Lecture 26 of 42

Reasoning under Uncertainty: Probability Review & Graphical Models Overview
Discussion: Fuzzy Sets and Soft Computing

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:

Lecture Outline

- Reading for Next Class: Sections 14.1 – 14.2 (p. 492 – 499), R&N 2nd
- Last Class: Uncertainty, Probability, 13 (p. 462-486), R&N 2nd
  - Where uncertainty is encountered: reasoning, planning, learning (later)
  - Sources: sensor error, incomplete/inaccurate domain theory, randomness
- Today: Probability Intro, Continued, Chapter 13, R&N 2nd
  - 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
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 ∈ Ω) is measure of belief in X = x
  ⇒ Pr(X = x) = 0 or Pr(X = x) = 1: plain (aka categorical) beliefs
  ⇒ Can’t be revised; all other beliefs are subject to revision

Kolmogorov Axioms
1. ∀x ∈ Ω. 0 ≤ Pr(X = x) ≤ 1
2. Pr(Ω) = Σx∈Ω Pr(X = x) = 1
3. ∀X₁, X₂, ... ⇒ i ≠ j ⇒ Xᵢ ∩ Xⱼ = ∅.
   \[ Pr \left( \bigcup_{i} X_i \right) = \sum_{i} Pr(X_i) \]
- Joint Probability: Pr(X₁ ∩ X₂) = Prob. of Joint Event X₁ ∩ X₂
- Independence: Pr(X₁ ∩ X₂) = Pr(X₁) • Pr(X₂)
### Inference by Enumeration

Start with the joint distribution:

<table>
<thead>
<tr>
<th></th>
<th>toothache</th>
<th>¬ toothache</th>
</tr>
</thead>
<tbody>
<tr>
<td>catch</td>
<td>0.108</td>
<td>0.012</td>
</tr>
<tr>
<td>¬ catch</td>
<td>0.072</td>
<td>0.008</td>
</tr>
<tr>
<td>cavity</td>
<td>0.016</td>
<td>0.064</td>
</tr>
<tr>
<td>¬ cavity</td>
<td>0.144</td>
<td>0.576</td>
</tr>
</tbody>
</table>

For any proposition φ, sum the atomic events where it is true:

\[
P(\phi) = \sum_{\omega \models \phi} P(\omega)
\]

\[
P(\text{toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
\]

\[
P(\text{cavity} \lor \text{toothache}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28
\]

\[
P(\neg \text{cavity} \lor \text{toothache}) = \frac{P(\neg \text{cavity} \land \text{toothache})}{P(\text{toothache})} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4
\]

Can also compute conditional probabilities.

### Normalization

Denominator can be viewed as a normalization constant α

\[
P(\text{Cavity} | \text{toothache}) = \alpha \frac{P(\text{Cavity, toothache})}{P(\text{toothache})} = \alpha \frac{P(\text{Cavity, toothache})}{\alpha (0.108, 0.016) + \alpha (0.012, 0.064)} = \alpha (0.12, 0.08) = (0.6, 0.4)
\]

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables.

---

Adapted from slides © 2004 S. Russell & P. Norvig. Reused with permission.
**Evidential Reasoning — Inference by Enumeration Approach**

Let $X$ be all the variables. Typically, we want
the posterior joint distribution of the query variables $Y$
given specific values $e$ for the evidence variables $E$

Let the hidden variables be $H = X - Y - E$

Then the required summation of joint entries is done by summing out the hidden variables:

$$P(Y|E=e) = aP(Y, E=e) = a\sum_h P(Y, E=e, H=h)$$

The terms in the summation are joint entries because $Y$, $E$, and $H$ together exhaust the set of random variables

Obvious problems:
1) Worst-case time complexity $O(d^n)$ where $d$ is the largest arity
2) Space complexity $O(d^n)$ to store the joint distribution
3) How to find the numbers for $O(d^n)$ entries? ??

---

**INDEPENDENCE**

$A$ and $B$ are independent iff

$$P(A|B) = P(A) \quad \text{or} \quad P(B|A) = P(B) \quad \text{or} \quad P(A, B) = P(A)P(B)$$

$$P(\text{Toothache, Catch, Cavity, Weather}) = P(\text{Toothache, Catch, Cavity})P(\text{Weather})$$

32 entries reduced to 12; for $n$ independent biased coins, $2^n \rightarrow n$

Absolute independence powerful but rare

Dentistry is a large field with hundreds of variables, none of which are independent. What to do?
**Conditional Independence [1]**

\[ P(\text{Toothache, Cavity, Catch}) \text{ has } 2^3 - 1 = 7 \text{ independent entries} \]

If I have a cavity, the probability that the probe catches in it doesn’t depend on whether I have a toothache:

1. \[ P(\text{catch|toothache, cavity}) = P(\text{catch|cavity}) \]

The same independence holds if I haven’t got a cavity:

2. \[ P(\text{catch|toothache, ~cavity}) = P(\text{catch|~cavity}) \]

**Catch** is conditionally independent of **Toothache** given **Cavity**:

\[ P(\text{Catch|Toothache, Cavity}) = P(\text{Catch|Cavity}) \]

Equivalent statements:

\[ P(\text{Toothache,Catch, Cavity}) = P(\text{Toothache|Cavity}) \]
\[ P(\text{Toothache,Catch|Cavity}) = P(\text{Toothache|Cavity})P(\text{Catch|Cavity}) \]

---

**Conditional Independence [2]**

Write out full joint distribution using chain rule:

\[
\begin{align*}
P(\text{Toothache,Catch, Cavity}) &= P(\text{Toothache,Catch, Cavity})P(\text{Catch,Cavity}) \\
&= P(\text{Toothache,Catch, Cavity})P(\text{Catch|Cavity})P(\text{Cavity}) \\
&= P(\text{Toothache|Cavity})P(\text{Catch|Cavity})P(\text{Cavity})
\end{align*}
\]

I.e., \(2 + 2 + 1 = 5\) independent numbers (equations 1 and 2 remove 2)

In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in \(n\) to linear in \(n\).

Conditional independence is our most basic and robust form of knowledge about uncertain environments.
Bayes' Theorem (aka Bayes' Rule)

Product rule: \( P(a \land b) = P(a|b)P(b) = P(b|a)P(a) \)

\[ \Rightarrow \text{Bayes' rule } P(a|b) = \frac{P(b|a)P(a)}{P(b)} \]

or in distribution form:

\[ P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \alpha P(X|Y)P(Y) \]

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!

Bayes' Rule & Conditional Independence

\[ P(\text{Cavity} \land \text{toothache} \land \text{catch}) = \alpha P(\text{toothache} \land \text{catch}|\text{Cavity})P(\text{Cavity}) = \alpha P(\text{toothache}|\text{Cavity})P(\text{catch}|\text{Cavity})P(\text{Cavity}) \]

This is an example of a naive Bayes model:

\[ P(\text{Cause}, \text{Effect}_1, \ldots, \text{Effect}_n) = P(\text{Cause})P(\text{Effect}_1|\text{Cause}) \]

Total number of parameters is linear in \( n \)

Uncertain Reasoning Roadmap

- **Framework:** Interpretations of Probability [Cheeseman, 1985]
  - **Bayesian subjectivist view**
    - Measure of agent’s belief in proposition
    - Proposition denoted by random variable (range: sample space $\Omega$)
    - e.g., $Pr(\text{Outlook} = \text{Sunny}) = 0.8$
  - **Frequentist view:** probability is frequency of observations of event
  - **Logicist view:** probability is inferential evidence in favor of 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 ($p=1$), categorical rejection ($p=0$)
    - Conditional beliefs: depends on reviewer (use probabilistic model)
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


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

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 = Fuzziness of Meaning

Adapted from slide © 2008 L. A. Zadeh, UC Berkeley

Basic Formulas For Probabilities

- **Product Rule** (Alternative Statement of Bayes's Theorem)
  \[ P(A \mid B) = \frac{P(A \cap B)}{P(B)} \]
  - Proof: requires axiomatic set theory, as does Bayes's Theorem

- **Sum Rule**
  \[ P(A \cup B) = P(A) + P(B) - P(A \cap B) \]
  - 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 \( A \cup B \)

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

Basic Formulas For Probabilities
Bayes’s Theorem: Joint vs. Conditional Probability

- **Theorem**
  \[
P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} = \frac{P(h \wedge D)}{P(D)}
  \]
- **\(P(h)\)**: Prior Probability of Assertion (Hypothesis) \(h\)
  - Measures initial beliefs (BK) before any information is obtained (hence prior)
- **\(P(D)\)**: Prior Probability of Data (Observations) \(D\)
  - Measures probability of obtaining sample \(D\) (i.e., expresses \(D\))
- **\(P(h \mid D)\)**: Probability of \(h\) Given \(D\)
  - \(\mid\) denotes conditioning - hence \(P(h \mid D)\) conditional (aka posterior) probability
- **\(P(D \mid h)\)**: Probability of \(D\) Given \(h\)
  - Measures probability of observing \(D\) when \(h\) is correct (“generative” model)
- **\(P(h \wedge D)\)**: 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 \(P(X_i \mid E = e)\)
  - e.g., \(P(\text{NoGas} \mid \text{Gauge = empty, Lights = on, Starts = false})\)

Conjunctive queries: \(P(X_i, X_j \mid E = e) = P(X_i \mid E = e)P(X_j \mid X_i, E = e)\)

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

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?

Adapted from slide © 2004 S. Russell & P. Norvig. Reused with permission.
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
  - \( \alpha \equiv P(\text{false negative for } H_0 \equiv \text{Cancer}) = 0.02 \) (NB: for 1-point sample)
  - \( \beta \equiv P(\text{false positive for } H_0 \equiv \text{Cancer}) = 0.03 \) (NB: for 1-point sample)
  - \( P(\text{Cancer}) = 0.008 \quad P(+ \mid \text{Cancer}) = 0.98 \quad P(+ \mid \neg \text{Cancer}) = 0.03 \)
  - \( P(\neg \text{Cancer}) = 0.992 \quad P(\neg + \mid \text{Cancer}) = 0.02 \quad P(\neg + \mid \neg \text{Cancer}) = 0.97 \)
  - \( P(+ \mid H_0 \land H_0) = 0.0078, P(+ \mid H_0 \land H_0) = 0.0298 \Rightarrow h_{\text{MAP}} = H_A = \neg \text{Cancer} \)

Choosing Hypotheses

- Bayes’s Theorem
  \[
P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)} = \frac{P(h \land D)}{P(D)}
\]

- MAP Hypothesis
  - Generally want most probable hypothesis given training data
  - Define: \( \arg \max_{x \in D} f(x) \) = value of \( x \) in sample space \( D \) with highest \( f(x) \)
  - Maximum a posteriori hypothesis, \( h_{\text{MAP}} \)
    \[
h_{\text{MAP}} = \arg \max_{h \in H} P(h \mid D)
    = \arg \max_{h \in H} \frac{P(D \mid h)P(h)}{P(D)}
    = \arg \max_{h \in H} P(D \mid h)P(h)
\]

- ML Hypothesis
  - Assume that \( p(h_i) = p(h_j) \) for all pairs \( i, j \) (uniform priors, i.e., \( P_H \sim \text{Uniform} \))
  - Can further simplify and choose maximum likelihood hypothesis, \( h_{\text{ML}} \)
    \[
h_{\text{ML}} = \arg \max_{h \in H} P(D \mid h_i)
\]
Graphical Models of Probability

1. Conditional Independence
   - X is conditionally independent (CI) from Y given Z if \( P(X \mid Y, Z) = P(X \mid Z) \) for all values of X, Y, and Z
   - Example: \( P(\text{Thunder} \mid \text{Rain, Lightning}) = P(\text{Thunder} \mid \text{Lightning}) \iff T \perp R \mid L \)

2. 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): \( P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{parents}(X_i)) \)

Example BBN

Bayesian Network (Pearl, 1986) aka Burglar Alarm Network

Example BBN

\[
P(\text{Burglar} = \text{Yes}) = P(\text{Burglar} = \text{No})
\]

\[
P(\text{Burglar Alarm}) = P(\text{Burglar Alarm} = \text{Yes})
\]

\[
P(\text{Burglar Alarm} = \text{No})
\]

\[
P(\text{Burglar Alarm} = \text{Yes} \mid \text{Burglar} = \text{Yes}) = P(\text{Burglar Alarm} = \text{Yes} \mid \text{Burglar} = \text{No})
\]

\[
P(\text{Burglar Alarm} = \text{No} \mid \text{Burglar} = \text{Yes}) = P(\text{Burglar Alarm} = \text{No} \mid \text{Burglar} = \text{No})
\]

\[
P(\text{Burglar Alarm} = \text{No})
\]

\[
P(\text{Burglar Alarm} = \text{Yes} \mid \text{Burglar} = \text{Yes}) = P(\text{Burglar Alarm} = \text{Yes} \mid \text{Burglar} = \text{No})
\]

\[
P(\text{Burglar Alarm} = \text{No} \mid \text{Burglar} = \text{Yes}) = P(\text{Burglar Alarm} = \text{No} \mid \text{Burglar} = \text{No})
\]

\[
P(\text{Burglar Alarm} = \text{Yes}) = P(\text{Burglar Alarm} = \text{No})
\]

\[
P(\text{Burglar Alarm} = \text{Yes} \mid \text{Burglar} = \text{Yes}) = P(\text{Burglar Alarm} = \text{Yes} \mid \text{Burglar} = \text{No})
\]

\[
P(\text{Burglar Alarm} = \text{No} \mid \text{Burglar} = \text{Yes}) = P(\text{Burglar Alarm} = \text{No} \mid \text{Burglar} = \text{No})
\]

\[
P(\text{Burglar Alarm} = \text{Yes}) = P(\text{Burglar Alarm} = \text{No})
\]
Semantics of Bayesian Networks

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

\[ P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | \text{Parents}(X_i)) \]

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

\[ = P(\neg B)P(\neg E)P(A | \neg B \land \neg E)^{\forall}P(J | A)P(M | A) \]

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

This choice of parents guarantees the global semantics:

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

$$= \prod_{i=1}^{n} P(X_i|\text{Parents}(X_i)) \quad \text{by construction}$$

---

**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

---

Adapted from slide © 2004 S. Russell & P. Norvig. Reused with permission.
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

REFERENCES:
GRAPHICAL MODELS & INFERENCE

- Graphical Models
  - Bayesian (Belief) Networks tutorial – Murphy (2001)
    http://www.cs.berkeley.edu/~murphyk/Bayes/bayes.html
    http://research.microsoft.com/~heckerman

- Inference Algorithms
    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**: \[ P(A | B) = \frac{P(B | A) \times P(A)}{P(B)} \]
    - **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