Abduction, Uncertainty, and Probabilistic Reasoning

Chapter 15 and more
Introduction

• **Abduction** is a reasoning process that tries to form plausible explanations for abnormal observations
  – Abduction is distinct different from deduction and induction
  – Abduction is inherently uncertain
• Uncertainty becomes an important issue in AI research
• Some major formalisms for representing and reasoning about uncertainty
  – Mycin’s certainty factor (an early representative)
  – Probability theory (esp. Bayesian belief networks)
  – Dempster-Shafer theory
  – Fuzzy logic
  – Truth maintenance systems
Abduction

• **Definition** (Encyclopedia Britannica): reasoning that derives an explanatory hypothesis from a given set of facts
  – The inference result is a *hypothesis*, which if true, could *explain* the occurrence of the given facts

• **Examples**
  – Dendral, an expert system to construct 3D structure of chemical compounds
    • Fact: mass spectrometer data of the compound and its chemical formula
    • KB: chemistry, esp. strength of different types of bounds
    • Reasoning: form a hypothetical 3D structure which meet the given chemical formula, and would most likely produce the given mass spectrum if subjected to electron beam bombardment
– Medical diagnosis
  • Facts: symptoms, lab test results, and other observed findings (called manifestations)
  • KB: causal associations between diseases and manifestations
  • Reasoning: one or more diseases whose presence would causally explain the occurrence of the given manifestations

– Many other reasoning processes (e.g., word sense disambiguation in natural language process, image understanding, detective’s work, etc.) can also been seen as abductive reasoning.
Comparing abduction, deduction and induction

**Deduction:** major premise: All balls in the box are black
minor premise: These balls are from the box
conclusion: These balls are black

**Abduction:** rule: All balls in the box are black
observation: These balls are black
explanation: These balls are from the box

**Induction:** case: These balls are from the box
observation: These balls are black
hypothesized rule: All ball in the box are black

Induction: from specific cases to general rules
Abduction and deduction:
both from part of a specific case to other part of the case using general rules (in different ways)
Characteristics of abduction reasoning

1. Reasoning results are **hypotheses**, not theorems (may be false *even if* rules and facts are true),
   - e.g., misdiagnosis in medicine
2. There may be multiple plausible hypotheses
   - When given rules $A \Rightarrow B$ and $C \Rightarrow B$, and fact $B$
     both $A$ and $C$ are plausible hypotheses
   - Abduction is inherently uncertain
   - Hypotheses can be ranked by their plausibility if that can be determined
3. Reasoning is often a Hypothesize- and-test cycle
   - hypothesize phase: postulate possible hypotheses, each of which could explain the given facts (or explain most of the important facts)
   - test phase: test the plausibility of all or some of these hypotheses
One way to test a hypothesis $H$ is to query if some thing that is currently unknown but can be predicted from $H$ is actually true.

- If we also know $A \Rightarrow D$ and $C \Rightarrow E$, then ask if $D$ and $E$ are true.
- If it turns out $D$ is true and $E$ is false, then hypothesis $A$ becomes more plausible (support for $A$ increased, support for $C$ decreased)

### 4. Reasoning is non-monotonic

- Plausibility of hypotheses can increase/decrease as new facts are collected (deductive inference determines if a sentence is true but would never change its truth value)
- Some hypotheses may be discarded, and new ones may be formed when new observations are made
Source of Uncertainty

• Uncertain data (noise)
• Uncertain knowledge (e.g, causal relations)
  – A disorder may cause any and all POSSIBLE manifestations in a specific case
  – A manifestation can be caused by more than one POSSIBLE disorders
• Uncertain reasoning results
  – Abduction and induction are inherently uncertain
  – Default reasoning, even in deductive fashion, is uncertain
  – Incomplete deductive inference may be uncertain
Probabilistic Inference

• Based on probability theory (especially Bayes’ theorem)
  – Well established discipline about uncertain outcomes
  – Empirical science like physics/chemistry, can be verified by experiments

• Probability theory is too rigid to apply directly in many applications
  – Some assumptions have to be made to simplify the reality
  – Different formalisms have been developed in which some aspects of the probability theory are changed/modified.

• We will briefly review the basics of probability theory before discussing different approaches to uncertainty

• The presentation uses diagnostic process (an abductive and evidential reasoning process) as an example
Probability of Events

• Sample space and events
  – Sample space $S$: (e.g., all people in an area)
  – Events $E_1 \subseteq S$: (e.g., all people having cough)
    $E_2 \subseteq S$: (e.g., all people having cold)

• Prior (marginal) probabilities of events
  – $P(E) = \frac{|E|}{|S|}$ (frequency interpretation)
  – $P(E) = 0.1$ (subjective probability)
  – $0 \leq P(E) \leq 1$ for all events
  – Two special events: $\emptyset$ and $S$: $P(\emptyset) = 0$ and $P(S) = 1.0$

• Boolean operators between events (to form compound events)
  – Conjunctive (intersection): $E_1 \cap E_2$ ($E_1 \cap E_2$)
  – Disjunctive (union): $E_1 \cup E_2$ ($E_1 \cup E_2$)
  – Negation (complement): $\sim E$ ($E^C = S - E$)
• Probabilities of compound events
  – $P(\sim E) = 1 - P(E)$ because $P(\sim E) + P(E) = 1$
  – $P(E_1 \lor E_2) = P(E_1) + P(E_2) - P(E_1 \land E_2)$
  – But how to compute the joint probability $P(E_1 \land E_2)$?

• Conditional probability (of $E_1$, given $E_2$)
  – How likely $E_1$ occurs in the subspace of $E_2$

\[
P(E_1|E_2) = \frac{|E_1 \land E_2|}{|E_2|} = \frac{|E_1 \land E_2|}{|E_2|/|S|} = \frac{P(E_1 \land E_2)}{P(E_2)}
\]

\[
P(E_1 \land E_2) = P(E_1|E_2)P(E_2)
\]
• Independence assumption
  – Two events $E_1$ and $E_2$ are said to be independent of each other if
    \[ P(E_1 \mid E_2) = P(E_1) \] (given $E_2$ does not change the likelihood of $E_1$)
  – It can simplify the computation
    \[ P(E_1 \land E_2) = P(E_1 \mid E_2)P(E_2) = P(E_1)P(E_2) \]
    \[ P(E_1 \lor E_2) = P(E_1) + P(E_2) - P(E_1 \land E_2) \]
    \[ = P(E_1) + P(E_2) - P(E_1)P(E_2) \]
    \[ = 1 - (1 - P(E_1)(1 - P(E_2)) \]

• Mutually exclusive (ME) and exhaustive (EXH) set of events
  – ME: $E_i \land E_j = \emptyset$ ($P(E_i \land E_j) = 0), i, j = 1, \ldots, n, i \neq j$
  – EXH: $E_1 \lor \ldots \lor E_n = S$ ($P(E_1 \lor \ldots \lor E_n) = 1$)
Bayes’ Theorem

• In the setting of diagnostic/evidential reasoning

\[ \begin{array}{c}
H_i \quad P(H_i) \\
P(E_j | H_i) \\
E_1 \quad E_j \quad E_m
\end{array} \]

- Know prior probability of hypothesis \( P(H_i) \)
- Conditional probability \( P(E_j | H_i) \)
- Want to compute the *posterior probability* \( P(H_i | E_j) \)

• Bayes’ theorem (formula 1): \( P(H_i | E_j) = P(H_i) P(E_j | H_i) / P(E_j) \)

• If the purpose is to find which of the \( n \) hypotheses \( H_1, \ldots, H_n \) is more plausible given \( E_j \), then we can ignore the denominator and rank them use *relative likelihood*

\[ rel(H_i | E_j) = P(E_j | H_i) P(H_i) \]
• \( P(E_j) \) can be computed from \( P(E_j \mid H_i) \) and \( P(H_i) \), if we assume all hypotheses \( H_1, \ldots, H_n \) are ME and EXH

\[
P(E_j) = P(E_j \land (H_1 \lor \ldots \lor H_n)) \quad \text{(by EXH)}
\]

\[
= \sum_{i=1}^{n} P(E_j \land H_i) \quad \text{(by ME)}
\]

\[
= \sum_{i=1}^{n} P(E_j \mid H_i)P(H_i)
\]

• Then we have another version of Bayes’ theorem:

\[
P(H_i \mid E_j) = \frac{P(E_j \mid H_i)P(H_i)}{\sum_{k=1}^{n} P(E_j \mid H_k)P(H_k)} = \frac{\text{rel}(H_i \mid E_j)}{\sum_{k=1}^{n} \text{rel}(H_k \mid E_j)}
\]

where \( \sum_{k=1}^{n} P(E_j \mid H_k)P(H_k) \), the sum of relative likelihood of all \( n \) hypotheses, is a normalization factor
Probabilistic Inference for simple diagnostic problems

• Knowledge base:
  \[ E_1,\ldots,E_m : \text{evidence/manifestation} \]
  \[ H_1,\ldots,H_n : \text{hypotheses/disorders} \]

  \( E_j \) and \( H_i \) are binary and hypotheses form a ME & EXH set

  \[ P(E_j \mid H_i), i = 1,\ldots,n, j = 1,\ldots,m \]

  conditional probabilities

• Case input: \( E_1,\ldots,E_l \)

• Find the hypothesis \( H_i \) with the highest posterior probability \( P(H_i \mid E_1,\ldots,E_l) \)

• By Bayes’ theorem

  \[ P(H_i \mid E_1,\ldots,E_l) = \frac{P(E_1,\ldots,E_l \mid H_i)P(H_i)}{P(E_1,\ldots,E_l)} \]

• Assume all pieces of evidence are conditionally independent, given any hypothesis

  \[ P(E_1,\ldots,E_l \mid H_i) = \prod_{j=1}^{l} P(E_j \mid H_i) \]
• The relative likelihood

\[ rel(H_i \mid E_1, \ldots, E_l) = P(E_1, \ldots, E_l \mid H_i) P(H_i) = P(H_i) \prod_{j=1}^{l} P(E_j \mid H_i) \]

• The absolute posterior probability

\[ P(H_i \mid E_1, \ldots, E_l) = \frac{\text{rel}(H_i \mid E_1, \ldots, E_l)}{\sum_{k=1}^{n} \text{rel}(H_k \mid E_1, \ldots, E_l)} = \frac{P(H_i) \prod_{j=1}^{l} P(E_j \mid H_i)}{\sum_{k=1}^{n} P(H_k) \prod_{j=1}^{l} P(E_j \mid H_k)} \]

• Evidence accumulation (when new evidence discovered)

\[ rel(H_i \mid E_1, \ldots, E_l, E_{l+1}) = P(E_{l+1} \mid H_i) \text{rel}(H_i \mid E_1, \ldots, E_l) \]

\[ rel(H_i \mid E_1, \ldots, E_l, \sim E_{l+1}) = (1 - P(E_{l+1} \mid H_i)) \text{rel}(H_i \mid E_1, \ldots, E_l) \]
Assessment of Assumptions

• Assumption 1: hypotheses are mutually exclusive and exhaustive
  – Single fault assumption (one and only hypothesis must true)
  – Multi-faults do exist in individual cases
  – Can be viewed as an approximation of situations where hypotheses are independent of each other and their prior probabilities are very small

\[ P(H_1 \land H_2) = P(H_1)P(H_2) \approx 0 \text{ if both } P(H_1) \text{ and } P(H_2) \text{ are very small} \]

• Assumption 2: pieces of evidence are conditionally independent of each other, given any hypothesis
  – Manifestations themselves are not independent of each other, they are correlated by their common causes
  – Reasonable under single fault assumption
  – Not so when multi-faults are to be considered
Limitations of the simple Bayesian system

• Cannot handle well hypotheses of multiple disorders
  – Suppose $H_1, \ldots, H_n$ are independent of each other
  – Consider a composite hypothesis $H_1 \land H_2$
  – How to compute the posterior probability (or relative likelihood)

$$P(H_1 \land H_2 \mid E_1, \ldots, E_l)?$$
  – Using Bayes’ theorem

$$P(H_1 \land H_2 \mid E_1, \ldots, E_l) = \frac{P(E_1, \ldots, E_l \mid H_1 \land H_2)P(H_1 \land H_2)}{P(E_1, \ldots, E_l)}$$

$$P(H_1 \land H_2) = P(H_1)P(H_2)$$ because they are independent

$$P(E_1, \ldots, E_l \mid H_1 \land H_2) = \prod_{j=1}^{l} P(E_j \mid H_1 \land H_2)$$
  assuming $E_j$ are independent, given $H_1 \land H_2$

How to compute $P(E_j \mid H_1 \land H_2)$?
– Assuming $H_1, \ldots, H_n$ are independent, given $E_1, \ldots, E_I$?

$$P(H_1 \cap H_2 \mid E_1, \ldots, E_I) = P(H_1 \mid E_1, \ldots, E_I) \times P(H_2 \mid E_1, \ldots, E_I)$$

but this is a very unreasonable assumption

E and B are independent  
But when A is given, they are (adversely) dependent 
because they become competitors to explain A  
P(B\mid A, E) \ll P(B\mid A)

\begin{itemize}
  \item Cannot handle causal chaining  
  \begin{itemize}
    \item Ex. A: weather of the year  
      B: cotton production of the year  
      C: cotton price of next year  
    \item Observed: A influences C  
    \item The influence is not direct (A -> B -> C)  
  \end{itemize}
  \begin{itemize}
    \item P(C\mid B, A) = P(C\mid B): instantiation of B blocks influence of A on C  
  \end{itemize}
  \item Need a better representation and a better assumption
\end{itemize}
Bayesian Belief Networks (BBN)

• Definition: A BBN = (DAG, CPD)
  – **DAG**: directed acyclic graph
    nodes: random variables of interest (binary or multi-valued)
    arcs: direct causal/influential relations between nodes
  – **CPD**: conditional probability distribution at each node $x_i$
    \[
    P(x_i | \pi_i) \quad \text{where } \pi_i \text{ is the set of all parent nodes of } x_i
    \]
  – For root nodes $\pi_i = \emptyset$, so $P(x_i | \pi_i) = P(x_i)$
    Since roots are not influenced by anyone, they are considered independent of each other

• Example BBN

\[
\begin{align*}
P(A) &= 0.001 \\
P(B|A) &= 0.3 \\
P(C|A) &= 0.2 \\
P(D|B,C) &= 0.1 \\
P(D|B,\neg C) &= 0.01 \\
P(D|\neg B,C) &= 0.01 \\
P(D|\neg B,\neg C) &= 0.00001 \\
P(E|C) &= 0.4 \\
P(E|\neg C) &= 0.002
\end{align*}
\]
• Independence assumption
  – \( P(x_i | \pi_i, q) = P(x_i | \pi_i) \)
    where \( q \) is any set of variables (nodes) other than \( x_i \) and its successors
  – \( \pi_i \) blocks influence of other nodes on \( x_i \) and its successors (\( q \) influences \( x_i \) only through variables in \( \pi_i \))
  – With this assumption, the complete joint probability distribution of all variables in the network can be represented by (recovered from) local CPD by chaining these CPD

\[
P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \pi_i)
\]

\[
P(A, B, C, D, E) = P(E|A, B, C, D) P(A, B, C, D) \quad \text{by Bayes’ theorem}
\]
\[
= P(E|C) P(A, B, C, D) \quad \text{by indep. assumption}
\]
\[
= P(E|C) P(D|A, B, C) P(A, B, C)
\]
\[
= P(E|C) P(D|B, C) P(C|A, B) P(A, B)
\]
\[
= P(E|C) P(D|B, C) P(C|A) P(B|A) P(A)
\]
Inference with BBN

• Belief update
  
  − Original belief (no variable is instantiated): the prior probability $P(x_i)$
    If $x_i$ is a root, then $P(x_i)$ is given in BBN.
    Otherwise, $P(x_i) = \sum_{\pi_i} P(x_i | \pi_i)P(\pi_i)$

    $P(x_i | \pi_i)$ is given, but computer $P(\pi_i)$ is complicated
    Ex: $P(B, C) = P(A, B, C) + P(\sim A, B, C)$
      $= P(B | A, C)P(A, C) + P(B | \sim A, C)P(\sim A, C)$
      $= P(B | A)P(C | A)P(A) + P(B | \sim A)P(C | \sim A)P(\sim A)$

  − When some variables are instantiated (say $x_j$ has value $X_j$),
    beliefs on all other variable $x_i$ is changed to $P(x_i | X_j)$
    $P(x_i | X_j)$ can be computed from the joint probability distribution
    Ex: $d = D$ and $e = E$

    $$P(A | D, E) = \frac{P(A, D, E)}{P(D, E)} = \frac{\sum_{b,c} P(A, b, c, D, E)}{\sum_{a,b,c} P(a, b, c, D, E)}$$

    This approach is not computationally feasible with large network
– **Algorithmic approach** (Pearl and others)
  - Singly connected network, SCN (also known as poly tree)
    - there is at most one undirected path between any two nodes
      (i.e., the network is a tree if the direction of arcs are ignored)
  - The influence of the instantiated variable spreads to the rest of the network along the arcs
    - The instantiated variable influences its predecessors and successors differently
    - Computation is linear to the diameter of the network (the longest undirected path)
  - For non-SCN (network with general structure)
    - Conditioning: find the network’s smallest cutset \( C \) (a set of nodes whose removal will render the network singly connected) for each instantiation of \( C \), compute the belief update with the SCN algorithm
    - Combine the results from all possible instantiation of \( C \).
    - Computationally expensive (finding the smallest cutset is itself NP-hard, and total number of possible instantiations of \( C \) is \( O(2^{|C|}) \).)
– *Stochastic simulation*

- Randomly generate large number of instantiations of ALL variables $I_k^{(n)}$ according to CPD
- Only keep those instantiations $I_k^{(n)}$ which are consistent with the values of given instantiated variables
- Updated belief of those un-instantiated variables as their frequencies in the pool of recorded $I_k^{(n)}$
- The accuracy of the results depend on the size of the pool (asymptotically approaches the exact results)
• MAP problems
  
  – Let $X$ denote the set of all variables in a BBN, $V \subseteq X$ the set of instantiated variables, $U = X - V$ the set of all un-instantiated variables. Then the MAP (maximum a posteriori probability) problem is to find the most probable instantiation of $U$, given $V$, i.e.,
  
  $$\max_u (P(U | V))$$

  – This is an optimization problem
  
  – Algorithms developed for exact solutions for different special BBN (Peng, Cooper, Pearl) have exponential complexity
  
  – Other techniques for approximate solutions
    • Genetic algorithms
    • Neural networks
    • Simulated annealing
    • Mean field theory
Noisy-Or BBN

• A special BBN of binary variables (Peng & Reggia, Cooper)
  – Each link $x_i \rightarrow x_j$ is associated with a probability value called 
    **causal strength** $c_{ij}$ that measures the strength of $x_i$ alone may 
    cause $x_j$, i.e., $c_{ij} = P(x_i \mid x_j$ is true and all others in $\pi_i$ are false)
  – Causation independence: parent nodes influence a child 
    independently

• Advantages:
  – One-to-one correspondence between causal links and causal 
    strengths
  – Easy for humans to understand (acquire and evaluate KB)
  – Fewer # of probabilities needed in KB 
    Complete joint prob. distribution: $2^n$
    General BBN: $\sum_{i=1}^{n} 2^{\mid \pi_i \mid}$
    Noisy-Or BBN: $\sum_{i=1}^{n} \mid \pi_i \mid$
  – Computation is less expensive

• Disadvantage: less expressive (less general)
Learning BBN (from case data)

• Need for learning
  – Experts’ opinions are often biased, inaccurate, and incomplete
  – Large databases of cases become available

• What to learn
  – Learning CPD when DAG is known (easy)
  – Learning DAG (hard)

• Difficulties in learning DAG from case data
  – There are too many possible DAG when # of variables is large (more than exponential)
    n = 3, # of possible DAG = 25
    n = 10, # of possible DAG = 4*10^18
  – Missing values in database
  – Noisy data
• Approaches
  – *Early effort*: Based on variable dependencies (Pearl)
    • Find all pairs of variables that are dependent of each other (applying standard statistical method on the database)
    • Eliminate (as much as possible) indirect dependencies
    • Determine directions of dependencies
    • Learning results are often incomplete (learned BBN contains indirect dependencies and undirected links)
  – *Bayesian approach* (Cooper)
    • Find the most probable DAG, given database DB, i.e., 
      \[
      \max(P(DAG|DB)) \text{ or } \max(P(DAG, DB))
      \]
    • Based on some assumptions, a formula is developed to compute \(P(DAG, DB)\) for a given pair of DAG and DB
    • A hill-climbing algorithm (K2) is developed to search a (sub)optimal DAG
    • Extensions to handle some form of missing values
– *Minimum description length* (MDL) (Lam)
  
  - Sacrifices accuracy for simpler (less dense) structure
    - Case data not always accurate
    - Fewer links imply smaller CPD tables and less expensive inference
  
  - $L = L_1 + L_2$ where
    - $L_1$: the length of the encoding of DAG (smaller for simpler DAG)
    - $L_2$: the length of the encoding of the difference between DAG and DB
      (smaller for better match of DAG with DB)
    - Smaller $L_2$ implies more accurate (and more complex) DAG, and thus larger $L_1$
  
  - Find DAG by heuristic best-first search, that Minimizes $L$

– *Neural network approach* (Neal, Peng)
  
  - For noisy-or BBN
  
  - Maximizing $L = \ln \prod_{V^r \in D} P(\tilde{V} = V^r)$ where
    - $D$: case database; $V^r$: case in $D$; $\tilde{V}$: state vector of the learned network
    - $L$ measures the similarity of the two distributions: one in $D$, another in the learned network
Dempster-Shafer theory

- A variation of Bayes’ theorem to represent ignorance
- Uncertainty and ignorance
  - Suppose two events $A$ and $B$ are ME and EXH, given an evidence $E$
    - $A$: having cancer  $B$: not having cancer  $E$: smoking
  - By Bayes’ theorem: our beliefs on $A$ and $B$, given $E$, are measured by $P(A|E)$ and $P(B|E)$, and $P(A|E) + P(B|E) = 1$
  - In reality,
    - I may have some belief in $A$, given $E$
    - I may have some belief in $B$, given $E$
    - I may have some belief not committed to either one,
  - The uncommitted belief (ignorance) should not be given to either $A$ or $B$, even though I know one of the two must be true, but rather it should be given to “$A$ or $B$”, denoted \{A, B\}
  - Uncommitted belief may be given to $A$ and $B$ when new evidence is discovered
• Representing ignorance
  – Frame of discernment: \( \theta = \{ h_1, \ldots, h_n \} \), a set of ME and EXH hypotheses. The power set \( 2^\theta \) is organized as a lattice of super/subset relations. Each node \( S \) is a subset of hypotheses (\( S \subseteq \theta \))
  – Ex: \( \theta = \{ A, B, C \} \)

Each node \( S \) is associated with a basic probability assignment \( m(S) \)
\[
0 \leq m(S) \leq 1;
m(\emptyset) = 0;
\sum_{S \subseteq \theta} m(S) = 1
\]

• Belief function
\[
Bel(S) = \sum_{S' \subseteq S} m(S'); \quad Bel(\emptyset) = 0; \quad Bel(\theta) = 1
\]
\[
Bel(\{A, B\}) = m(\{A, B\}) + m(\{A\}) + m(\{B\}) + m(\emptyset)
= 0.1 + 0.1 + 0.2 + 0 = 0.4
\]
\[
Bel(\{A, B\}^C) = Bel(\{C\}) = 0.3
\]
– Plausibility (upper bound of belief of a node)

All belief not committed to $S^C$ may be committed to $S$

\[
Pls(S) = 1 - Bel(S^C)
\]

\[
Pls(\{A, B\}) = 1 - Bel(\{C\}) = 1 - 0.3 = 0.7
\]

[\[Bel(S), \quad Pls(S)\]] belief interval

\[
\begin{align*}
\text{Lower bound} & \quad \text{Upper bound} \\
\text{(known belief)} & \quad \text{(maximally possible)}
\end{align*}
\]
• **Evidence combination** (how to use D-S theory)

– Each piece of evidence has its own $m(.)$ function for the same $\theta$

\[ \theta = \{A, B\} : A : \text{having cancer}; B : \text{not having cancer} \]

\[
\begin{align*}
\{A, B\} & : 0.3 \\
\{A\} & : 0.2 \\
\{B\} & : 0.5 \\
\{\emptyset\} & : 0
\end{align*}
\]

\[
\begin{align*}
\{A, B\} & : 0.1 \\
\{A\} & : 0.7 \\
\{B\} & : 0.2 \\
\{\emptyset\} & : 0
\end{align*}
\]

\[
m_1(S) \quad m_2(S)
\]

$E_1 : \text{smoking} \quad E_2 : \text{living in high radiation area}$

– Belief based on combined evidence can be computed from

\[
m(S) = m_1(S) \oplus m_2(S) = \frac{\sum_{X \cap Y = S} m_1(X)m_2(Y)}{1 - \sum_{X \cap Y = \emptyset} m_1(X)m_2(Y)}
\]

normalization factor

incompatible combination
\[
m(\{A\}) = \frac{m_1(\{A\})m_2(\{A\}) + m_1(\{A\})m_2(\{A,B\}) + m_1(\{A,B\})m_2(\{A\})}{1-[m_1(\{A\})m_2(\{B\}) + m_1(\{B\})m_2(\{A\})]} \\
= \frac{0.2 \cdot 0.7 + 0.2 \cdot 0.1 + 0.3 \cdot 0.7}{1-\left[0.2 \cdot 0.2 + 0.5 \cdot 0.7\right]} = \frac{0.37}{0.61} = 0.607
\]

\[
m(\{B\}) = \frac{m_1(\{B\})m_2(\{B\}) + m_1(\{B\})m_2(\{A,B\}) + m_1(\{A,B\})m_2(\{B\})}{1-[m_1(\{A\})m_2(\{B\}) + m_1(\{B\})m_2(\{A\})]} \\
= \frac{0.5 \cdot 0.2 + 0.5 \cdot 0.1 + 0.3 \cdot 0.2}{1-\left[0.2 \cdot 0.2 + 0.5 \cdot 0.7\right]} = \frac{0.21}{0.61} = 0.344
\]

\[
m(\{A, B\}) = \frac{m_1(\{A, B\})m_2(\{A, B\})}{0.61} = \frac{0.03}{0.61} = 0.049
\]
– Ignorance is reduced
  from $m_1(\{A,B\}) = 0.3$ to $m(\{A,B\}) = 0.049$
– Belief interval is narrowed
  A: from $[0.2, 0.5]$ to $[0.607, 0.656]$
  B: from $[0.5, 0.8]$ to $[0.344, 0.393]$

• Advantage:
  – The only formal theory about ignorance
  – Disciplined way to handle evidence combination

• Disadvantages
  – Computationally very expensive (lattice size $2^{|\Theta|}$)
  – Assuming hypotheses are ME and EXH
  – How to obtain $m(.)$ for each piece of evidence is not clear, except subjectively
Fuzzy sets and fuzzy logic

- Ordinary set theory
  
  \[ f_A(x) = \begin{cases} 
  1 & \text{if } x \in A \\
  0 & \text{otherwise} 
  \end{cases} \]

  \( f_A(x) \) is called the characteristic or membership function of set \( A \)

  Predicate \( A(x) = \begin{cases} 
  1 & \text{if } x \in A \\
  0 & \text{otherwise} 
  \end{cases} \)

  When it is uncertain if \( x \in A \), use probability \( P(x \in A) \)

- There are sets that are described by vague linguistic terms (sets without hard, clearly defined boundaries), e.g., tall-person, fast-car
  - Continuous
  - Subjective (context dependent)
  - Hard to define a clear-cut 0/1 membership function
• Fuzzy set theory
  – Relax $f_A(x)$ from binary $\{0, 1\}$ to continuous $[0, 1]$ stands for the degree $x$ is thought to belong to set $A$

  $\begin{align*}
  \text{height(john)} &= 6'5'' \quad \text{Tall(john)} = 0.9 \\
  \text{height(harry)} &= 5'8'' \quad \text{Tall(harry)} = 0.5 \\
  \text{height(joe)} &= 5'1'' \quad \text{Tall(joe)} = 0.1
  \end{align*}$

  – Examples of membership functions

\begin{itemize}
  \item Set of teenagers
  \item Set of young people
  \item Set of mid-age people
\end{itemize}
• Fuzzy logic: many-value logic
  – Fuzzy predicates (degree of truth)  \( F_A(x) = y \) if \( f_A(x) = y \)
  – Connectors/Operators
    negation : \( \neg F_A(x) = 1 - F_A(x) \)
    conjunction : \( F_A(x) \land F_B(x) = \min\{F_A(x), F_B(x)\} \)
    disjunction : \( F_A(x) \lor F_B(x) = \max\{F_A(x), F_B(x)\} \)

• Compare with probability theory
  – Prob. Uncertainty of outcome,
    • Based on large # of repetitions or instances
    • For each experiment (instance), the outcome is either true or false
      (without uncertainty or ambiguity)
      unsure before it happens but sure after it happens

Fuzzy: vagueness of conceptual/linguistic characteristics
• Unsure even after it happens
  whether a child of tall mother and short father is tall
unsure before the child is born
unsure after grown up (height = 5’6”)
– Empirical vs subjective (testable vs agreeable)
– Fuzzy set connectors may lead to unreasonable results
  • Consider two events A and B with \( P(A) < P(B) \)
  • If \( A \Rightarrow B \) (or \( A \subseteq B \)) then
    \[
    P(A \cap B) = P(A) = \min\{P(A), P(B)\}
    \]
    \[
    P(A \cup B) = P(B) = \max\{P(A), P(B)\}
    \]
  • Not the case in general
    \[
    P(A \cap B) = P(A)P(B|A) \leq P(A)
    \]
    \[
    P(A \cup B) = P(A) + P(B) - P(A \cap B) \geq P(B)
    \]
    (equality holds only if \( P(B|A) = 1 \), i.e., \( A \Rightarrow B \))
  
– Something prob. theory cannot represent
  • Tall(john) = 0.9, \( \neg\text{Tall}(\text{john}) = 0.1 \)
    Tall(john) \( \cap \neg\text{Tall}(\text{john}) = \min\{0.1, 0.9\} = 0.1 \)
    john’s degree of membership in the fuzzy set of “median-height people” (both Tall and not-Tall)
  • In prob. theory: \( P(\text{john} \in \text{Tall} \cap \text{john} \notin \text{Tall}) = 0 \)
Uncertainty in rule-based systems

- Elements in Working Memory (WM) may be uncertain because
  - Case input (initial elements in WM) may be uncertain
    Ex: the CD-Drive does not work 70% of the time
  - Decision from a rule application may be uncertain even if the
    rule’s conditions are met by WM with certainty
    Ex: flu => sore throat with high probability

- Combining symbolic rules with numeric uncertainty: Mycin’s
  Uncertainty Factor (CF)
  - An early attempt to incorporate uncertainty into KB systems
  - $CF \in [-1, 1]$
  - Each element in WM is associated with a CF: certainty of that
    assertion
  - Each rule $C_1, \ldots, C_n \Rightarrow Conclusion$ is associated with a CF:
    certainty of the association (between $C_1, \ldots, C_n$ and $Conclusion$).
– CF propagation:
  • Within a rule: each $C_i$ has $CF_i$, then the certainty of Action is $\min\{CF_1, \ldots, CF_n\} \times CF$-of-the-rule
  • When more than one rules can apply to the current WM for the same Conclusion with different CFs, the largest of these CFs will be assigned as the CF for Conclusion
  • Similar to fuzzy rule for conjunctions and disjunctions

– Good things of Mycin’s CF method
  • Easy to use
  • CF operations are reasonable in many applications
  • Probably the only method for uncertainty used in real-world rule-base systems

– Limitations
  • It is in essence an ad hoc method (it can be viewed as a probabilistic inference system with some strong, sometimes unreasonable assumptions)
  • May produce counter-intuitive results.