

CmpE 540 Principles of Artificial Intelligence

Pınar Yolum
pinar.yolum@boun.edu.tr

Department of
Computer Engineering
Boğaziçi University

Reasoning under Uncertainty

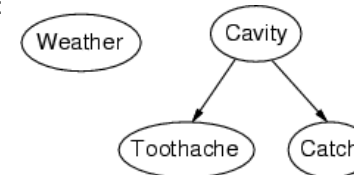
Bayesian Networks

A directed graph in which the following holds:

1. A set of random variables makes up the nodes of the network. Variables may be discrete or continuous.
2. A set of directed links or arrows connects pairs of nodes. If there is an arrow from node X to node Y , X is said to be a parent of Y .
3. Each node X_i has an associated conditional probability distribution $P(X_i | \text{Parents}(X_i))$ that quantifies the effect of the parents on the node.
4. The graph has no directed cycles (hence is a directed, acyclic graph, or DAG).

Example

- Topology of network encodes conditional independence assertions:



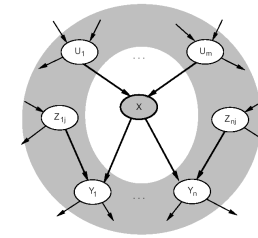
- *Weather* is independent of the other variables
- *Toothache* and *Catch* are conditionally independent given *Cavity*

Properties of Bayesian Networks

- Space-efficient data structure for encoding all of the information in the **full joint probability distribution** for the set of random variables defining a domain.
- Represents all of the direct causal relationships between variables
- Space efficient because it exploits the fact that in many real-world problem domains the dependencies between variables are generally local, so there are a lot of conditionally independent variables
- Can be used to reason

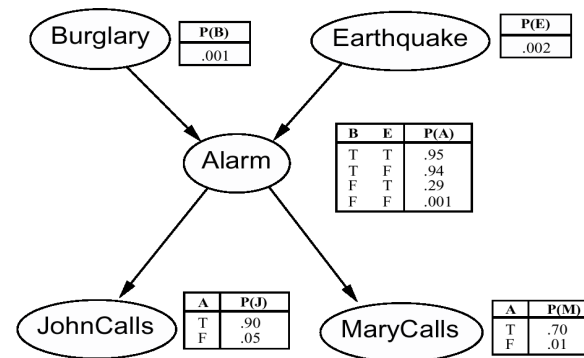
Markov blanket

- Each node is conditionally independent of all others given its **Markov blanket**:
parents + children + childrens parents



Burglary Example

Burglary Example



Reasoning with Bayesian Networks

- Inference in Bayesian networks means computing the probability distribution of a set of query variables, given a set of evidence variables.
- **predictive reasoning (causal reasoning)**
Forward (top-down) from causes to effects
- **diagnostic reasoning**
Backward (bottom-up) from effects to causes

Semantics

- **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

Example

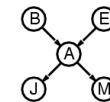
- The probability of the event that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both John and Mary call.

$$\begin{aligned} &P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) \\ &= P(j|a)P(m|a)P(a|\neg b \wedge \neg e)P(\neg b)P(\neg e) \\ &= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.00062 \end{aligned}$$

Compactness

- A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values

- Each row requires one number p for $X_i = \text{true}$ (the number for $X_i = \text{false}$ is just $1-p$)



- If each variable has no more than k parents, the complete network requires $O(n \cdot 2^k)$ numbers

- I.e., grows linearly with n , vs. $O(2^n)$ for the full joint distribution

- For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

Constructing Belief Networks

- 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:

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

Example

- Suppose we choose the ordering M, J, A, B, E



$$\mathbf{P}(J | M) = \mathbf{P}(J)?$$

Example

- Suppose we choose the ordering M, J, A, B, E

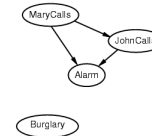


$$\mathbf{P}(J | M) = \mathbf{P}(J)? \text{ No}$$

$$\mathbf{P}(A | J, M) = \mathbf{P}(A | J)? \quad \mathbf{P}(A | J, M) = \mathbf{P}(A)?$$

Example

- Suppose we choose the ordering M, J, A, B, E



$$\mathbf{P}(J | M) = \mathbf{P}(J)? \text{ No}$$

$$\mathbf{P}(A | J, M) = \mathbf{P}(A | J)? \quad \mathbf{P}(A | J, M) = \mathbf{P}(A)? \text{ No}$$

$$\mathbf{P}(B | A, J, M) = \mathbf{P}(B | A)?$$

$$\mathbf{P}(B | A, J, M) = \mathbf{P}(B)?$$

Example

- Suppose we choose the ordering M, J, A, B, E



$P(J | M) = P(J)$? **No**

$P(A | J, M) = P(A | J)$? $P(A | J, M) = P(A)$? **No**

$P(B | A, J, M) = P(B | A)$? **Yes**

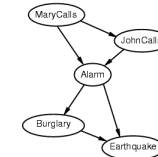
$P(B | A, J, M) = P(B)$? **No**

$P(E | B, A, J, M) = P(E | A)$?

$P(E | B, A, J, M) = P(E | A, B)$?

Example

- Suppose we choose the ordering M, J, A, B, E



$P(J | M) = P(J)$? **No**

$P(A | J, M) = P(A | J)$? $P(A | J, M) = P(A)$? **No**

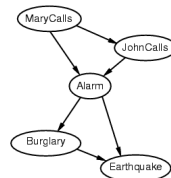
$P(B | A, J, M) = P(B | A)$? **Yes**

$P(B | A, J, M) = P(B)$? **No**

$P(E | B, A, J, M) = P(E | A)$? **No**

$P(E | B, A, J, M) = P(E | A, B)$? **Yes**

Example contd.



- Deciding conditional independence is hard in noncausal directions
- (Causal models and conditional independence seem hardwired for humans!)
- Network is less compact: $1 + 2 + 4 + 2 + 4 = 13$ numbers needed

Fuzzy Logic

Fuzzy Thinking

- Experts rely on common sense when they solve problems.
- How can we represent expert knowledge that uses vague terms?
- Fuzzy logic is not logic that is fuzzy, but logic that is used to describe fuzziness.
- Fuzzy logic is the theory of fuzzy sets, sets that calibrate vagueness.
- Fuzzy logic is based on the idea that all things admit of degrees.
 - Temperature, height, speed, distance, beauty - all come on a sliding scale.
 - The motor is running really hot.
 - Ali is a very tall person.

Fuzzy Logic

- Set of mathematical principles for knowledge representation based on degrees of membership.
- Unlike two-valued Boolean logic, fuzzy logic is **multi-valued**.
- It deals with **degrees of membership** and **degrees of truth**.
- Fuzzy logic uses the continuum of logical values between 0 (completely false) and 1 (completely true). Instead of just black and white, it employs the spectrum of colors, accepting that things can be partly true and partly false at the same time.

Why use fuzzy logic?

Pros:

- Tolerant of imprecise data
- Universal approximation: can model arbitrary nonlinear functions
- Based on linguistic terms
- Convenient way to express expert and common sense knowledge

Cons:

- Not a cure-all
- Crisp/precise models can be more efficient and even convenient
- Other approaches might be formally verified to work

Fuzzy sets

- **Boolean/Crisp set A** is a mapping for the elements of S to the set {0, 1}, i.e., $A: S \rightarrow \{0, 1\}$
- Characteristic function:

$$\mu_A(x) = \begin{cases} 1 & \text{if } x \text{ is an element of set } A \\ 0 & \text{if } x \text{ is not an element of set } A \end{cases}$$

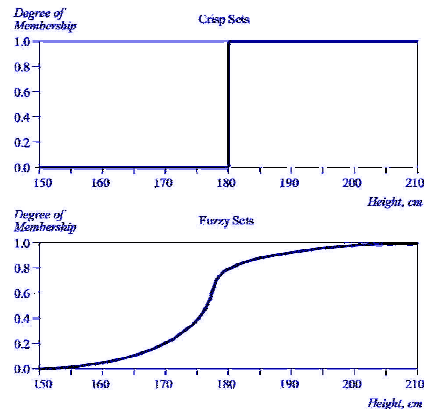
Fuzzy sets

- **Fuzzy set F** is a mapping for the elements of S to the interval [0, 1], i.e., $F: S \rightarrow [0, 1]$
- Characteristic function: $0 \leq \mu_F(x) \leq 1$
- 1 means full membership, 0 means no membership and anything in between, e.g., 0.5 is called **graded membership**

Fuzzy Set Example: Tall Men

Name	Height, cm	Degree of Crisp	Membership Fuzzy
Ali	208	1	1.00
Cemal	205	1	1.00
Hasan	198	1	0.98
İsmail	181	1	0.82
Orhan	179	0	0.78
Ahmet	172	0	0.24
Abdullah	167	0	0.15
Barış	158	0	0.06
Çetin	155	0	0.01
Sezai	152	0	0.00

Fuzzy Set Example: Tall Men

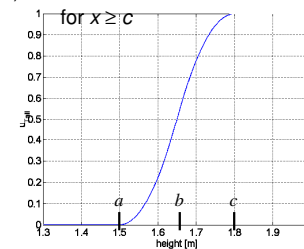


Membership functions: S-function

- The S-function can be used to define fuzzy sets

- $S(x, a, b, c) =$

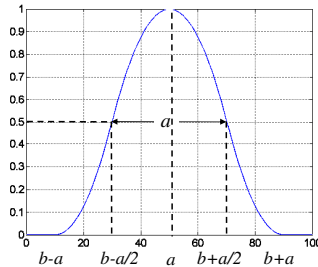
- 0 for $x \leq a$
- $2(x-a/c-a)^2$ for $a \leq x \leq b$
- $1 - 2(x-c/c-a)^2$ for $b \leq x \leq c$
- 1 for $x \geq c$



Membership functions: P-Function

- $P(x, a, b) =$
 - $S(x, b-a, b-a/2, b)$ for $x \leq b$
 - $1 - S(x, b, b+a/2, a+b)$ for $x \geq b$

E.g., **close** (to a)



Fuzzy set operators

- Equality $A = B$
 $\mu_A(x) = \mu_B(x)$ for all $x \in X$
- Complement A'
 $\mu_{A'}(x) = 1 - \mu_A(x)$ for all $x \in X$
- Containment $A \subseteq B$
 $\mu_A(x) \leq \mu_B(x)$ for all $x \in X$
- Union $A \cup B$
 $\mu_{A \cup B}(x) = \max(\mu_A(x), \mu_B(x))$ for all $x \in X$
- Intersection $A \cap B$
 $\mu_{A \cap B}(x) = \min(\mu_A(x), \mu_B(x))$ for all $x \in X$

Fuzzy inference

- Fuzzy logical operations
- Fuzzy rules
- Fuzzification
- Implication
- Aggregation
- Defuzzification

Fuzzy Inference

- Process of mapping from a given input to an output using fuzzy set theory
- In Fuzzy Logic system all rules fire in parallel
- Rules may fire partially
- **Monotonic Selection** - Truth membership grade of rule consequent can be estimated directly from the truth membership grade in the antecedent

Fuzzy Inference

- Rules may have multiple antecedents
 - All parts are calculated simultaneously and resolved to a single number using set operations
- Rules may have multiple consequents
 - All parts are affected equally by the antecedents
 - The output of each rule is a fuzzy set
 - Need to obtain single crisp number representing expert system output
 - Aggregate output fuzzy sets into single output fuzzy set
 - Defuzzify resulting set into single number

Fuzzy Rules

- Conclusions that fuzzy systems arrive at are fuzzy facts with degrees of membership
 - E.g. risk is low with membership of 0.5
- Outcome must however be a concrete decision e.g. loan money etc
- Process of transforming fuzzy fact into crisp fact is defuzzification

Defuzzification

- Converts fuzzy value into single crisp value
- Fuzzy set may not be easily translated into crisp values
- Methods
 - Max-membership
 - Centroid method
 - Weighted average method

Fuzzy logical operations

- AND, OR, NOT, etc.
- **NOT** $A = A' = 1 - \mu_A(x)$
- **A AND B** $= A \cap B = \min(\mu_A(x), \mu_B(x))$
- **A OR B** $= A \cup B = \max(\mu_A(x), \mu_B(x))$

min(A,B)			max(A,B)			1-A	
A	B	A and B	A	B	A or B	A	not A
0	0	0	0	0	0	0	1
0	1	0	0	1	1	1	0
1	0	0	1	0	1	1	0
1	1	1	1	1	1	1	0

If-Then Rules

- Use fuzzy sets and fuzzy operators as the **subjects** and **verbs** of fuzzy logic to form rules.

if x is A then y is B

where A and B are linguistic terms defined by fuzzy sets on the sets X and Y respectively.

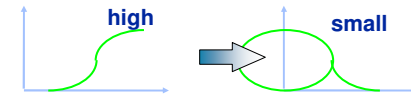
This reads

if x == A then y = B

Fuzzy If-Then Rules

- Mamdani style**

If pressure is high then volume is small



- Sugeno style**

If speed is medium then resistance = 5*speed



Evaluation of fuzzy rules

- In Boolean logic: $p \Rightarrow q$
if p is true then q is true
- In fuzzy logic: $p \Rightarrow q$
if p is true to some degree then q is true to some degree.
 $0.5p \Rightarrow 0.5q$ (partial premise implies partially)
- How?

Evaluation of fuzzy rules (cont'd)

- Apply implication function to the rule
- Most common way is to use min to “chop-off” the consequent (prod can be used to scale the consequent)

Summary: If-Then rules

- Fuzzify inputs
Determine the degree of membership for all terms in the premise.
If there is one term then this is the degree of support for the consequence.
- Apply fuzzy operator
If there are multiple parts, apply logical operators to determine the degree of support for the rule.
- Apply implication method
Use degree of support for rule to shape output fuzzy set of the consequence.
- How do we then combine several rules?

Multiple rules

- We aggregate the outputs into a single fuzzy set which combines their decisions.
- The input to aggregation is the list of truncated fuzzy sets and the output is a single fuzzy set for each variable.
- **Aggregation rules:** max, sum, etc.
- As long as it is commutative then the order of rule execution is irrelevant.

max-min rule of composition

- Given N observations E_i over X and hypothesis H_i over Y we have N rules:

if E_1 then H_1

if E_2 then H_2

if E_N then H_N

$$\forall \mu_H = \max[\min(\mu_{E_1}), \min(\mu_{E_2}), \dots, \min(\mu_{E_N})]$$

Defuzzify the output

- Take a fuzzy set and produce a single crisp number that represents the set.
- Practical when making a decision, taking an action etc.

Fuzzy Inference System

- Mamdani-style inference
 - Step 1: Fuzzification of input variables
 - Step 2: Rule Evaluation
 - Step 3: Aggregation of Rule Outputs
 - Step 4: Defuzzification

Fuzzy Inference - Mamdani

Rule1:	IF x is A3 OR y is B1 THEN z is C1	Rule1:	IF project_funding is adequate OR project_staffing is small THEN risk is low
Rule2:	IF x is A2 AND y is B2 THEN z is C2	Rule2:	IF project_funding is marginal AND project_staffing is large THEN risk is normal
Rule3:	IF x is A1 THEN z is C3	Rule3:	IF project_funding is inadequate THEN risk is high

x = project funding,
A1=inadequate, A2=marginal, A3=adequate on universe of discourse X
y=project staffing,
B1=small, B2=large on universe of discourse Y
z=risk
C1=low, C2=normal, C3=high on universe of discourse Z

Fuzzy Inference - Mamdani

- Step1 : Fuzzification
 - Take crisp inputs x_1 and y_1 and determine the degree to which inputs belong to fuzzy sets
 - x_1 and y_1 are limited to universes of discourse X and Y
 - Experts determine range of universe
 - Some can be measured directly, others only on basis of expert opinion
- Our example
 - Universe X and Y are from 0 to 100%
 - Suppose our crisp input x_1 is 35% and has been rated as falling into inadequate and marginal to degrees of 0.5 and 0.2 respectively
 - Crisp input y_1 is 60% which falls into small and large with degrees of 0.1 and 0.7 respectively

Fuzzy Inference - Mamdani

- Step 2: Rule Evaluation
 - Take fuzzified inputs and apply them to antecedents of rules
 - If rule has multiple antecedents then fuzzy operator is used to get single number to represent result
 - Result is then applied to consequent membership function
 - Result can be produced by **clipping or scaling**

Fuzzy Inference - Mamdani

- Rule1:
 - IF x is A3 (0.0) or y is B1(0.1) THEN z is C1(0.1)
- Rule2:
 - IF x is A2(0.2) AND y is B2(0.7) THEN z is C2(0.2)
- Rule3:
 - IF x is A1(0.5) THEN z is C3(0.5)

Fuzzy Inference - Mamdani

- Step3: Aggregation of Rule outputs
 - Unification of outputs of all rules
 - Take rule outputs and combine into a single fuzzy set
 - One fuzzy set for each variable
- Our Example
 - z is C1 (0.1), z is C2(0.2), z is C3(0.5)

Fuzzy Inference - Mamdani

- Step4: Defuzzification
 - Need to provide a crisp number
 - Most popular technique is centroid technique
 - Finds the point where a vertical line would slice the aggregate set into two equal masses

$$COG = \frac{\sum \mu_i(a) \cdot x_i}{\sum \mu_i(a)}$$

Fuzzy Inference- Mamdani

- Our Problem

$$COG = \frac{(0+10+20) \times 0.1 + (30+40+50+60) \times 0.2 + (70+80+90+100) \times 0.5}{0.1+0.1+0.1+0.2+0.2+0.2+0.2+0.5+0.5+0.5+0.5}$$

= 67.4

So

$$z = 67.4$$

Limitations of fuzzy logic

- How to determine the membership functions? Usually requires fine-tuning of parameters
- Defuzzification can produce undesired results

Distinctions to Probabilities

- Why both fuzzy sets and probabilities use real numbers to “describe” a degree of membership they differ:
 - membership functions are not necessarily based on statistic distributions
 - fuzzy logic deals with deterministic plausibilities
 - probabilities deal more with non-deterministic but stochastic events and their likelihoods

Dempster-Shafer Theory

Dempster-Shafer Theory

- mathematical theory of evidence
 - uncertainty is modeled through a range of probabilities
 - instead of a single number indicating a probability
 - sound theoretical foundation
 - allows distinction between belief, disbelief, ignorance (non-belief)
 - certainty factors are a special case of DS theory

Frame of Discernment

- *Universe of Discourse* θ , also called a *Frame of Discernment*, is a set of mutually exclusive alternatives.
 - Given the example of determining the disease of a patient, θ would be the set consisting of all possible diseases.
- Subsets of θ are the class of general propositions in the domain.
 - For example, the proposition "The disease is infectious" corresponds to the set of the elements of θ which are infectious, i.e. {"Influenza", "Small Pox", ...}.

DS Theory Notation

- *environment* $\Theta = \{O_1, O_2, \dots, O_n\}$
 - set of objects O_i that are of interest
 - $\Theta = \{O_1, O_2, \dots, O_n\}$
- *frame of discernment FD (Universe of Discourse)*
 - A set of mutually exclusive alternatives
- *mass probability function m*
 - assigns a value from $[0, 1]$ to every item
 - describes the degree of belief in analogy to the mass of a physical object
- *mass probability m(A)*
 - portion of the total mass probability that is assigned to a specific element A of FD

Belief and Certainty

- *belief* $Bel(A)$ in a set A
 - sum of the mass probabilities of all the proper subsets of A
 - all the mass that supports A
 - likelihood that one of its members is the conclusion
 - also called support function
- *plausibility* $Pls(A)$
 - maximum belief of A
 - upper bound for the range of belief
- *certainty* $Cer(A)$
 - interval $[Bel(A), Pls(A)]$
 - also called evidential interval
 - expresses the range of belief

Combination of Mass Probabilities

- combining two masses in such a way that the new mass represents a consensus of the contributing pieces of evidence
 - set intersection puts the emphasis on common elements of evidence, rather than conflicting evidence
- $m_1 \oplus m_2 (C) = \sum_{X \cap Y} m_1(X) * m_2(Y)$
 $= C m_1(X) * m_2(Y) / (1 - \sum X \cap Y)$
 $= C m_1(X) * m_2(Y)$

where

X, Y are hypothesis subsets and
 C is their intersection $C = X \cap Y$
 \oplus is the orthogonal or direct sum

Differences Probabilities - DF Theory

Aspect	Probabilities	Dempster-Shafer
Aggregate Sum	$\sum_i P_i = 1$	$m(\Theta) \leq 1$
Subset $X \subseteq Y$	$P(X) \leq P(Y)$	$m(X) > m(Y)$ allowed
relationship $X, \neg X$	$P(X) + P(\neg X) = 1$	$m(X) + m(\neg X) \leq 1$

(ignorance)

Evidential Reasoning

- extension of DS theory that deals with uncertain, imprecise, and possibly inaccurate knowledge
- also uses evidential intervals to express the confidence in a statement
 - lower bound is called support (Spt) in evidential reasoning, and belief (Bel) in Dempster-Shafer theory
 - upper bound is plausibility (Pls)

Evidential Intervals

Meaning	Evidential Interval
Completely true	[1, 1]
Completely false	[0, 0]
Completely ignorant	[0, 1]
Tends to support	[Bel, 1] where $0 < \text{Bel} < 1$
Tends to refute	[0, Pls] where $0 < \text{Pls} < 1$
Tends to both support and refute	[Bel, Pls] where $0 < \text{Bel} \leq \text{Pls} < 1$

refute

Bel: belief; lower bound of the evidential interval

Pls: plausibility; upper bound

Advantages and Problems of Dempster-Shafer

- advantages
 - clear, rigorous foundation
 - ability to express confidence through intervals
 - certainty about certainty
 - proper treatment of ignorance
- problems
 - non-intuitive determination of mass probability
 - very high computational overhead
 - may produce counterintuitive results due to normalization