



## Lecture 27 of 42

### Uncertain Reasoning: Probability Review Discussion: Probabilistic Reasoning Apps

Friday, 27 October 2006

William H. Hsu

Department of Computing and Information Sciences, KSU

KSOL course page: <http://snipurl.com/v9v3>

Course web site: <http://www.kddresearch.org/Courses/Fall-2006/CIS730>

Instructor home page: <http://www.cis.ksu.edu/~bhsu>

Reading for Next Class:

Chapter 13, Russell & Norvig 2<sup>nd</sup> edition



## Lecture Outline

- Today's Reading: Sections 12.5 – 12.8, R&N 2e
- Next Week's Reading: Chapter 13, Sections 14.1 – 14.2, R&N 2e
- Today: Intro to Uncertain Reasoning
  - \* Nondeterminism in the real world
  - \* Incomplete domain theories
  - \* Observation errors: sensor, measurement, estimation
  - \* Actuator errors
- Probability Review
  - \* Kolmogorov axioms
  - \* Conditioning
- Next Week: Graphical models
  - \* Bayesian networks and causality
  - \* Inference and learning
  - \* BNJ interface (<http://bnj.sourceforge.net>)





## Looking Ahead [1]: Planning and Learning Roadmap

- Bounded Indeterminacy (12.3)
- Four Techniques for Dealing with Nondeterministic Domains
- 1. Sensorless aka Conformant Planning: “Be Prepared” (12.3)
  - \* Idea: be able to respond to any situation (universal planning)
  - \* Coercion
- 2. Conditional aka Contingency Planning: “Review the Situation” (12.4)
  - \* Idea: be able to respond to many typical alternative situations
  - \* Actions for sensing
- 3. Execution Monitoring and Replanning: “The Show Must Go On” (12.5)
  - \* Idea: be able to resume momentarily failed plans
  - \* Plan revision
- 4. Continuous Planning: “Always in Motion, The Future Is” (12.6)
  - \* Lifetime planning (and learning!)
  - \* Formulate new goals



## Probability: Basic Definitions and Axioms

- Sample Space ( $\Omega$ ): Range of a Random Variable  $X$
- Probability Measure  $Pr(\bullet)$ 
  - \*  $\Omega$  denotes a range of “events”;  $X: \Omega$
  - \* Probability  $Pr$ , or  $P$ , is a *measure* over  $2^\Omega$
  - \* In a general sense,  $Pr(X = x \in \Omega)$  is a measure of belief in  $X = x$ 
    - $\Rightarrow P(X = x) = 0$  or  $P(X = x) = 1$ : plain (aka categorical) beliefs (can't be revised)
    - $\Rightarrow$  All other beliefs are subject to revision
- Kolmogorov Axioms
  - \* 1.  $\forall x \in \Omega . 0 \leq P(X = x) \leq 1$
  - \* 2.  $P(\Omega) \equiv \sum_{x \in \Omega} P(X = x) = 1$
  - \* 3.  $\forall \mathbf{X}_1, \mathbf{X}_2, \dots \exists i \neq j \Rightarrow \mathbf{X}_i \wedge \mathbf{X}_j = \emptyset$ .
$$P\left(\bigcup_{i=1}^{\infty} \mathbf{X}_i\right) = \sum_{i=1}^{\infty} P(\mathbf{X}_i)$$
- Joint Probability:  $P(X_1 \wedge X_2) \equiv$  Probability of the Joint Event  $X_1 \wedge X_2$
- Independence:  $P(X_1 \wedge X_2) = P(X_1) \cdot P(X_2)$





## Basic Formulas for Probabilities

- **Product Rule** (Alternative Statement of Bayes's Theorem)

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

Cause → Effect  
"predictive"

- \* Proof: requires axiomatic set theory, as does Bayes's Theorem

- **Sum Rule**

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

- \* Sketch of proof (immediate from axiomatic set theory)

⇒ Draw a Venn diagram of two sets denoting events A and B

⇒ Let  $A \cup B$  denote the event corresponding to  $A \vee B$ ...

Forecast  
prognosis:



- **Theorem of Total Probability**

- \* Suppose events  $A_1, A_2, \dots, A_n$  are mutually exclusive and exhaustive

⇒ Mutually exclusive:  $i \neq j \Rightarrow A_i \wedge A_j = \emptyset$

⇒ Exhaustive:  $\sum P(A_i) = 1$

- \* Then  $P(B) = \sum_{i=1}^n P(B|A_i) \cdot P(A_i)$

- \* Proof: follows from product rule and 3<sup>rd</sup> Kolmogorov axiom



## Bayes's Theorem [1]

Product rule  $P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$

$$\Rightarrow \text{Bayes' rule } P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Why is this useful???

For assessing diagnostic probability from causal probability:

$$P(\text{Cause}|\text{Effect}) = \frac{P(\text{Effect}|\text{Cause})P(\text{Cause})}{P(\text{Effect})}$$

E.g., let  $M$  be meningitis,  $S$  be stiff neck:

$$P(M|S) = \frac{P(S|M)P(M)}{P(S)} = \frac{0.8 \times 0.0001}{0.1} = 0.0008$$

Note: posterior probability of meningitis still very small!

Adapted from slides by S. Russell, UC Berkeley



## Looking Ahead [2]: Uncertain Reasoning Roadmap

- Framework: Interpretations of Probability [Cheeseman, 1985]
  - \* Bayesian subjectivist view
    - ⇒ A measure of an agent's belief in a proposition
    - ⇒ Proposition denoted by random variable (sample space: range)
    - ⇒ e.g.,  $Pr(\text{Outlook} = \text{Sunny}) = 0.8$
  - \* Frequentist view: probability is the *frequency of observations* of an event
  - \* Logicist view: probability is inferential evidence in favor of a proposition
- Some Applications
  - \* HCI: learning natural language; intelligent displays; decision support
  - \* Approaches: prediction; sensor and data fusion (e.g., bioinformatics)
- Prediction: Examples
  - \* Measure *relevant parameters*: temperature, barometric pressure, wind speed
  - \* Make statement of the form  $Pr(\text{Tomorrow's-Weather} = \text{Rain}) = 0.5$
  - \* College admissions:  $Pr(\text{Acceptance}) = p$ 
    - ⇒ Plain beliefs: unconditional acceptance ( $p = 1$ ) or categorical rejection ( $p = 0$ )
    - ⇒ Conditional beliefs: depends on reviewer (use probabilistic model)



## Automated Reasoning using Probabilistic Models: Inference Tasks

Simple queries: compute posterior marginal  $P(X_i | \mathbf{E} = e)$

e.g.,  $P(\text{NoGas} | \text{Gauge} = \text{empty}, \text{Lights} = \text{on}, \text{Starts} = \text{false})$

Conjunctive queries:  $P(X_i, X_j | \mathbf{E} = e) = P(X_i | \mathbf{E} = e)P(X_j | X_i, \mathbf{E} = e)$

Optimal decisions: decision networks include utility information;  
probabilistic inference required for  $P(\text{outcome} | \text{action}, \text{evidence})$

Value of information: which evidence to seek next?

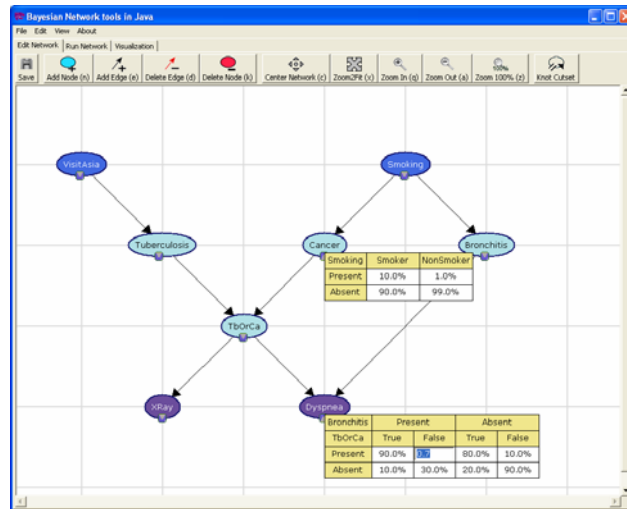
Sensitivity analysis: which probability values are most critical?

Explanation: why do I need a new starter motor?

Adapted from slides by S. Russell, UC Berkeley



## Looking Ahead [3]: Bayesian Network tools in Java (BNJ)

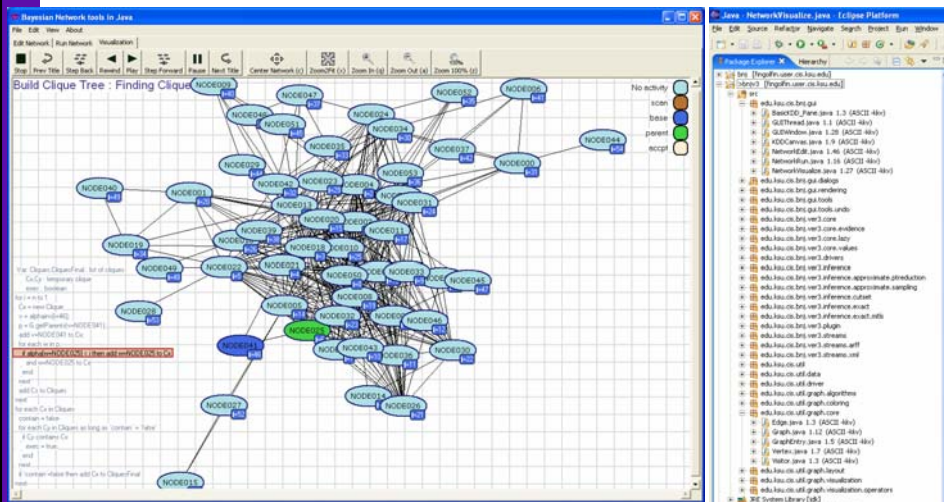


© 2004 KSU BNJ Development Team

Asia (Chest Clinic) Network



## BNJ Core [2] Graph Architecture



© 2004 KSU BNJ Development Team

CPCS-54 Network



## Bayes's Theorem [2]

- Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} = \frac{P(h \wedge D)}{P(D)}$$

- $P(h) \equiv$  Prior Probability of Assertion (Hypothesis)  $h$ 
  - \* Measures initial beliefs (BK) before any information is obtained (hence prior)
- $P(D) \equiv$  Prior Probability of Data (Observations)  $D$ 
  - \* Measures probability of obtaining sample  $D$  (i.e., expresses  $D$ )
- $P(h | D) \equiv$  Probability of  $h$  Given  $D$ 
  - \*  $|$  denotes conditioning - hence  $P(h | D)$  is a conditional (aka posterior) probability
- $P(D | h) \equiv$  Probability of  $D$  Given  $h$ 
  - \* Measures probability of observing  $D$  given that  $h$  is correct ("generative" model)
- $P(h \wedge D) \equiv$  Joint Probability of  $h$  and  $D$ 
  - \* Measures probability of observing  $D$  and of  $h$  being correct



## Bayesian Inference: Query Answering (QA)

- Answering User Queries
  - \* Suppose we want to perform intelligent inferences over a database  $DB$ 
    - $\Rightarrow$  Scenario 1:  $DB$  contains records (instances), some "labeled" with answers
    - $\Rightarrow$  Scenario 2:  $DB$  contains probabilities (annotations) over propositions
  - \* QA: an application of probabilistic inference

- QA Using Prior and Conditional Probabilities: Example

- \* Query: *Does patient have cancer or not?*
- \* Suppose: patient takes a lab test and result comes back positive
  - $\Rightarrow$  Correct + result in only 98% of the cases in which disease is actually present
  - $\Rightarrow$  Correct - result in only 97% of the cases in which disease is not present
  - $\Rightarrow$  Only 0.008 of the entire population has this cancer
- \*  $\alpha \equiv P(\text{false negative for } H_0 \equiv \text{Cancer}) = 0.02$  (NB: for 1-point sample)
- \*  $\beta \equiv P(\text{false positive for } H_0 \equiv \text{Cancer}) = 0.03$  (NB: for 1-point sample)

$$P(\text{Cancer}) = 0.008 \quad P(+ | \text{Cancer}) = 0.98 \quad P(+ | \neg \text{Cancer}) = 0.03$$

$$P(\neg \text{Cancer}) = 0.992 \quad P(- | \text{Cancer}) = 0.02 \quad P(- | \neg \text{Cancer}) = 0.97$$

- \*  $P(+ | H_0) P(H_0) = 0.0078$ ,  $P(+ | H_A) P(H_A) = 0.0298 \Rightarrow h_{MAP} = H_A \equiv \neg \text{Cancer}$





## Choosing Hypotheses

- Bayes's Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} = \frac{P(h \wedge D)}{P(D)}$$

- MAP Hypothesis

- \* Generally want most probable hypothesis given the training data
- \* Define:  $\arg \max_{x \in \Omega} [f(x)]$   $\equiv$  the value of  $x$  in the sample space  $\Omega$  with the highest  $f(x)$
- \* Maximum a posteriori hypothesis,  $h_{MAP}$

$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(h|D) \\ &= \arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg \max_{h \in H} P(D|h)P(h) \end{aligned}$$

- ML Hypothesis

- \* Assume that  $p(h_i) = p(h_j)$  for all pairs  $i, j$  (uniform priors, i.e.,  $P_H \sim$  Uniform)
- \* Can further simplify and choose the maximum likelihood hypothesis,  $h_{ML}$

$$h_{ML} = \arg \max_{h_j \in H} P(D|h_j)$$



## Automated Reasoning using Probabilistic Models: Inference Tasks

Simple queries: compute posterior marginal  $P(X_i|\mathbf{E}=e)$   
e.g.,  $P(\text{NoGas}|\text{Gauge} = \text{empty}, \text{Lights} = \text{on}, \text{Starts} = \text{false})$

Conjunctive queries:  $P(X_i, X_j|\mathbf{E}=e) = P(X_i|\mathbf{E}=e)P(X_j|X_i, \mathbf{E}=e)$

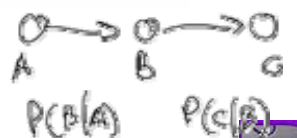
Optimal decisions: decision networks include utility information;  
probabilistic inference required for  $P(\text{outcome}|\text{action}, \text{evidence})$

Value of information: which evidence to seek next?

Sensitivity analysis: which probability values are most critical?

Explanation: why do I need a new starter motor?

MYCIN  
↓  
NEOMYCIN



Adapted from slides by S. Russell, UC Berkeley



## Choosing Hypotheses

- Bayes's Theorem

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} = \frac{P(h \wedge D)}{P(D)}$$



- MAP Hypothesis

- Generally want most probable hypothesis given the training data
- Define:  $\arg \max_{x \in \Omega} [f(x)]$   $\equiv$  the value of  $x$  in the sample space  $\Omega$  with the highest  $f(x)$
- Maximum a posteriori hypothesis,  $h_{MAP}$

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

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

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

- ML Hypothesis

- Assume that  $p(h_i) = p(h_j)$  for all pairs  $i, j$  (uniform priors, i.e.,  $P_H \sim$  Uniform)
- Can further simplify and choose the maximum likelihood hypothesis,  $h_{ML}$

$$h_{ML} = \arg \max_{h_j \in H} P(D|h_j)$$



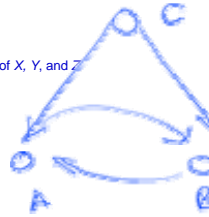
## Graphical Models of Probability

- Conditional Independence

- $X$  is conditionally independent (CI) from  $Y$  given  $Z$  iff  $P(X|Y, Z) = P(X|Z)$  for all values of  $X, Y$ , and  $Z$
- Example:  $P(\text{Thunder} | \text{Rain}, \text{Lightning}) = P(\text{Thunder} | \text{Lightning}) \Leftrightarrow T \perp R | L$

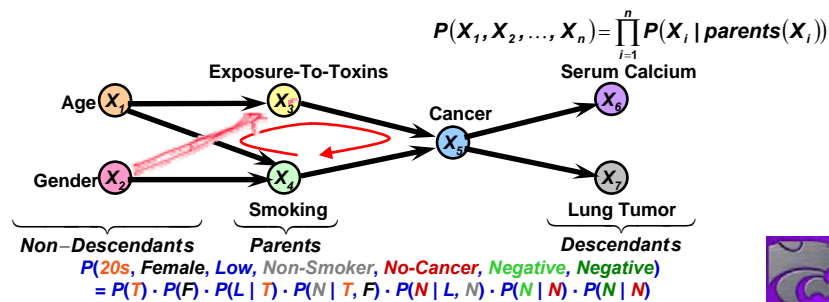
- Bayesian (Belief) Network

- Acyclic directed graph model  $B = (V, E, \Theta)$  representing CI assertions over  $\Theta$
- Vertices (nodes)  $V$  denote events (each a random variable)
- Edges (arcs, links)  $E$  denote conditional dependencies



- Markov Condition for BBNs (Chain Rule):

- Example BBN





## Semantics of Bayesian Networks

“Global” semantics defines the full joint distribution as the product of the local conditional distributions:

$$P(X_1, \dots, X_n) = \prod_{i=1}^n P(X_i | \text{Parents}(X_i))$$

e.g.,  $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$  is given by??  
 $= P(\neg B)P(\neg E)P(A|\neg B \wedge \neg E)P(J|A)P(M|A)$

“Local” semantics: each node is conditionally independent of its nondescendants given its parents

Theorem: Local semantics  $\Leftrightarrow$  global semantics

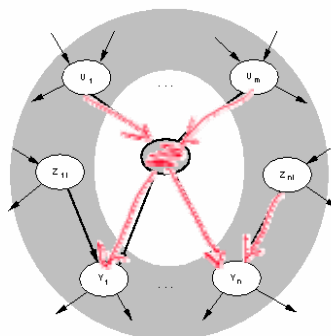


Adapted from slides by S. Russell, UC Berkeley



## Markov Blanket

Each node is conditionally independent of all others given its Markov blanket: parents + children + children's parents



Adapted from slides by S. Russell, UC Berkeley



## Constructing Bayesian Networks: The Chain Rule of Inference

Need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics

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

This choice of parents guarantees the global semantics:

$$\begin{aligned} \mathbf{P}(X_1, \dots, X_n) &= \prod_{i=1}^n \mathbf{P}(X_i | X_1, \dots, X_{i-1}) \text{ (chain rule)} \\ &= \prod_{i=1}^n \mathbf{P}(X_i | \text{Parents}(X_i)) \text{ by construction} \end{aligned}$$

Adapted from slides by S. Russell, UC Berkeley

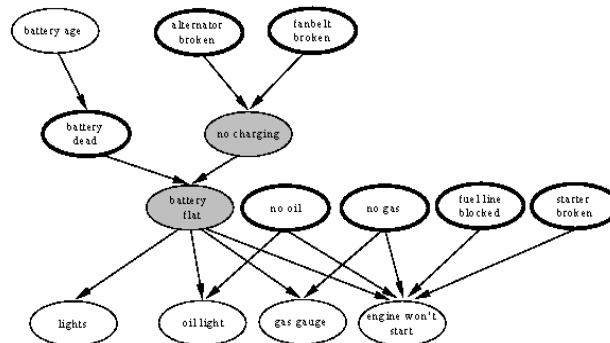


## Example: Evidential Reasoning for Car Diagnosis

Initial evidence: engine won't start

Testable variables (thin ovals), diagnosis variables (thick ovals)

Hidden variables (shaded) ensure sparse structure, reduce parameters



Adapted from slides by S. Russell, UC Berkeley





## Terminology

- Introduction to Reasoning under Uncertainty
  - \* Probability foundations
  - \* Definitions: subjectivist, frequentist, logistic
  - \* (3) Kolmogorov axioms
- Bayes's Theorem
  - \* Prior probability of an event
  - \* Joint probability of an event
  - \* Conditional (posterior) probability of an event
- Maximum *A Posteriori* (MAP) and Maximum Likelihood (ML) Hypotheses
  - \* MAP hypothesis: highest conditional probability given observations (data)
  - \* ML: highest likelihood of generating the observed data
  - \* ML estimation (MLE): estimating parameters to find ML hypothesis
- Bayesian Inference: Computing Conditional Probabilities (CPs) in A Model
- Bayesian Learning: Searching Model (Hypothesis) Space using CPs



## Summary Points

- Introduction to Probabilistic Reasoning
  - \* Framework: using probabilistic criteria to search  $H$
  - \* Probability foundations
    - ⇒ Definitions: subjectivist, objectivist; Bayesian, frequentist, logistic
    - ⇒ Kolmogorov axioms
- Bayes's Theorem
  - \* Definition of conditional (posterior) probability
  - \* Product rule
- Maximum *A Posteriori* (MAP) and Maximum Likelihood (ML) Hypotheses
  - \* Bayes's Rule and MAP
  - \* Uniform priors: allow use of MLE to generate MAP hypotheses
  - \* Relation to version spaces, candidate elimination
- Next Week: Chapter 14, Russell and Norvig
  - \* Later: Bayesian learning: MDL, BOC, Gibbs, Simple (Naïve) Bayes
  - \* Categorizing text and documents, other applications

