Lecture 30 of 42

Reasoning under Uncertainty: Inference and Software Tools, Part 2 of 2
Discussion: *BNJ*

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/CIS730
Instructor home page: http://www.cis.ksu.edu/~bhsu

Reading for Next Class:

Kevin Murphy's survey on BNs, learning: http://bit.ly/2S6Z1L
BNJ homepage: http://bnj.sourceforge.net

Lecture Outline

- Reading for Next Class: Sections 14.1 – 14.2 (p. 492 – 499), R&N 2e
- Last Class: Uncertainty, Probability, 13 (p. 462-486), R&N 2e
  - Where uncertainty is encountered: reasoning, planning, learning (later)
  - Sources: sensor error, incomplete/inaccurate domain theory, randomness
- Today: Probability Intro, Continued, Chapter 13, R&N 2e
  - Why probability
    - Axiomatic basis: Kolmogorov
    - With utility theory: sound foundation of rational decision making
  - Joint probability
  - Independence
  - Probabilistic reasoning: inference by enumeration
  - Conditioning
    - Bayes's theorem (*aka* Bayes' rule)
    - Conditional independence
- Coming Week: More Applied Probability, Graphical Models

Acknowledgements

Stuart J. Russell
Professor of Computer Science
Chair, Department of Electrical Engineering and Computer Sciences
Smith-Zadeh Prof. in Engineering

Peter Norvig
Director of Research

Lotfali Asker-Zadeh (Lotfi A. Zadeh)
Professor of Computer Science
Department of Electrical Engineering and Computer Sciences
Director, Berkeley Initiative in Soft Computing
University of California - Berkeley

Russell, S. J.
University of California, Berkeley
http://www.eecs.berkeley.edu/~russell/

Norvig, P.
http://norvig.com/

Slides from:
http://aima.eecs.berkeley.edu

Zadeh, L. A.
University of California, Berkeley
http://bit.ly/39shSQ

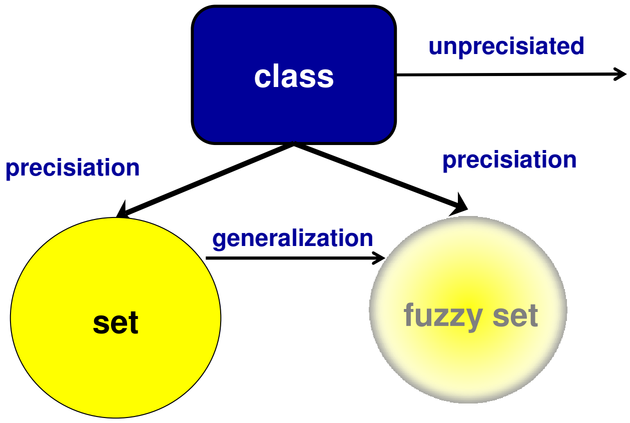
Probability: Basic Definitions and Axioms

- **Sample Space (Ω): Range of Random Variable *X***
- **Probability Measure *Pr*(•)**
  - Ω denotes range of observations; *X*: Ω
  - Probability *Pr*, or *P*: *measure* over power set 2^{Ω} - event space
  - In general sense, $Pr(X = x \in \Omega)$ is measure of belief in $X = x$
    - ⇨ $P(X = x) = 0$ or $P(X = x) = 1$: plain (*aka* categorical) beliefs
    - ⇨ Can't be revised; all other beliefs are subject to revision
- **Kolmogorov Axioms**
  - 1. $\forall x \in \Omega \,.\, 0 \le P(X = x) \le 1$
  - 2. $P(\Omega) \equiv \sum_{x \in \Omega} P(X = x) = 1$
  - 3. $\forall X_1, X_2, \ldots \ni i \ne j \Rightarrow X_i \wedge X_j = \varnothing$.

$$P\left(\bigcup_{i=1}^{\infty} X_i\right) = \sum_{i=1}^{\infty} P(X_i)$$

- **Joint Probability: $P(X_1 \wedge X_2) \equiv$ Prob. of Joint Event $X_1 \wedge X_2$**
- **Independence: $P(X_1 \wedge X_2) = P(X_1) \bullet P(X_2)$**

Non-Probabilistic Representation [1]: Concept of Fuzzy Set



Informally, a fuzzy set, *A*, in a universe of discourse, *U*, is a class with a fuzzy boundary.

Adapted from slide © 2008 L. A. Zadeh, UC Berkeley http://bit.ly/39shSQ

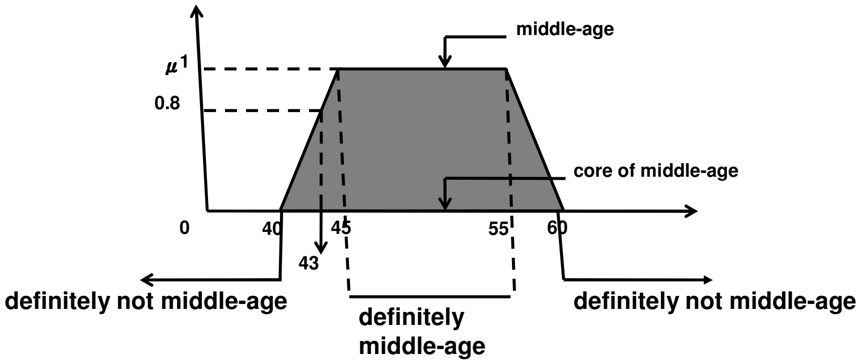
Non-Probabilistic Representation [2]: Precisiation & Degree of Membership

- Set *A* in *U*: Class with Crisp Boundary
- Precisiation: Association with Function whose Domain is *U*
- Precisiation of Crisp Sets
  - Through association with (Boolean-valued) characteristic function
  - $c_A: U \rightarrow \{0, 1\}$
- Precisiation of Fuzzy Sets
  - Through association with membership function
  - $\mu_A: U \rightarrow [0, 1]$
  - $\mu_A(u)$, $u \in U$, represents grade of membership of *u* in *A*
- Degree of Membership
  - Membership in *A*: matter of degree
  - "In fuzzy logic everything is or is allowed to be a matter of degree." – Zadeh

Adapted from slide © 2008 L. A. Zadeh, UC Berkeley http://bit.ly/39shSQ

Non-Probabilistic Representation [3]: Fuzzy Set Example – Middle-Age

- "Linguistic" Variables: Qualitative, Based on Descriptive Terms
- Imprecision of Meaning = Elasticity of Meaning
- Elasticity of Meaning = <u>Fuzziness</u> of Meaning



Adapted from slide © 2008 L. A. Zadeh, UC Berkeley http://bit.ly/39shSQ

Basic Formulas for Probabilities

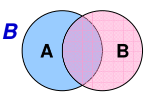
- <u>Product Rule</u> (Alternative Statement of Bayes's Theorem)

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

  - Proof: requires axiomatic set theory, as does Bayes's Theorem
- <u>Sum Rule</u>

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

  - 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 $A \vee B$...



- Theorem of Total Probability
  - Suppose events $A_1, A_2, \ldots, A_n$ are mutually exclusive and exhaustive
    - <u>Mutually exclusive</u>: $i \neq j \Rightarrow A_i \wedge A_j = \emptyset$
    - <u>Exhaustive</u>: $\sum P(A_i) = 1$
  - Then $P(B) = \sum_{i=1}^{n} P(B \mid A_i) \cdot P(A_i)$
  - Proof: follows from product rule and 3rd Kolmogorov axiom

Bayes's Theorem: Joint *vs.* Conditional Probability

- **Theorem**

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

- **$P(h) \equiv$ 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)**
- **$P(h \mid D) \equiv$ Probability of h Given D**
  - **| denotes conditioning - hence P(h | D) conditional (*aka* posterior) probability**
- **$P(D \mid h) \equiv$ Probability of D Given h**
  - **Measures probability of observing D when 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**

Automated Reasoning using Probability: Inference Tasks

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

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

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

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?

Bayesian Inference: Assessment

- **Answering User Queries**
  - Suppose we want to perform intelligent inferences over a database *DB*
    - Scenario 1: *DB* contains records (instances), some "labeled" with answers
    - 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
    - Correct + result in only 98% of cases in which disease is actually present
    - Correct - result in only 97% of cases in which disease is not present
    - Only 0.008 *of the entire population* has this cancer
  - $\alpha \equiv P$(false negative for $H_0 \equiv Cancer$) = 0.02 (*NB*: for 1-point sample)
  - $\beta \equiv P$(false positive for $H_0 \equiv Cancer$) = 0.03 (*NB*: for 1-point sample)

$$P(Cancer) = 0.008 \quad P(+ \mid Cancer) = 0.98 \quad P(+ \mid \neg Cancer) = 0.03$$
$$P(\neg Cancer) = 0.992 \quad P(- \mid Cancer) = 0.02 \quad P(- \mid \neg Cancer) = 0.97$$

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

Choosing Hypotheses

- **Bayes's Theorem**

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

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

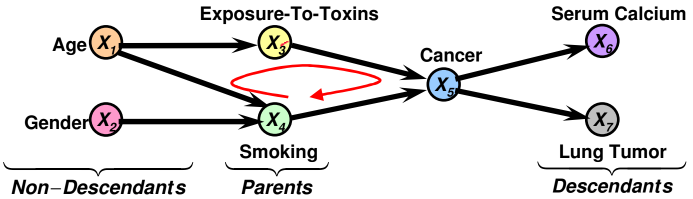
$$h_{MAP} = \arg\max_{h \in H} P(h \mid D) = \arg\max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)} = \arg\max_{h \in H} P(D \mid h)P(h)$$

- **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 maximum likelihood hypothesis, h_{ML}

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

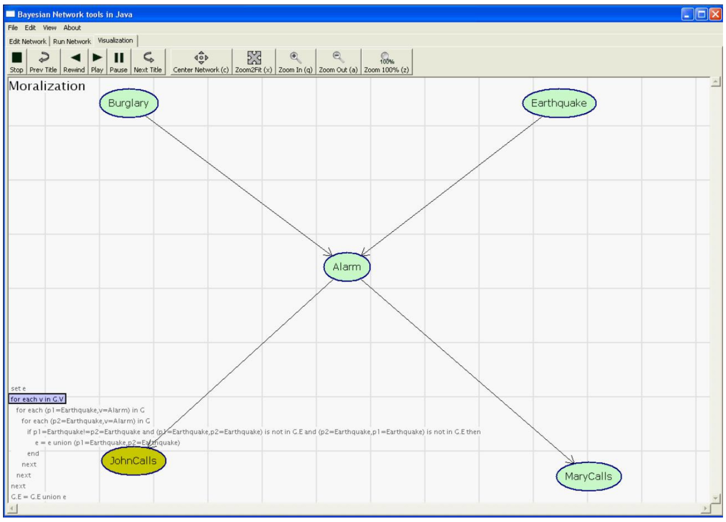
Graphical Models of Probability

- **Conditional Independence**
  - *X* is conditionally independent (CI) from *Y* given *Z* iff $P(X \mid Y, Z) = P(X \mid Z)$ for all values of *X, Y*, and *Z*
  - Example: $P(Thunder \mid Rain, Lightning) = P(Thunder \mid Lightning) \Leftrightarrow T \perp R \mid L$
- **Bayesian (Belief) Network**
  - Acyclic directed graph model $B = (V, E, \Theta)$ representing *CI assertions* over Θ
  - Vertices (nodes) *V*: denote events (each a random variable)
  - Edges (arcs, links) *E*: denote conditional dependencies
- **Markov Condition for BBNs (Chain Rule):** $P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid parents(X_i))$
- **Example BBN**



$P(20s, Female, Low, Non\text{-}Smoker, No\text{-}Cancer, Negative, Negative)$
$= P(T) \cdot P(F) \cdot P(L \mid T) \cdot P(N \mid T, F) \cdot P(N \mid L, N) \cdot P(N \mid N) \cdot P(N \mid N)$

Burglar Network (Pearl, 1986) *aka* Burglar Alarm Network



Burglar aka Burglar Alarm Network

Pearl (1986)

BNJ 3 © 2004 Kansas State University

Semantics of Bayesian Networks

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

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

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

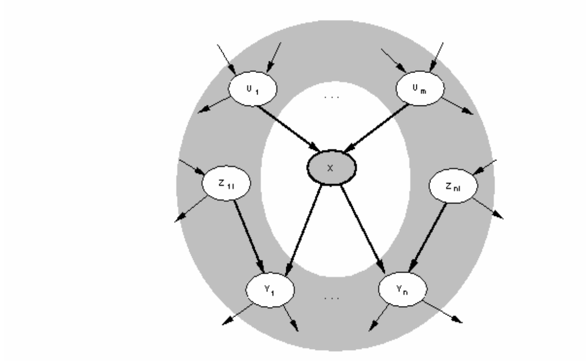
$$= P(\neg B)P(\neg E)P(A|\neg B \wedge \neg E)P(J|A)P(M|A)$$

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

Theorem: Local semantics $\Leftrightarrow$ global semantics

Markov Blanket

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



Constructing Bayesian Networks: Chain Rule in Inference & Learning

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

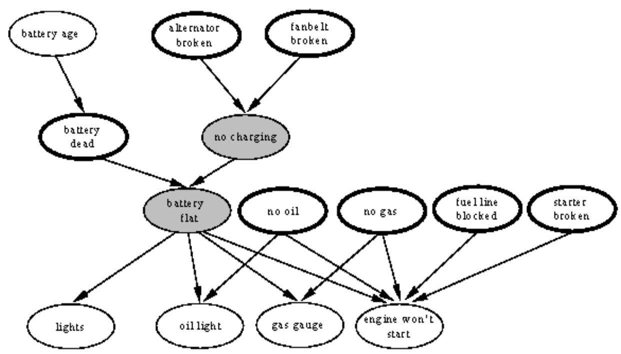
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that
   $\mathbf{P}(X_i | Parents(X_i)) = \mathbf{P}(X_i | X_1, \ldots, X_{i-1})$

This choice of parents guarantees the global semantics:

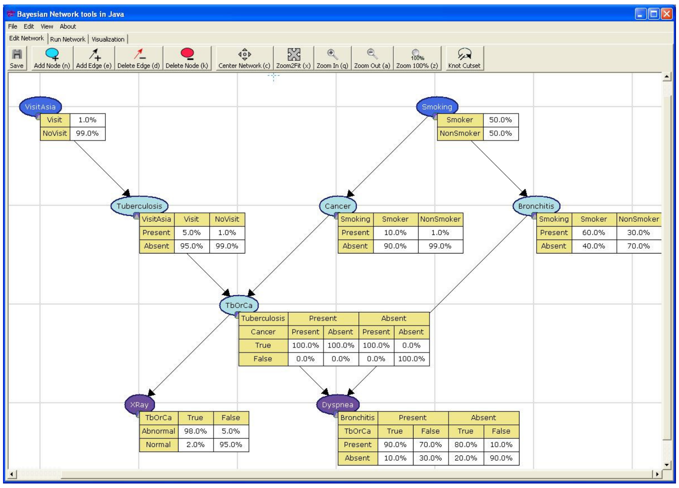
$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | X_1, \ldots, X_{i-1}) \text{ (chain rule)}$$
$$= \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i)) \text{ by construction}$$

Evidential Reasoning: Example – Car Diagnosis

Initial evidence: engine won't start
Testable variables (thin ovals), diagnosis variables (thick ovals)
Hidden variables (shaded) ensure sparse structure, reduce parameters



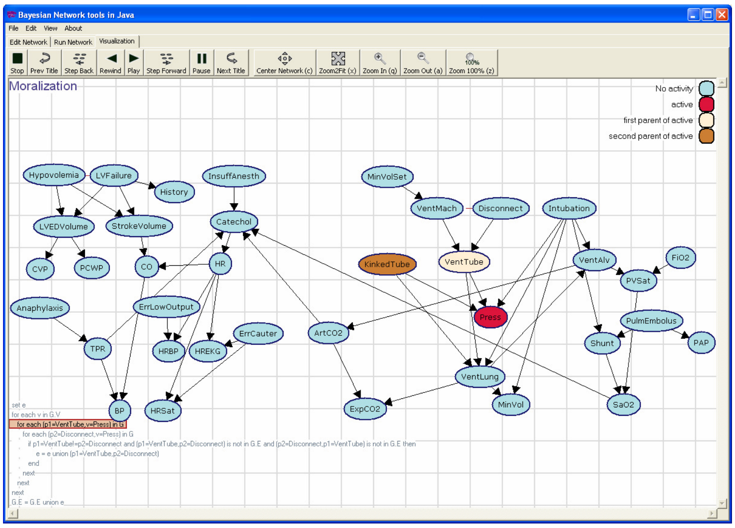
Asia Network
aka Chest Clinic Network



Asia aka Chest Clinic **Network**

BNJ 3 **© 2004 Kansas State University**

ALARM:
A Logical Alarm Reduction Mechanism



ALARM **Network**

© 1990
Cooper *et al.*

BNJ 3 **© 2004 Kansas State University**

Tools for Building Graphical Models

- **Commercial Tools: *Ergo, Netica, TETRAD, Hugin***
- ***Bayes Net Toolbox (BNT)* – Murphy (1997-present)**
  - **Distribution page** http://http.cs.berkeley.edu/~murphyk/Bayes/bnt.html
  - **Development group** http://groups.yahoo.com/group/BayesNetToolbox
- ***Bayesian Network tools in Java (BNJ)* – Hsu *et al.* (1999-present)**
  - **Distribution page** http://bnj.sourceforge.net
  - **Development group** http://groups.yahoo.com/group/bndev
  - **Current (re)implementation projects for KSU KDD Lab**
    - **Continuous state: Minka (2002) – Hsu, Guo, Li**
    - **Formats: XML BNIF (MSBN), Netica – Barber, Guo**
    - **Space-efficient DBN inference – Meyer**
    - **Bounded cutset conditioning – Chandak**

Bayesian Network tools in Java

References: Graphical Models & Inference

- **Graphical Models**
  - **Bayesian (Belief) Networks tutorial – Murphy (2001)**
    http://www.cs.berkeley.edu/~murphyk/Bayes/bayes.html
  - **Learning Bayesian Networks – Heckerman (1996, 1999)**
    http://research.microsoft.com/~heckerman
- **Inference Algorithms**
  - **Junction Tree (Join Tree, L-S, *Hugin*): Lauritzen & Spiegelhalter (1988)**
    http://citeseer.nj.nec.com/huang94inference.html
  - **(Bounded) Loop Cutset Conditioning: Horvitz & Cooper (1989)**
    http://citeseer.nj.nec.com/shachter94global.html
  - **Variable Elimination (Bucket Elimination, *ElimBel*): Dechter (1986)**
    http://citeseer.nj.nec.com/dechter96bucket.html
  - **Recommended Books**
    - **Neapolitan (1990) – *out of print*; see Pearl (1988), Jensen (2001)**
    - **Castillo, Gutierrez, Hadi (1997)**
    - **Cowell, Dawid, Lauritzen, Spiegelhalter (1999)**
  - **Stochastic Approximation**
    http://citeseer.nj.nec.com/cheng00aisbn.html

Terminology

- Uncertain Reasoning: Inference Task with Uncertain Premises, Rules
- Probabilistic Representation
  - Views of probability
    - Subjectivist: measure of belief in sentences
    - Frequentist: likelihood ratio
    - Logicist: counting evidence
  - Founded on Kolmogorov axioms
    - Sum rule
    - Prior, joint *vs.* conditional
    - Bayes's theorem & product rule: P(A | B) = (P(B | A) * P(A)) / P(B)
  - Independence & conditional independence
- Probabilistic Reasoning
  - Inference by enumeration
  - Evidential reasoning

Summary Points

- Last Class: Reasoning under Uncertainty and Probability (Ch. 13)
  - Uncertainty is pervasive
  - What are we uncertain about?
- Today: Chapter 13 Concluded, Preview of Chapter 14
  - Why probability
    - Axiomatic basis: Kolmogorov
    - With utility theory: sound foundation of rational decision making
  - Joint probability
  - Independence
  - Probabilistic reasoning: inference by enumeration
  - Conditioning
    - Bayes's theorem (*aka* Bayes' rule)
    - Conditional independence
- Coming Week: More Applied Probability
  - Graphical models as KR for uncertainty: Bayesian networks, *etc.*
  - Some inference algorithms for Bayes nets