Uncertainty

Uncertainty

- Our knowledge is incomplete and changing to include new facts
- How can we solve problems in the face of this incompleteness?
- Logic fails in such cases - but humans are capable of solving problems in these situations
- We need methods that can handle uncertainty

Uncertainty

- Ambiguity
  - Things can be interpreted in more than one way
- Incompleteness
  - Some information missing

Sources of Uncertainty

- Incorrectness
  - Human error
  - Equipment error
  - False negative - rejecting a true hypothesis
  - False positive - accepting a false hypothesis
- Measurement
  - Precision, how precise e.g. cm or mm
  - Accuracy - calibration

Sources of Uncertainty

- Random
  - Due to noise, loose wires etc.
- Systematic
  - Due to error in reading
    - e.g. user giving inches not cm
- Reasoning
  - Wrongly formulated or learnt rules

Example of reasoning with uncertainty

- Abbott, Babbitt and Cabot are suspects in a murder case
- *Abbott's alibi:* registration in a respectable hotel in Albany
- *Babbitt's alibi:* was visiting his brother-in-law in Brooklyn at the time of the murder
- *Cabot's alibi:* watching a ski meet in the Catskills (but no witnesses)

Example of reasoning with uncertainty (Cont'd)

- We have the following beliefs:
  - Abbott, Babbitt, or Cabot committed the murder
  - Abbott did not commit the murder (alibi)
  - Babbitt did not commit the murder (alibi)
  - Cabot may have committed the murder (his alibi is weak)
- We must solve this problem based on "beliefs" using "belief states"
  - I believe that Cabot is the murderer because he has the weakest alibi

Example of reasoning with uncertainty (Cont'd)

- Now consider that Cabot is able to document his alibi (someone took a photograph of him watching the ski meet)
  - Cabot's alibi is now stronger
- Babbitt's brother-in-law might be lying
  - Babbitt's alibi is now weaker
- Abbott's registry could be forged
  - Abbott's alibi is now weaker
- How do we deal with such uncertainties? What conclusion(s) should we draw?

Symbolic Approaches

Monotonic Reasoning

- Conventional reasoning mechanisms, such as logic, can make inferences based on information that has the following properties:
  - Inferences are made based on that the knowledge used is 'complete', i.e. all facts and rules required for these inferences have been included in the knowledge base.
  - Facts that are necessary to solve the problem are present or can be derived from the knowledge provided in the knowledge base.
  - Facts and knowledge are consistent, i.e. new facts or knowledge does not invalidate old ones.
  - Facts are either known to be true or false
  - Inferences can be established as either true or false.
- Such a reasoning mechanism is called *monotonic reasoning*
- If any of the above properties is not satisfied, conventional logic-based reasoning mechanisms become inadequate.

Nonmonotonic Reasoning

- Nonmonotonic reasoning addresses the following issues:
  - extending the knowledge base so that inferences can be made on the basis of lack of information as well as the presence of it. It makes distinction between: *it is known that a fact is not true* and *the fact is not known*.
  - updating information in the knowledge base when a new fact is added. New inferences will be made based on the new facts.
  - resolving conflicts when several inconsistent inferences are made. This is solved either by assigned preferences or allowing for multi-solution inferences.

Nonmonotonic Reasoning Mechanisms

- There are a number of reasoning mechanisms which uses nonmonotonic reasoning, such as
  - *nonmonotonic logic,*
  - *closed world assumption,*
  - *Truth Maintenance Systems*

Non-monotonic Logic

- A system where statements can be T, F or neither (knowledge does not have to be complete)
- There is no formalism like Resolution, so we must develop more “ad hoc” methods.
- Add an operator M to first order predicate calculus
- M means “True if consistent”

Non-monotonic Logic

- Now we are able to prove things T and F with statements that include M because it means that we assume any future information will be consistent. If not, our proof does not hold
- Example:
  - $\forall$X, Y: related (X,Y) $\wedge$M Get Along (X,Y) $\Rightarrow$Will Defend (X, Y)
  - That is, we assume that if X and Y are related then X will defend Y. However, if we learn that X does not get along with Y then this becomes false.

Default Logic

- A logic where assumptions are made when rules are introduced

$\frac{A : B}{C}$ ---> “If A is provable and it is consistent to assume B then conclude C”

- This is different from Non-monotonic logic by requiring the assumption to be made at the time a rule is introduced (i.e. you cannot add new rules of this form)

Abduction

- Standard deduction (*Modus Ponens*) states that if you know A is true and you have the rule $A \Rightarrow B$, then you can conclude B is true. This is truth preserving
- In Abduction, we turn this around to read if B is true and you have a rule A $\Rightarrow$ B then you can conclude A is true. This is not truth preserving.

Abduction

- Abduction is more often used when we have several rules A1 $\Rightarrow$ B, A2 $\Rightarrow$ B, A3 $\Rightarrow$ B, A4 $\Rightarrow$ B, etc...
- If B is true, what caused it (what was responsible for B being true)? Our answer is A1 or A2 or A3 or A4, but which one?
- We can use our beliefs to pick the right one.
- Abduction is often used for explanations:
  - diagnosis/credit assignment
  - theory understanding/legal reasoning
  - recognition/perception/natural language understanding

Reconsider the murder case

- We assume Cabot's alibi is good because of the photograph.
- We assume that Abbott really went to the hotel because of the hotel's reputation.
- We believe that Babbitt's brother might lie (he has been known to lie before)
- Therefore, we conclude that Babbitt is the murderer

Minimalist Reasoning

- In our belief model, we only know what we can currently prove to be true (we do not include things that may later be proven false -- i.e. no assumptions).
- **Closed World Assumption (CWA)** - if we do not know something is true, then it is false.
- **Circumscription** - Limiting conclusions (or interpretations) that should be derived given new axioms

A problem with CWA

- CWA does not take into account interactions between pieces of knowledge
  - consider our murder case, we know that murderer(A) or murderer(B) or murderer(C) and we have alibi(A), alibi(B), alibi(C), or NOT murderer(A), NOT murderer(B), NOT murderer(C), thus we have an inconsistency

Implementations

- Augmenting a problem solver with “UNLESS”
- Dependency-Directed Backtracking
- Justification-based Truth Maintenance System
- Assumption-based Truth Maintenance System

Using UNLESS

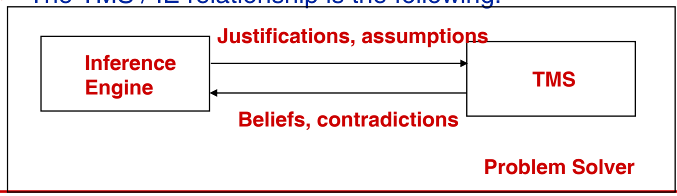
- Augment rules with the *Unless* clause which introduces exceptions to a rule
- Use *forward chaining* if the problem is data-driven
- Use *backward chaining* if the problem is goal-driven - this approach will not be able to determine if/when new information becomes available and thus is not as useful

Dependency-Directed Backtracking

- Consider the following example:
  - We want to prove F
  - We have a nonmonotonic rule If $A \Rightarrow F$
  - We have no reason not to assume A, so we assume A is true
  - $F \Rightarrow G \wedge H$
  - By some other facts, we can also derive $M \wedge N$
  - We learn A is not true. Backtracking would remove F, G, H, $M \wedge N$. F, $G \wedge H$ may not be true, but $M \wedge N$ were derived independently!

Truth Maintenance Systems

- A Truth Maintenance System (TMS) is responsible for:
- Enforcing logical relations among beliefs.
  - Generating explanations for conclusions.
  - Finding solutions to search problems
  - Supporting default reasoning.
  - Identifying causes for failure and recover from inconsistencies.
  - The TMS / IE relationship is the following:

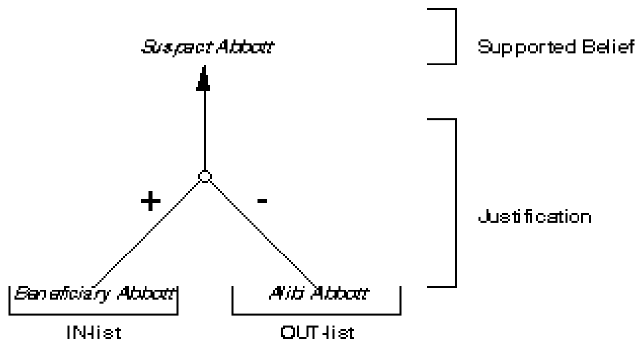


Families of TMSs

- There are several families of TMSs, which differ in the representation scheme they use and the functionality they support:
  - **Justification-based TMSs.** The language used is limited to Horn formulas.
  - **Logic-based TMSs.** These use a full propositional logic language.
  - **Assumption-based TMSs.** Language limited to Horn formulas, but several alternatives (contexts) can be explored at the same time.
  - **Non-monotonic JTMSs.** Language limited to Horn formulas, but allow non-monotonic justifications, thus making it possible to implement default reasoning.
  - **Clause Management Systems.** Their representational power is equivalent to LTMSs, but like ATMSs can support several contexts at the same time.
  - **Contradiction-tolerant TMSs.** Language limited to Horn formulas, but support non-monotonic and plausible reasoning and deal explicitly with contradictions in a single context.

Justification-Based TMS

- To justify (believe) an assertion, all items in the "IN-list" must be believed while all items in the "OUT-list" must be disbelieved.



A breadth-first approach

- JTMS - depth-first search of possible beliefs with dependency-directed backtracking
- Here, backtracking is avoided by considering multiple belief states at one time
- As the search space is expanded, any states that are inconsistent are pruned away reducing some of the search

Numerical Approaches

Probabilistic Reasoning

Statistical Reasoning

- Nonmonotonic reasoning can be used to model belief systems in which facts are believed to be true, false or unknown.
- However, for some problems, it is useful to be able to describe beliefs that are not certain but for which there is some supporting evidence.

Classes of Problems

- There are two broad classes of problems.
  - The first class contains problems for which there is a genuine randomness of the world.
    - Playing cards or throwing dice are examples of such problems.
    - These problems can be statistically modeled and it is possible to predict the likelihood of various outcomes, though complete certainty is not possible.
  - The second class of problems the relevant world is not random.
    - Many ‘common sense tasks’ fall into this class, such as, in certain domains, fault diagnostic and engineering design.
    - The difficulty is that there are many more possible exceptions than we can enumerate explicitly. For problems like these, statistical measures may serve as a means to ‘summarize’ the world rather than modeling every eventuality or exceptions.

Examples of Probabilistic Knowledge

- Consider that your grass is wet, what caused it?
  - There is a 30% chance of rain
  - We run the sprinkler 3 times per week
  - The water main might break one chance in 1000
- We need ways to represent and reason over this form of knowledge

Why Reason Probabilistically?

- In many problem domains it isn't possible to create complete, consistent models of the world. Therefore agents (and people) must act in uncertain worlds (which the real world is).
- We want an agent to make rational decisions even when there is not enough information to prove that an action will work.
- Some of the reasons for reasoning under uncertainty:
  - True uncertainty. E.g., flipping a coin.
  - Theoretical ignorance. There is no complete theory which is known about the problem domain. E.g., medical diagnosis.
  - Laziness. The space of relevant factors is very large, and would require too much work to list the complete set of antecedents and consequents. Furthermore, it would be too hard to use the enormous rules that resulted.
  - Practical ignorance. Uncertain about a particular individual in the domain because all of the information necessary for that individual has not been collected.
- Probability theory will serve as the formal language for representing and reasoning with uncertain knowledge.

Representing Belief about Propositions

- Rather than reasoning about the truth or falsity of a proposition, reason about the belief that a proposition or event is true or false
- For each primitive proposition or event, attach a **degree of belief** to the sentence
- Use probability theory as a formal means of manipulating degrees of belief
- Given a proposition, A, assign a probability, P(A), such that

  $0 <= P(A) <= 1$, where if A is true, P(A)=1, and if A is false, P(A)=0.
- Proposition A must be either true or false, but P(A) summarizes our degree of belief in A being true/false.

Examples

- P(Weather=Sunny) = 0.7 means that we believe that the weather will be Sunny with 70% certainty.
  - In this case Weather is a random variable that can take on values in a domain such as {Sunny, Rainy, Snowy, Cloudy}.
- P(Cavity=True) = 0.05 means that we believe there is a 5% chance that a person has a cavity.
  - Cavity is a Boolean random variable since it can take on possible values True and False.
- $P(A=a \wedge B=b) = P(A=a, B=b) = 0.2$, where A=My_Mood, a=happy, B=Weather, and b=rainy, means that there is a 20% chance that when it's raining my mood is happy.

Obtaining and Interpreting Probabilities

- There are several senses in which probabilities can be obtained and interpreted, among them the following:
  - **Frequentist Interpretation:** The probability is a property of a population of similar events. e.g., if set $S = P \cup N$, and $P \cap N$ is the empty set, then the probability of an object being in set P is |P|/|S|. Hence, in this interpretation probabilities come from experiments and determining the population associated with a given proposition.
  - **Subjectivist Interpretation:** A subjective degree of belief in a proposition or the occurrence of an event. E.g., the probability that you'll pass final exam based on your own subjective evaluation of the amount of studying you've done and your understanding of the material. Hence, in this interpretation probabilities characterize the agent's beliefs.

Concepts

- We will assume that in a given problem domain, the programmer and expert identify all of the relevant propositional variables that are needed to reason about the domain.
- Each of these will be represented as a **random variable**, i.e., a variable that can take on values from a set of mutually exclusive and exhaustive values called the **sample space** or **partition** of the random variable.

Types of Random Variables

- **Boolean random variables**, always have the domain {true; false}.
- **Discrete random variables**, which include Boolean random variables as a special case, take on values from a finite domain.
  - For example, the domain of Weather might be {sunny; rainy; cloudy; snow}.
  - The values in the domain must be mutually exclusive and exhaustive.
- **Continuous random variables** take on values from the real numbers. The domain may be the entire real line or some subset such as the interval [0,1].

Axioms of Probability Theory

1. All probabilities are between 0 and 1. For any proposition *a*,

   $0 \leq P(a) \leq 1$

2. Necessarily true (i.e., valid) propositions have probability 1, and necessarily false (i.e., unsatisfiable) propositions have probability 0.

   P(true) = 1 P(false) = 0

3. The probability of a disjunction is given by

   $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

Prior probability

- The **unconditional** or **prior probability** associated with a proposition A is the degree of belief accorded to it in the absence of any other information, and is written as P(A).
- P(A) can only be used when there is no other information. As soon as some new information is known, we have to reason with the conditional probability of a given that new information.

Joint Probability Distribution

- Given an application domain in which we have determined a sufficient set of random variables to encode all of the relevant information about that domain, we can completely specify all of the possible probabilistic information by constructing the **full joint probability distribution**,

  $(V_1=v_1, V_2=v_2, \ldots, V_n=v_n)$,
- which assigns probabilities to all possible combinations of values to all random variables.

Properties of Joint Probability Table

- With *n* Boolean variables the table will be of size 2^n.
- If *n* variables each had *k* possible values, then the table would be size k^n.
- The sum of the probabilities in the right column must equal *1* since we know that the set of all possible values for each variable are known.
- This means that for *n* Boolean random variables, the table has 2^n-1 values that must be determined to completely fill in the table.
- If all of the probabilities are known for a full joint probability distribution table, then we can compute *any* probabilistic statement about the domain.

Conditional Probabilities

- Conditional probabilities formalize the process of accumulating evidence and updating probabilities based on new evidence.
- If P(A|B) = 1, this is equivalent to the sentence in Propositional Logic B => A.
- Similarly, if P(A|B) =0.9, then this is like saying B => A with 90% certainty.
- Given several measurements and other "evidence", $E_1, \ldots, E_k$, we will formulate queries as
  $P(Q \mid E_1, E_2, \ldots, E_k)$
- meaning "what is the degree of belief that Q is true given that we know $E_1, \ldots, E_k$ *and nothing else*."
- Conditional probability is defined as:
  $P(A|B) = P(A \cap B)/P(B)$

Important Rules Related to Conditional Probability

- Rewriting the definition of conditional probability, we get the Product Rule: $P(A \cap B) = P(A|B)P(B)$
- Chain Rule:
  - P(A,B,C,D) = P(A|B,C,D)P(B|C,D)P(C|D)P(D),
  - which generalizes the product rule for a joint probability of an arbitrary number of variables.
  - Note that ordering the variables results in a different expression, but all have the same resulting value.
- Conditionalized version of the Chain Rule: P(A,B|C) = P(A|B,C)P(B|C)

Bayes Rule

$P(A \wedge B) = P(A \mid B)P(B)$

$P(A \wedge B) = P(B \mid A)P(A)$

Equating the two right-hand sides and dividing by P(B), we get Bayes' Rule:

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

- Bayes's Rule is the basis for probabilistic reasoning because given a prior model of the world in the form of P(A) and a new piece of evidence B, Bayes's Rule says how the new piece of evidence decreases the ignorance about the world by defining P(A|B).

Using Bayes's Rule for Medical Diagnosis

- A doctor knows that the disease meningitis causes the patient to have a stiff neck, say, 50% of the time.
- The doctor also knows some unconditional facts: the prior probability of a patient having meningitis is 1/50,000, and the prior probability of any patient having a stiff neck is 1/20.
- Letting s be the proposition that the patient has a stiff neck and m be the proposition that the patient has meningitis, we have

$P(s \mid m) = 0.5$, $P(m) = 1/50000$, $P(s) = 1/20$

$$P(m \mid s) = \frac{P(s \mid m)P(m)}{P(s)} = \frac{0.5 \text{ X } 1/50000}{1/20} = 0.0002$$

Combining Multiple Evidence using Bayes's Rule

- Generalizing Bayes's Rule for two pieces of evidence, B and C, we get:

  $P(A|B \wedge C) = ((P(A)P(B \wedge C \mid A))/P(B,C) = P(A) * [P(B|A)/P(B)] * [P(C \mid A \wedge B)/P(C|B)]$
- Again, this shows how the conditional probability of A is updated given B and C.
- The problem is that it may be hard in general to obtain or compute P(C | A,B).
- But this difficulty is circumvented if we know evidence B and C are conditionally independent or unconditionally independent.

Independence

- **A is (unconditionally) independent of B** if $P(A|B) = P(A)$. In this case, $P(A \wedge B) = P(A)P(B)$.
- **A is conditionally independent of B given C** if $P(A|B \wedge C) = P(A|C)$ and, symmetrically, $P(B|A \wedge C) = P(B|C)$.
  - What this means is that if we know P(A|C), we also know $P(A|B \wedge C)$, so we don't need to store this case. Furthermore, it also means that
  - $P(A \wedge B|C) = P(A|C)P(B|C)$.

Bayes's Rule with Multiple, Independent Evidence

- Assuming conditional independence of B and C given A, we can simplify Bayes's Rule for two pieces of evidence B and C:

  $P(A \mid B,C) = (P(A)P(B,C \mid A))/P(B,C) = (P(A)P(B|A)P(C|A))/(P(B)P(C|B)) =$

  $P(A) * [P(B|A)/P(B)] * [P(C|A)/P(C|B)] =$

  $(P(A) * P(B|A) * P(C|A))/P(B,C)$
- Furthermore, if B and C are (unconditionally) independent, then $P(C|B) = P(C)$, so

$P(A \mid B,C) = P(A) * [P(B|A)/P(B)] * [P(C|A)/P(C)]$

Naive Bayes Classifier

- Say we have a random variable, C, which represents the possible ways to classify an input pattern of features that have been measured.
- The domain of C is the set of possible classifications, e.g., it might be the possible diagnoses in a medical domain.
- Say the possible values for C are {a, b, c}, and the features we have measured are $E_1, E_2, ..., E_n$.
- Then we can compute $P(C=a \mid E_1, ..., E_n)$, $P(C=b \mid E_1, ..., E_n)$ and $P(C=c \mid E_1, ..., E_n)$ assuming $E_1, ..., E_n$ are independent using the above formula, and choose the value for C that gives the maximum probability.

Example

- Consider the medical domain consisting of three Boolean variables: PickledLiver, Jaundice, Bloodshot, where the first indicates if a given patient has the "disease" PickledLiver, and the second and third describe symptoms of the patient.
- We'll assume that Jaundice and Bloodshot are independent.
- The doctor wants to determine the likelihood that the patient has a PickledLiver.
- Based on no other information, she knows that the prior probability P(PickledLiver) = 10^{-17}. So, this represents the doctor's initial belief in this diagnosis.

Example (Cont'd)

- However, after examination, she determines that the patient has jaundice.
- She knows that P(Jaundice) = 2^{-10} and P(Jaundice | PickledLiver) = 2^{-3}, so she computes the new updated probability in the patient having PickledLiver as:
- P(PickledLiver | Jaundice) = $P(P)P(J|P)/P(J)$

  $= (2^{-17} * 2^{-3})/2^{-10}$

  $= 2^{-10}$
- So, based on this new evidence, the doctor increases her belief in this diagnosis from 2^{-17} to 2^{-10}.

Example (Cont'd)

- Next, she determines that the patient's eyes are bloodshot, so now we need to add this new piece of evidence and update the probability of PickledLiver given Jaundice and Bloodshot.
- Say,
  - P(Bloodshot) = 2^{-6} and
  - P(Bloodshot | PickledLiver) = 2^{-1}.
- Then, she computes the new conditional probability:

P(PickledLiver | Jaundice, Bloodshot)

$= (P(P)P(J|P)P(B|P))/(P(J)P(B))$

$= 2^{-10} * [2^{-1} / 2^{-6}] = 2^{-5}$

- So, after taking both symptoms into account, the doctor's belief that the patient has a PickledLiver is 2^{-5}.

Certainty Factors

Mycin

- Mycin is the oldest and best known expert system
- Its function is to diagnose certain bacterial infections and recommend a drug treatment

Mycin

- Rules are of the form:

IF: 1) The strain of the organism is gramneg

AND

2) The morphology of the rod is coccus

THEN:

There is strongly suggestive evidence (0.8) that the genus of the organism is neisseria

Handling uncertainty: Certainty Factors

- Everything in Mycin has a certainty factor associated with it
- CFs range from –1..1
  - -1 means definitely not
  - +1 means definitely
  - 0 means equally likely and unlikely

Certainty Factors

- User can input CFs in the range –10..10, which are converted to the range –1..1
- These CFs are propagated through the rules
- CFs in the range –0.2..0.2 are not propagated as their values are too small

Propagating CFs

- For a rule of the form:
  - If A and B and C then D
- The CF associated with D is found by taking the minimum CF of A, B and C and then multiplying it by the CF associated with the rule

Propagating CFs

- For a rule of the form:
  - If A or B or C then D
- The maximum CF associated with A, B or C is taken and then multiplied by the CF associated with the rule

Propagating CFs

- If two CFs are derived for the same conclusion they are combined as follows:
- New CF = CF1 + CF2 – (CF1)(CF2)
  - If both CF1 & CF2 > 0
- New CF = CF1 + CF2 + (CF1)(CF2)
  - If both CF1 & CF2 < 0
- New CF = (CF1 + CF2) / (1- min(|CF1|,|CF2|))
  - Otherwise

Comments on Certainty Factors

- These calculations are not probabilistic
- They assume that the conclusions are independent
- These problems are a restriction on the viability of Mycin, however as long as the same conclusion isn't reached too often this shouldn't create a problem
- Mycin has a very shallow search space

Examples

- If the following are known:
  - A CF(0.4)
  - B CF(0.7)
  - C CF(0.9)
- And the rule
  - If A and B and C then E(0.4)

Examples

- For ands we take min
- So min of 0.4, 0.7, 0.9 = 0.4
- Multiply this by the CF for the rule (0.4)
  - Gives 0.16
- This is < 0.2 and so is ignored

Examples

- Given:
  - A CF(0.4)
  - B CF(0.7)
  - C CF(0.9)
- And the rule:
  - If A or B or C then E(0.4)

Examples

- For or's we take max
- So max of 0.4, 0.7, 0.9 gives 0.9.
- Multiplying by CF of rule (0.4)
  - Gives 0.36
- This is > 0.2 so the CF for E becomes 0.36

Examples

- Given:
  - A CF(0.7)
  - B CF(0.8)
  - C CF(0.5)
  - E CF(0.4)
- And the rule
  - If A or B or C then E (0.7)

Examples

- Or's so max of 0.7, 0.8, 0.5 = 0.8
- Multiply this by CF for rule 0.8*0.7 = 0.56
- > 0.2 so new CF for E is 0.56
- Already have a CF for E of 0.4
- Need to combine them

Examples

- As both > 0 1st rule applies
  - New CF = CF1 + CF2 – (CF1)(CF2)

    = 0.4 + 0.56 – 0.4 * 0.56

    = 0.96 – 0.224

    = 0.736

  New CF for E is therefore 0.736

  This reinforces existing CF

Theoretical Flaws in Mycin

- Mycin assumes that a failure to establish the truth of something, despite attempts to do so means that it is false
- The threshold of 0.2 is arbitrary
- Mycin detects and rejects circularity in rules such as:
  - If A then B
  - If B then C
  - If C then A
- By ignoring the first rule
- This is not always the best approach

Theoretical flaws in Mycin

- Self referencing rules cause a problem
- Example
  - Ident = pseudomonas CF(0.3)
  - Burn = true CF(1.0)
  - Skinless = true CF(0.3)

Theoretical flaws in Mycin

- Given the following rules:
  1) If burn = true and ident = pseudomonas then
     - ident = pseudomonas with CF(1.0)
  2) If skinless = true and ident = pseudomonas then
     - ident = pseudomonas with CF(1.0)

Theoretical flaws in Mycin

- If rule 1 fires first the CF of the premise is 0.3, so the rule gives a new CF for pseudomonas of 0.3*1 = 0.3
- Combining this with the existing CF gives
  - 0.3 + 0.3 – 0.3*0.3 = 0.51
- After applying rule 2 this is updated to
  - 0.51 + 0.3 – 0.51 * 0.3 = 0.657

Theoretical flaws in Mycin

- If rule 2 fires first the CF of the premise is 0.3 so the rule gives a new CF for pseudomonas of 0.3*1.0 = 0.3
- Combining this with the existing CF gives
  - 0.3 + 0.3 – 0.3 * 0.3 = 0.51
- After applying rule 1 this is updated to
  - 0.51 + 0.51 – 0.51 * 0.51 = 0.7599
- The ordering of the rules is ∴ important

Bayesian Networks

A Bayesian network is 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 $P(X_i \mid Parents(X_i))$ 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).

Bayesian Networks

Properties of Bayesian Networks

- Bayesian Networks are a 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
- Captures both qualitative and quantitative relationships between variables
- Can be used to reason

Semantics

- **Global semantics** defines the full joint distribution as the product of the local conditional distributions

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

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

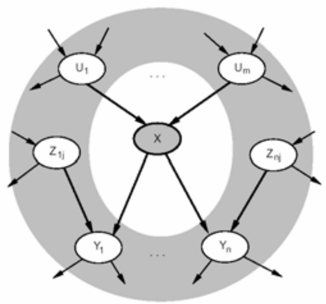
Compactness

- In Bayesian networks, it is reasonable to suppose that in most domains each random variable is directly influenced by at most *k* others, for some constant *k*.
- If we assume n Boolean variables for simplicity, then the amount of information needed to specify each conditional probability table will be at most 2^k numbers, and the complete network can be specified by $n2^k$ numbers.
- In contrast, the joint distribution contains 2^n numbers.
- Example:
  - Suppose we have 30 nodes *(n=30)* and each has at most 5 parents *(k=5)*.
  - Then the Bayesian network requires **960** numbers, but the full joint distribution requires over **a billion**.

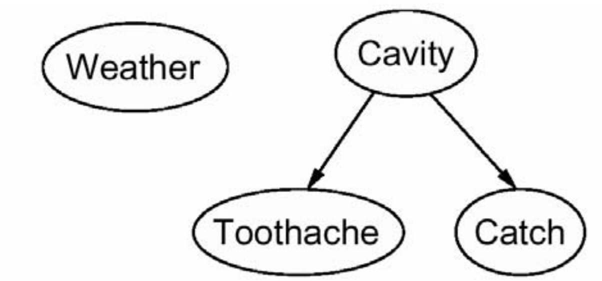
Markov blanket

- Each node is conditionally independent of all others given its Markov blanket:

parents + children + childrens parent

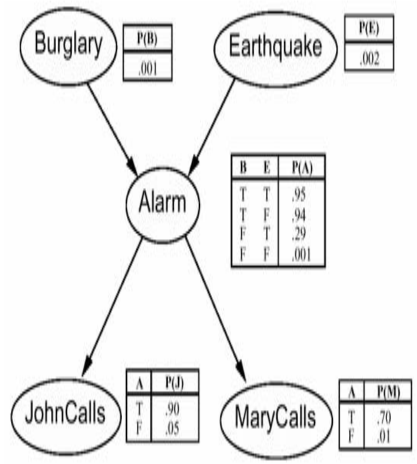


Example



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

Burglary Example



- You have a new burglar alarm installed at home. It is fairly reliable at detecting a burglary, but also responds on occasion to minor earthquakes.
- You also have two neighbors, John and Mary, who have promised to call you at work when they hear the alarm.
- John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm and calls then, too.
- Mary, on the other hand, likes rather loud music and sometimes misses the alarm altogether.
- Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.

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.

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

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

Constructing Belief Networks

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

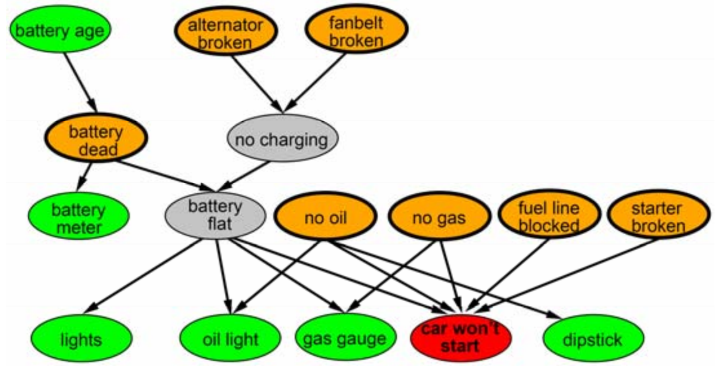
1. Choose an ordering of variables $X_1, \ldots, X_n$
2. For $i = 1$ to n
   add X_i to the network
   select parents from $X_1, \ldots, X_{i-1}$ such that
   $\mathbf{P}(X_i | Parents(X_i)) = \mathbf{P}(X_i | X_1, \ldots, X_{i-1})$

This choice of parents guarantees the global semantics:

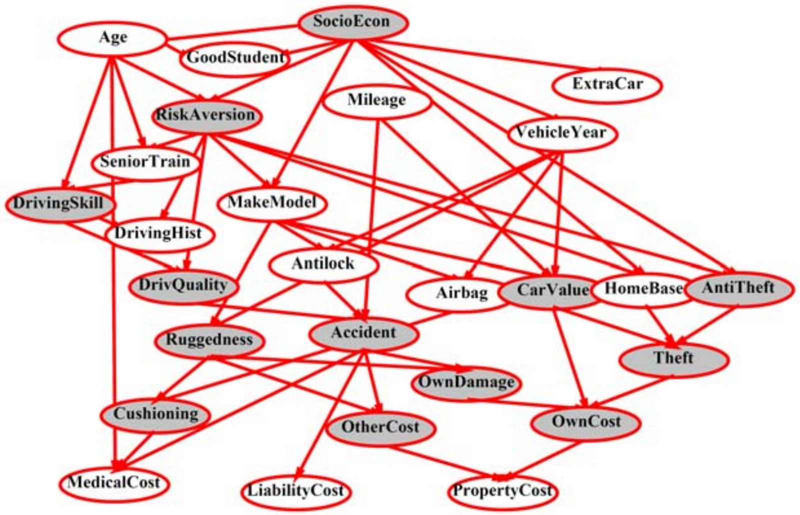
$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i | X_1, \ldots, X_{i-1})$ (chain rule)
$= \prod_{i=1}^{n} \mathbf{P}(X_i | Parents(X_i))$ by construction

Example Car diagnosis

- Initial evidence: car won't start
- Testable variables (green), “broken, so fix it" variables (orange)
- Hidden variables (gray) ensure sparse structure, reduce parameters



Example: Car insurance



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.

History

- Fuzzy, or multi-valued logic was introduced in the 1930s by Jan Lukasiewicz, a Polish philosopher.
- While classical logic operates with only two values 1 (true) and 0 (false), Lukasiewicz introduced logic that extended the range of truth values to all real numbers in the interval between 0 and 1.
- He used a number in this interval to represent the possibility that a given statement was true or false.
- This work led to an inexact reasoning technique often called possibility theory.

History

- In 1965 **Lotfi Zadeh,** published his famous paper "Fuzzy sets".
- Zadeh extended the work on possibility theory into a formal system of mathematical logic, and introduced a new concept for applying natural language terms.
- This new logic for representing and manipulating fuzzy terms was called **fuzzy logic.**

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:

- Conceptually easy to understand with “natural” mathematics
- Tolerant of imprecise data
- Universal approximation: can model arbitrary nonlinear functions
- Intuitive
- 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 → {0, 1}
- Characteristic function:

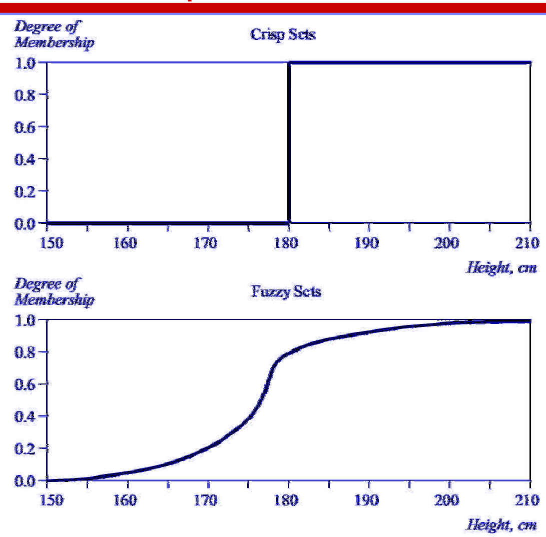
$$\mu_A(x) = \begin{cases} 1 & \text{if x is an element of set A} \\ 0 & \text{if x is not an element of set A} \end{cases}$$

- **Fuzzy set F** is a mapping for the elements of S to the interval [0, 1], i.e., F: S → [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

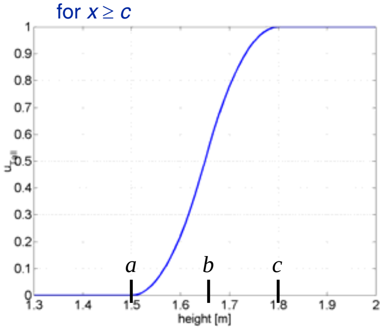
Name	Height, cm	Degree of *Crisp*	Membership *Fuzzy*
Ali	208	1	1.00
Cemal	205	1	1.00
Hasan	198	1	0.98
İsmail	181	1	0.82
Orhan	179	0	0.78
Ahmet	172	0	0.24
Abdullah	167	0	0.15
Barış	158	0	0.06
Çetin	155	0	0.01
Sezai	152	0	0.00

Fuzzy Set Example: Tall Men



Membership functions: S-function

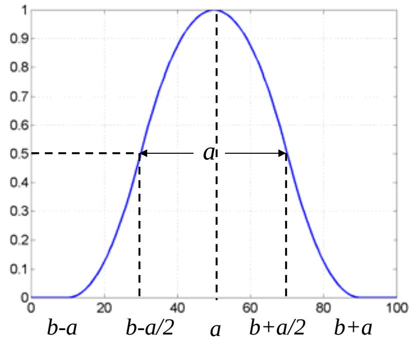
- The S-function can be used to define fuzzy sets
- S(x, a, b, c) =
  - 0 for $x \le a$
  - $2(x\text{-}a/c\text{-}a)^2$ for $a \le x \le b$
  - $1 - 2(x\text{-}c/c\text{-}a)^2$ for $b \le x \le c$
  - 1 for $x \ge c$



Membership functions: P-Function

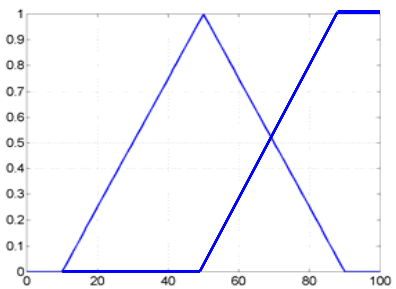
- P(x, a, b) =
  - S(x, b-a, b-a/2, b) for $x \le b$
  - 1 – S(x, b, b+a/2, a+b) for $x \ge b$

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



Simple membership functions

- Piecewise linear: triangular etc.
- Easier to represent and calculate ⇒ saves computation



Other representations of fuzzy sets

- A finite set of elements:

  $F = \mu_1/x_1 + \mu_2/x_2 + \ldots \mu_n/x_n$

  + means (Boolean) set union

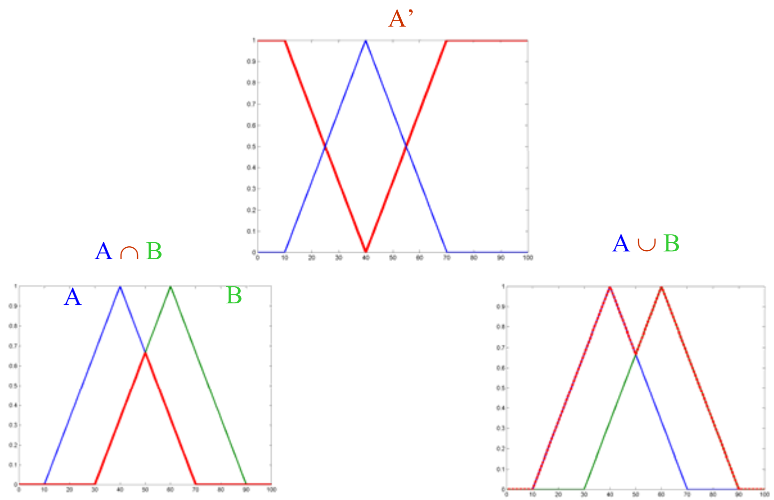
- For example:

  TALL = {0/1.0, 0/1.2, 0/1.4, 0.2/1.6, 0.8/1.7, 1.0/1.8}

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 $x \in X$

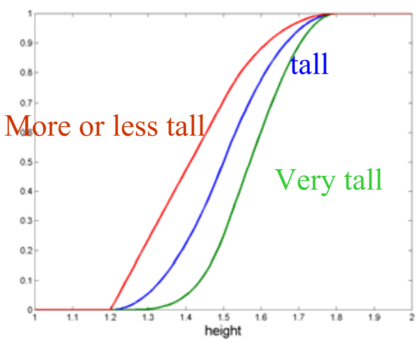
Example fuzzy set operations



Linguistic Hedges

- Modifying the meaning of a fuzzy set using hedges such as *very, more or less, slightly, etc.*

- *Very F = F^2*
- *More or less F = $F^{1/2}$*
- *etc.*



Fuzzy relations

- A fuzzy relation for N sets is defined as an extension of the crisp relation to include the membership grade.

  $R = \{\mu_R(x_1, x_2, \ldots x_N)/(x_1, x_2, \ldots x_N) \mid x_i \in X, i=1, \ldots N\}$

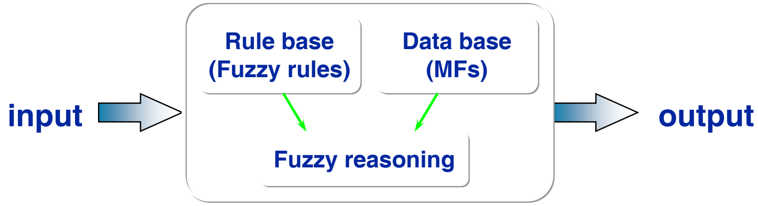
  which associates the membership grade, μ_R, of each tuple.

- E.g.

  Friend = {0.9/(Ali, Niyazi), 0.1/(Ali, Demir),
  0.8/(Ayşe, Melih), 0.3/(Ayşe, Cemal)}

Fuzzy Inference Systems (FIS)

- Also known as
  - Fuzzy models
  - Fuzzy associate memories (FAM)
  - Fuzzy controllers



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
    - Defuzzifies 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 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(μ_A (x), μ_B (x))
- A **OR** B = A $\cup$ B = max(μ_A (x), μ_B (x))

From the following truth tables it is seen that fuzzy logic is a **superset** of Boolean logic.

min(A,B)

A	B	A and B
0	0	0
0	1	0
1	0	0
1	1	1

max(A,B)

A	B	A or B
0	0	0
0	1	1
1	0	1
1	1	1

1-A

A	not A
0	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.

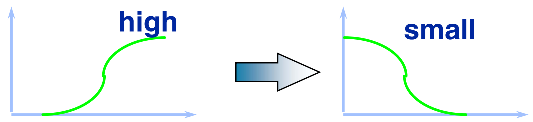
  This reads

  if x == A then y = B

Fuzzy If-Then Rules

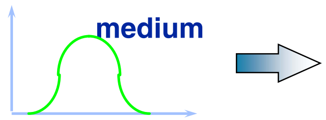
- **Mamdani style**

  If pressure is high then volume is small



- **Sugeno style**

  If speed is medium then resistance = 5*speed



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 => 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 exec 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

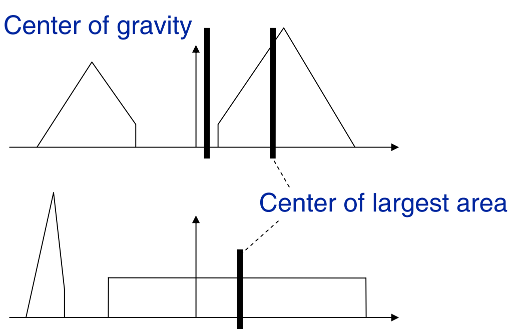
  if E_N then H_N

- $\mu_H = \max[\min(\mu_{E1}), \min(\mu_{E2}), \dots \min(\mu_{EN})]$

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.

$$I = \frac{\Sigma \mu_I x}{\Sigma \mu_I}$$



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

Rule2:

IF x is A2 AND y is B2 THEN z is C2

Rule3:

IF x is A1 THEN z is C3

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

Rule1:

IF project_funding is adequate OR project_staffing is small THEN risk is low

Rule2:

IF project_funding is marginal AND project_staffing is large THEN risk is normal

Rule3:

IF project_funding is inadequate THEN risk is high

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 x1 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

- Our Example
  - 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)
- Clipped or scaled
  - Rule1 = 0.0 + 0.1 - 0.0x0.1=0.1
  - Rule2= 0.2 x 0.7 = 0.14

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

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

Fuzzy Inference- Mamdani

- Our Problem

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

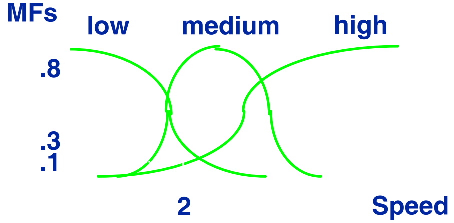
= 67.4

So

z = 67.4

Sugeno Style

If speed is low then resistance = 2
If speed is medium then resistance = 4*speed
If speed is high then resistance = 8*speed

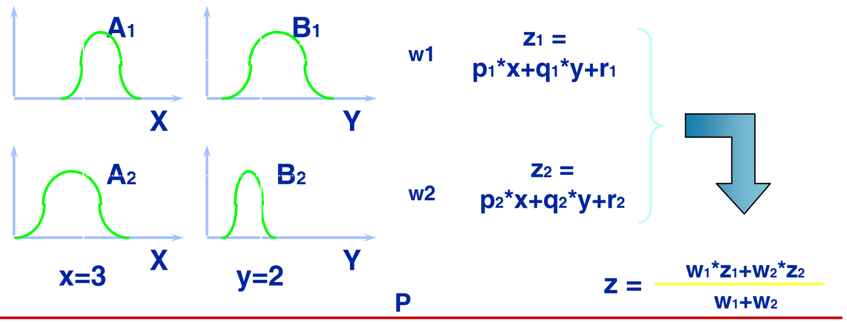


Rule 1: w1 = 0.3; r1 = 2
Rule 2: w2 = 0.8; r2 = 4*2
Rule 3: w3 = 0.1; r3 = 8*2

Resistance = Σ(wi*ri) / Σwi
= 7.12

First-Order Sugeno FIS

- **Rule base**
  If X is A_1 and Y is B_1 then $Z = p_1*x + q_1*y + r_1$
  If X is A_2 and Y is B_2 then $Z = p_2*x + q_2*y + r_2$
- **Fuzzy reasoning**



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* $\Theta = \{O_1, O_2, ..., O_n\}$
  - set of objects O_i that are of interest
  - $\Theta = \{O_1, O_2, ..., O_n\}$
- *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 in the frame of discernment
  - 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) $= \Sigma_{X \cap Y}\, m_1(X) * m_2(Y)$
  $= C\, m_1(X) * m_2(Y) / (1 - \Sigma X \cap Y)$
  $= C\, m_1(X) * m_2(Y)$

  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	$\sum_i Pi = 1$	$m(\Theta) \leq 1$
Subset $X \subseteq Y$	$P(X) \leq P(Y)$	m(X) > m(Y) allowed
relationship X, ¬X (ignorance)	$P(X) + P(\neg X) = 1$	$m(X) + m(\neg X) \leq 1$

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 < Bel < 1
Tends to refute	[0,Pls] where 0 < Pls < 1
Tends to both support and refute	[Bel,Pls] where 0 < Bel ≤ Pls< 1

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
  - usability somewhat unclear