability
- Bayesian Networks
- Inference
- Decision making
- Learning networks from data
- » Reasoning over time
- Applications

Reasoning over time

- Dynamic Bayesian networks
- Hidden Markov models
- Decision-theoretic planning
  - Markov decision problems
  - Structured representation of actions
  - The qualification problem & the frame problem
  - Causality (and the frame problem revisited)

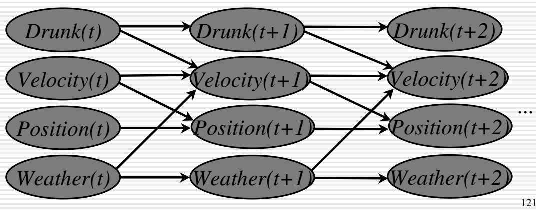
Dynamic environments



- Markov property:
  - past independent of future given current state;
  - a conditional independence assumption;
  - implied by fact that there are no arcs $t \rightarrow t+2$.

Dynamic Bayesian networks

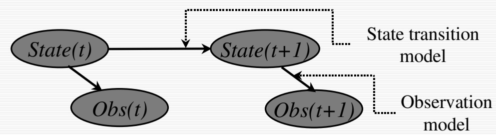
- State described via random variables.
- Each variable depends only on few others.



121

Hidden Markov model

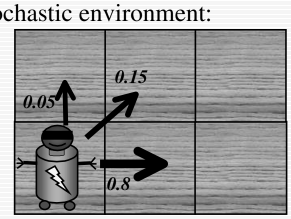
- An HMM is a simple model for a partially observable stochastic domain.



122

Hidden Markov models (HMMs)

Partially observable stochastic environment:



- Mobile robots:
  - states = location
  - observations = sensor input
- Speech recognition:
  - states = phonemes
  - observations = acoustic signal
- Biological sequencing:
  - states = protein structure
  - observations = amino acids

123

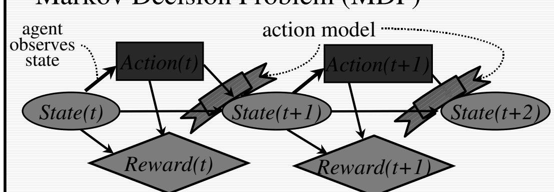
HMMs and DBNs

- HMMs are just very simple DBNs.
- Standard inference & learning algorithms for HMMs are instances of DBN algorithms
  - Forward-backward = polytree
  - Baum-Welch = EM
  - Viterbi = most probable explanation.

124

Acting under uncertainty

Markov Decision Problem (MDP)

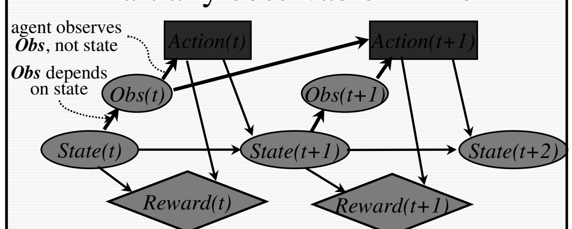


- Overall utility = sum of momentary rewards.
- Allows rich preference model, e.g.:

rewards corresponding to "get to goal asap" $= \begin{cases} +100 & \text{goal states} \\ -1 & \text{other states} \end{cases}$

125

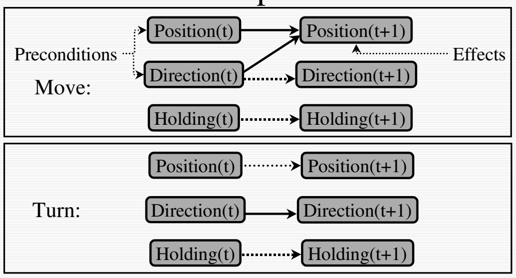
Partially observable MDPs



- The optimal action at time t depends on the entire history of previous observations.
- Instead, a distribution over *State(t)* suffices.

126

Structured representation



Probabilistic action model
- allows for exceptions & qualifications;
- persistence arcs: a solution to the frame problem.

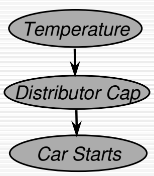
127

Causality

- Modeling the effects of interventions
- Observing vs. "setting" a variable
- A form of persistence modeling

128

Causal Theory

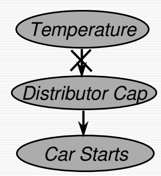


Cold temperatures can cause the distributor cap to become cracked.

If the distributor cap is cracked, then the car is less likely to start.

129

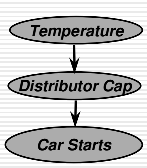
Setting vs. Observing



The car does not start. Will it start if we replace the distributor?

130

Predicting the effects of interventions

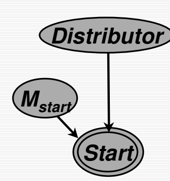


The car does not start. Will it start if we replace the distributor?

What is the probability that the car will start if I replace the distributor cap?

131

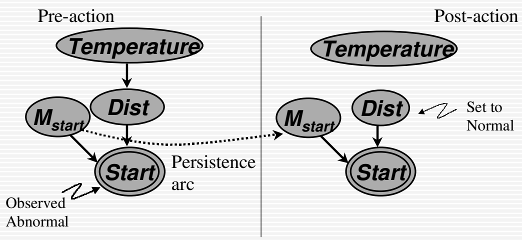
Mechanism Nodes



M_{start}	Distributor	Starts?
Always Starts	Cracked	Yes
Always Starts	Normal	Yes
Never Starts	Cracked	No
Never Starts	Normal	No
Normal	Cracked	No
Normal	Normal	Yes
Inverse	Cracked	Yes
Inverse	Normal	No

132

Persistence



Assumption:The mechanism relating *Dist* to *Start* is unchanged by replacing the *Distributor*.

133

Course Contents

- Concepts in Probability
- Bayesian Networks
- Inference
- Decision making
- Learning networks from data
- Reasoning over time
- » Applications

134

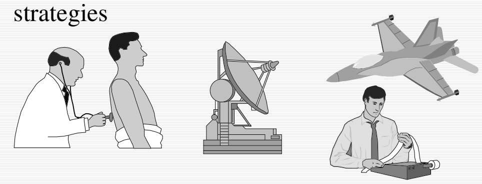
Applications

- Medical expert systems
  - Pathfinder
  - Parenting MSN
- Fault diagnosis
  - Ricoh FIXIT
  - Decision-theoretic troubleshooting
- Vista
- Collaborative filtering

135

Why use Bayesian Networks?

- Explicit management of uncertainty/tradeoffs
- Modularity implies maintainability
- Better, flexible, and robust recommendation strategies



136

Pathfinder

- Pathfinder is one of the first BN systems.
- It performs diagnosis of lymph-node diseases.
- It deals with over 60 diseases and 100 findings.
- Commercialized by Intellipath and Chapman Hall publishing and applied to about 20 tissue types.

137

Studies of Pathfinder Diagnostic Performance

- Naïve Bayes performed considerably better than certainty factors and Dempster-Shafer Belief Functions.
- Incorrect zero probabilities caused 10% of cases to be misdiagnosed.
- Full Bayesian network model with feature dependencies did best.

138

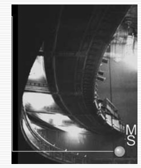
Commercial system: Integration

- Expert System with advanced diagnostic capabilities
  - uses key features to form the differential diagnosis
  - recommends additional features to narrow the differential diagnosis
  - recommends features needed to confirm the diagnosis
  - explains correct and incorrect decisions
- Video atlases and text organized by organ system
- “Carousel Mode” to build customized lectures
- Anatomic Pathology Information System

139

On Parenting: Selecting problem

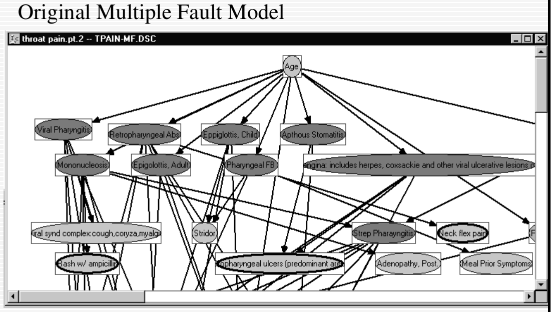
- Diagnostic indexing for Home Health site on Microsoft Network
- Enter symptoms for pediatric complaints
- Recommends multimedia content



140

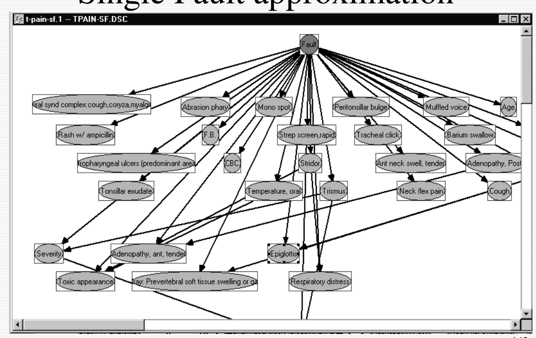
On Parenting : MSN

Original Multiple Fault Model



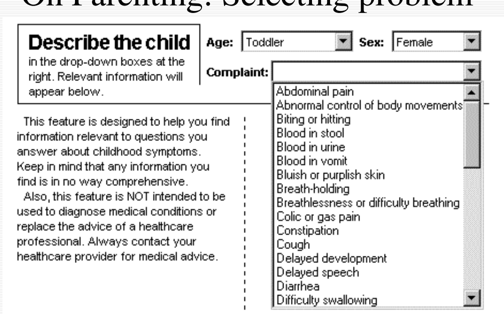
141

Single Fault approximation



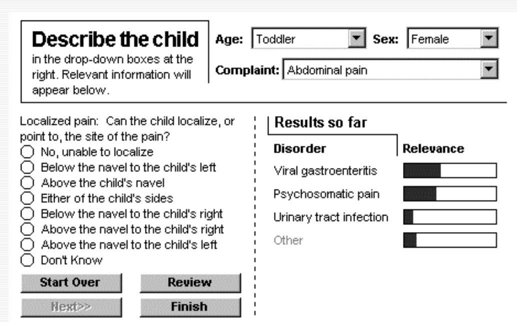
142

On Parenting: Selecting problem



143

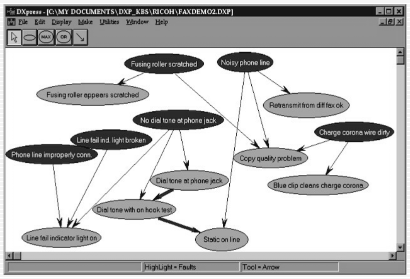
Performing diagnosis/indexing



144

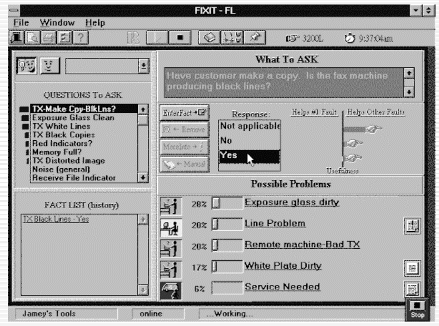
RICOH Fixit

■ Diagnostics and information retrieval



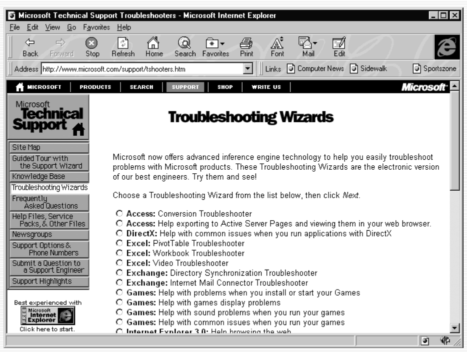
145

FIXIT: Ricoh copy machine



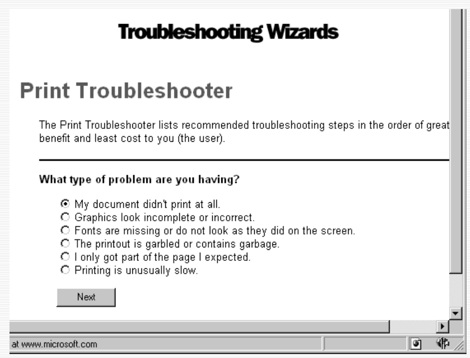
146

Online Troubleshooters



147

Define Problem

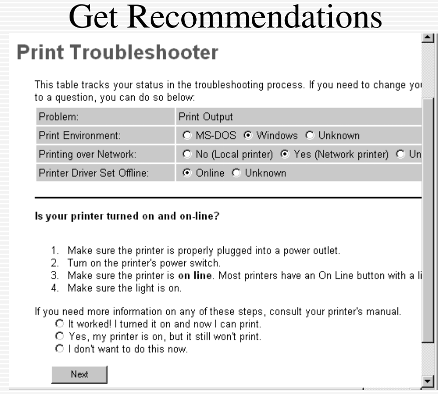


148

Gather Information

149

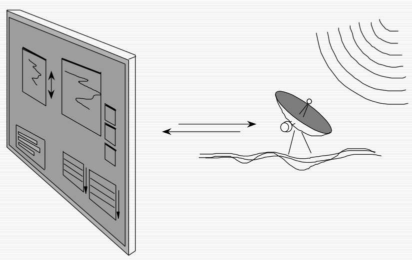
Get Recommendations



150

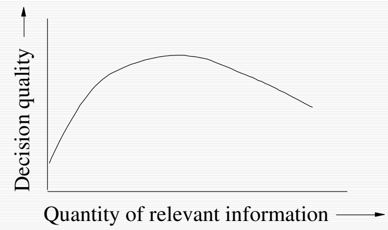
Vista Project: NASA Mission Control

Decision-theoretic methods for display for high-stakes aerospace decisions



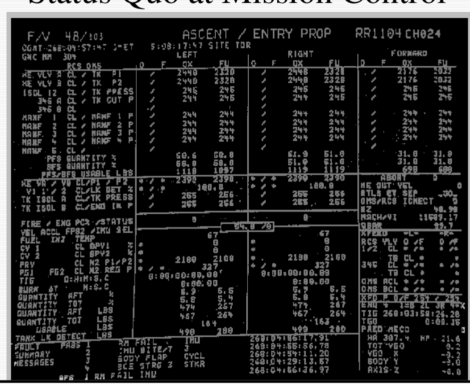
151

Costs & Benefits of Viewing Information



152

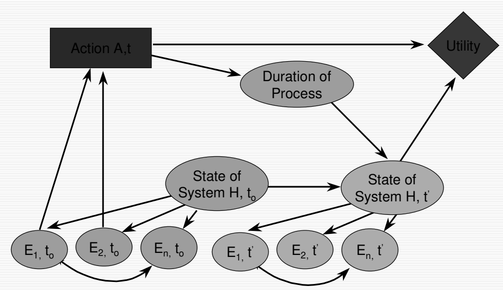
Status Quo at Mission Control



153

Time-Critical Decision Making

- **Consideration of time delay in temporal process**

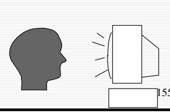


154

Simplification: Highlighting Decisions

- Variable threshold to control amount of highlighted information

Oxygen	15.6	14.2
Fuel Pres	10.5	11.8
Chamb Pres	5.4	4.8
He Pres	17.7	14.7
Delta v	33.3	63.3
Oxygen	10.2	10.6
Fuel Pres	12.8	12.5
Chamb Pres	0.0	0.0
He Pres	15.8	15.7
Delta v	32.3	63.3

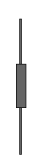


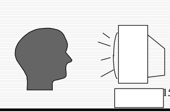
155

Simplification: Highlighting Decisions

- Variable threshold to control amount of highlighted information

Oxygen	15.6	14.2
Fuel Pres	10.5	11.8
Chamb Pres	5.4	4.8
He Pres	17.7	14.7
Delta v	33.3	63.3
Oxygen	10.2	10.6
Fuel Pres	12.8	12.5
Chamb Pres	0.0	0.0
He Pres	15.8	15.7
Delta v	32.3	63.3



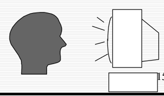


156

Simplification: Highlighting Decisions

- Variable threshold to control amount of highlighted information

Oxygen	15.6	14.2
Fuel Pres	10.5	11.8
Chamb Pres	5.4	4.8
He Pres	17.7	14.7
Delta v	33.3	63.3
Oxygen	10.2	10.6
Fuel Pres	12.8	12.5
Chamb Pres	0.0	0.0
He Pres	15.8	15.7
Delta v	32.3	63.3



157

What is Collaborative Filtering?

- A way to find cool websites, news stories, music artists etc
- Uses data on the preferences of many users, not descriptions of the content.
- Firefly, Net Perceptions (GroupLens), and others offer this technology.

158

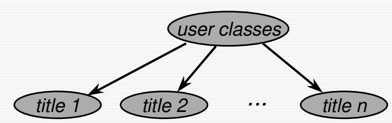
Bayesian Clustering for Collaborative Filtering

- Probabilistic summary of the data
- Reduces the number of parameters to represent a set of preferences
- Provides insight into usage patterns.
- Inference:

P(Like title i | Like title j, Like title k)

159

Applying Bayesian clustering



	class1	class2 ...
title1	p(like)=0.2	p(like)=0.8
title2	p(like)=0.7	p(like)=0.1
title3	p(like)=0.99	p(like)=0.01
...		

160

MSNBC Story clusters

Readers of commerce and technology stories (36%):
- E-mail delivery isn't exactly guaranteed
- Should you buy a DVD player?
- Price low, demand high for Nintendo

Readers of top promoted stories (29%):
- 757 Crashes At Sea
- Israel, Palestinians Agree To Direct Talks
- Fuhrman Pleads Innocent To Perjury

Sports Readers (19%):
- Umps refusing to work is the right thing
- Cowboys are reborn in win over eagles
- Did Orioles spend money wisely?

Readers of “Softer” News (12%):
- The truth about what things cost
- Fuhrman Pleads Innocent To Perjury
- Real Astrology

161

Top 5 shows by user class

Class 1
- Power rangers
- Animaniacs
- X-men
- Tazmania
- Spider man

Class 2
- Young and restless
- Bold and the beautiful
- As the world turns
- Price is right
- CBS eve news

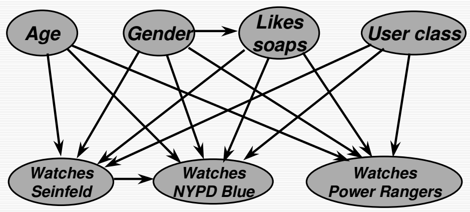
Class 3
- Tonight show
- Conan O’Brien
- NBC nightly news
- Later with Kinnear
- Seinfeld

Class 4
- 60 minutes
- NBC nightly news
- CBS eve news
- Murder she wrote
- Matlock

Class 5
- Seinfeld
- Friends
- Mad about you
- ER
- Frasier

162

Richer model



163

What's old?

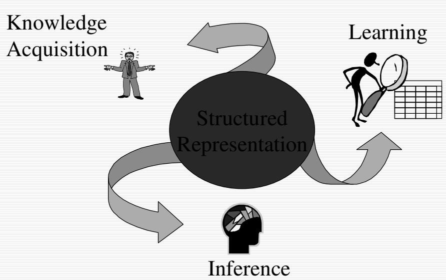
Decision theory & probability theory provide:

- principled models of belief and preference;
- techniques for:
  - integrating evidence (conditioning);
  - optimal decision making (max. expected utility);
  - targeted information gathering (value of info.);
  - parameter estimation from data.

164

What's new?

Bayesian networks exploit domain structure to allow compact representations of complex models.



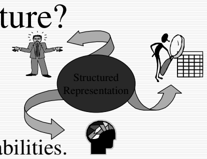
165

Some Important AI Contributions

- Key technology for diagnosis.
- Better more coherent expert systems.
- New approach to planning & action modeling:
  - planning using Markov decision problems;
  - new framework for reinforcement learning;
  - probabilistic solution to frame & qualification problems.
- New techniques for learning models from data.

166

What's in our future?



- Better models for:
  - preferences & utilities;
  - not-so-precise numerical probabilities.
- Inferring causality from data.
- More expressive representation languages:
  - structured domains with multiple objects;
  - levels of abstraction;
  - reasoning about time;
  - hybrid (continuous/discrete) models.

167