Artificial Intelligence

Unit-VI

Uncertainty measure : Probability Theory

V Semester CSE

By

Dr. Loshma Gunisetti

Associate Professor,CSE

Course Outcomes

After successful completion of this course, the student will be able to:

CO	**Course Outcomes**	**Knowledge Level**
1	Illustrate the concept of Intelligent systems and current trends in AI	K2
2	Apply Problem solving, Problem reduction and Game Playing techniques in AI	K3
3	Illustrate the Logic concepts in AI	K2
4	Explain the Knowledge Representation techniques in AI	K2
5	Describe Expert Systems and their applications	K2
6	**Illustrate Uncertainty Measures**	**K2**

Introduction

- Most Intelligent Systems have some degree of uncertainty associated with them
- Uncertainty in systems arises primarily because of problems in data, major causes being that data is missing or unavailable, data is present but unreliable or ambiguous due to errors in measurement, presence of multiple conflicting measurements and so on
- It is not always possible to represent data in a precise and consistent manner, data is generally based on defaults, which may have exceptions leading to errors in intelligent systems
- Uncertainty may also be caused by the represented knowledge since it might represent only the best guesses of the expert based on observations or statistical conclusion, which may not be appropriate in all situations
- Hence, Uncertainty management is incorporated in Intelligent Systems
- Uncertain information can be handled by Probability Theory, Fuzzy Logic, model and temporal logic

Probability Theory

- Used to estimate the degree of uncertainty
- Oldest method with strong mathematical basis
- Probability is defined as a way of turning an opinion or an expectation into a number lying between 0 and 1
- It reflects the likelihood of an event
- Assume a set S (sample space) consisting of independent events representing all possible outcomes of a random experiment
- Probability of an event A, $P(A) = \frac{\text{(Number of outcomes favourable to A)}}{\text{(Total number of possible outcomes)}}$
- Probability of all events { $A_1, A_2, \ldots, A_n$}, must sum up to certainty, i.e.

  $P(A_1)+P(A_2)+\ldots+P(A_n) = 1$
- Impossible event has probability of 0, while a certain event has a probability 1

Axioms of Probability

If S represents a sample space and A and B represent events, then the following axioms hold true, A' represents complement of set A

- $P(A) \geq 0$
- $P(S) = 1$
- $P(A') = 1 – P(A)$
- $P(A \cup B)=P(A) + P(B)$, if events A and B are mutually exclusive
- $P(A \cup B)=P(A) + P(B) – P(A \cap B)$, if A and B are not mutually exclusive. It is called addition rule of probability

Joint Probability

- Joint Probability is defined as the probability of occurrence of two independent events in conjunction i.e. Joint probability refers to the probability of both events occurring together written as P(A ⋂ B) or P(A and B)
- P(A and B)=P(A) * P(B)
- Two events are said to be independent if the occurrence of one event does not affect the probability of occurrence of the other
- P (A or B) = P(A) + P(B) **-** P(A) * P(B)

Joint Probabilities	**A**	**A´**
B	0.20	0.12
B´	0.65	0.03

P(A) = P(A and B) + P(A and B´**)**=0.20+0.65=0.85

P(B) = P(A and B) + P(A´ and B**)**=0.20+0.12=0.32

P(A or B) = P(A) + P(B) **-** P(A and B)

= P(A and B) + P(A and B´**) +** P(A and B)+P(A´ and B**) -** P(A and B)

= P(A and B´**) +** P(A and B)+P(A´ and B**)**

= 0.65+0.20+0.12=0.97

or

P(A or B) = 1 – P((A or B)´)= 1 – P(A´ and B´)=1**-**0.03=0.97

Conditional Probability

- Probability of an event can depend on the probability of another event
- The concept of conditional probability relates the probability of one event to the occurrence of another event
- It is defined as the Probability of the occurrence of an event H (hypothesis) provided an event E(evidence) is known to have occurred.
- It is denoted by $P(H \mid E)$

$$P(H \mid E) = \frac{\text{Number of events favourable to H which are also favourable to E}}{\text{Number of events favourable to E}}$$

$$= \frac{P(H \text{ and } E)}{P(E)}$$

- If events H and E are independent, then $P(H \mid E) = P(H)$ and $P(E \mid H) = P(E)$

Conditional Probability

Example: Probability of a person chosen at random being literate is 0.40 and probability of any person chosen at random having age > 60 years as 0.005. Find the probability of the fact that a person chosen at random of age > 60 is literate

Solution:

P(X is literate and age of X >60) = P(X is literate) * P(age of X>60) =0.40*0.005=0.002

Probability that a person chosen at random having age > 60 years is literate:

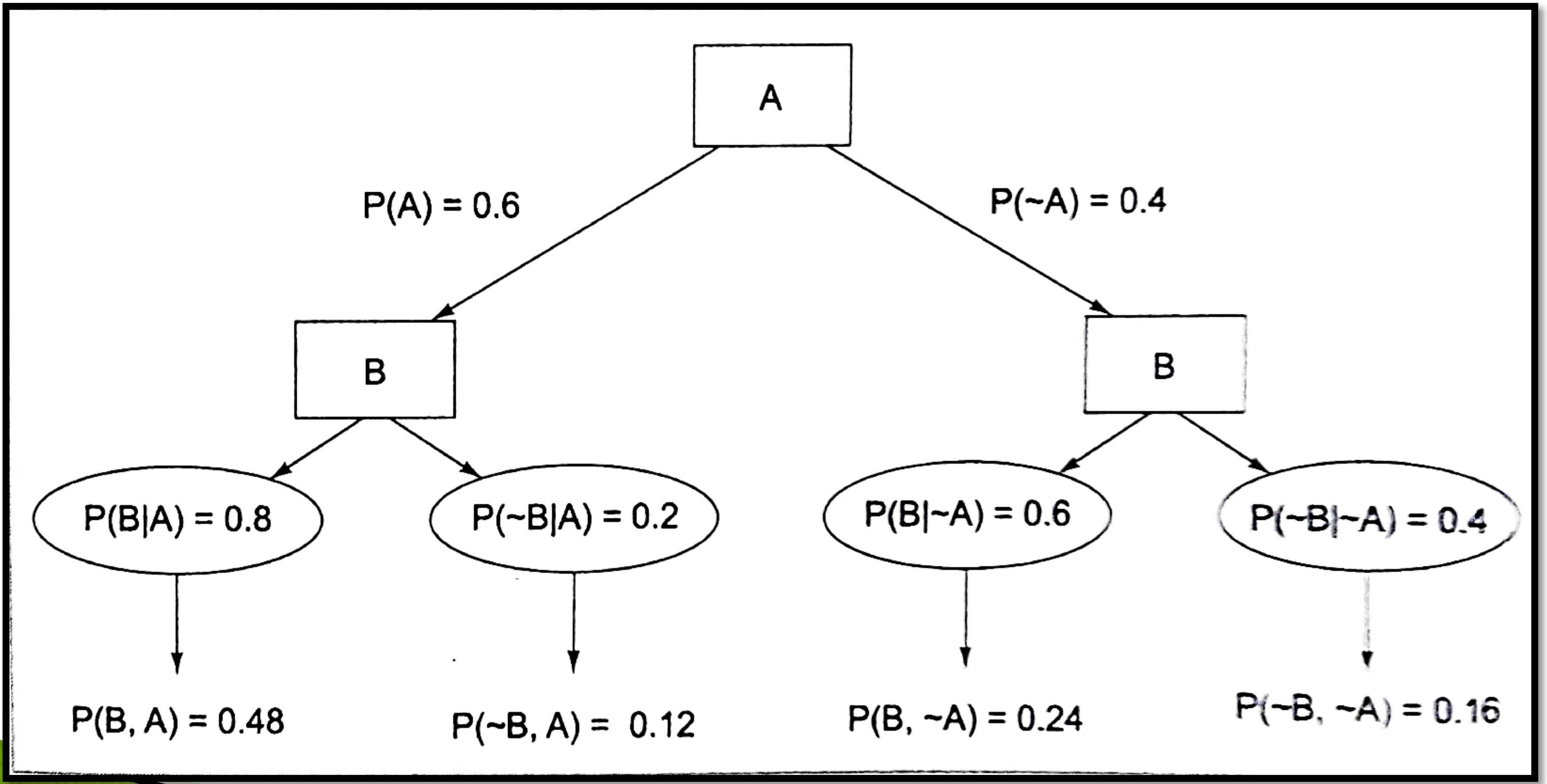
$$P(\text{X is literate} \mid \text{age of X} > 60) = \frac{P(\text{X is literate and age of X} > 60\ )}{P(\text{age of X} > 60)}$$

$$= \frac{0.002}{0.005} = 0.4$$

Conditional Probability

Example: A: Sun is bright today and B : Sun will be bright tomorrow

Given P(sunny_today) = 0.6 , P(sunny_tomorrow | sunny_today) = 0.8



Joint Probability Distribution

Bayes' Theorem

- Developed in 1763 by Thomas Bayes
- It provides a mathematical model for reasoning where prior beliefs are combined with evidence to get estimates of uncertainty
- It relies on the concept that one should incorporate the prior probability of an event into the interpretation of a new situation
- Bayes' theorem can be written as :

$$\mathbf{P(H|E) = \frac{P(E|H) * P(H)}{P(E)}}$$

Proof of Bayes' Theorem

Using Conditional Probability,

P(H|E) = P(H and E) / P(E)

=> P(H|E) * P(E) = P(H and E) ---(1)

Similarly,

P(E|H) = P(E and H) / P(H)

=> P(E|H) * P(H) = P(E and H) ---(2)

From (1) and (2),

P(H|E) * P(E) = P(E|H) * P(H)

Hence we obtain,

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E)} \qquad \text{=> Bayes' Theorem}$$

Bayes’ Theorem

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E|H) * P(H) + P(E|\sim H) * P(\sim H)}$$

As P(E) = P(E and H) + P(E and ~H)

Using Conditional Probability, we obtain,

P(E and H) = P(E|H) *P(H)

P(E and ~H) = P(E|~H) *P(~H)

Therefore P(E) = P(E|H) *P(H) + P(E|~H) *P(~H)

- P(H) is the prior probability of H
- P(H|E) is known as the conditional probability of H, given E
- P(E|H) is known as the conditional probability of E, given H
- P(E) is the prior probability of E and acts as a normalizing constant

Bayes' Theorem Example

Suppose we are given the probability of Mike has a cold as 0.25, the probability of Mike was observed sneezing when he had cold in the past was 0.9 and the probability of Mike was observed sneezing when he did not have cold as 0.20. Find the probability of Mike having a cold given that he sneezes.

Sol. H : Mike has a cold, P(H)=0.25

E : Mike sneezes, P(H|E) to be calculated

P(Mike was observed sneezing | Mike has a cold) =P(E|H) = 0.9

P(Mike was observed sneezing | Mike does not have a cold) = P(E|~H)=0.2

P(Mike has a cold | Mike was observed sneezing) =P(H|E)

$$P(H|E) = \frac{P(E|H) * P(H)}{P(E|H) * P(H) + P(E|\sim H) * P(\sim H)} = \frac{0.9 * 0.25}{0.9*0.25+0.2*0.75} = \frac{0.225}{0.375} = 0.6$$

Similarly

P(H|~E) =[P(~E|H)*P(H)]/P(~E) = [(1-0.9) *0.25] /(1-0.375) =0.025/0.625=0.04

Extension of Bayes' Theorem

One Hypothesis and Two Evidences

Consider a given hypothesis H and two evidences E1 and E2

$$P(H \mid E1 \text{ and } E2) = \frac{P(E1|H) * P(E2|H) * P(H)}{P(E1 \text{ and } E2)}$$

We can derive different forms of the formula :

Form 1:

$$P(H \mid E1 \text{ and } E2) = \frac{P(H) * P(E1|H) * P(E2|H \text{ and } E1)}{P(E1) * P(E2|E1)}$$

Form 2:

$$P(H \mid E1 \text{ and } E2) = \frac{P(E2|H \text{ and } E1) * P(H|E1)}{P(E2|E1)}$$

Extension of Bayes’ Theorem

One Hypothesis and Multiple Evidences

Consider a given hypothesis H and n evidences

$$P(H \mid E1 \text{ and } \ldots \text{and } En) = \frac{P(H \text{ and } E1 \text{ and } \ldots \text{ and } En)}{P(E1 \text{ and } \ldots \text{ and } En)}$$

$$P(H \mid E1, \ldots , En) = \frac{P(H, E1, \ldots, En)}{P(E1, \ldots, En)}$$

Bayes’ theorem to express conditional probability involving n>2 evidences which are assumed to be independent of each other is defined as :

$$P(H \mid E1, \ldots En) = \frac{P(E1|H)* \ldots *P(En|H) *P(H)}{P(E1, \ldots, En)}$$

Chain Evidence :

If E1 is an evidence of H and E2 is an evidence of E1,

$$P(H|E1) = \frac{P(E1|H) * P(E2|E1) * P(H)}{P(E1|E2)*P(E2)}$$

Extension of Bayes' Theorem

Multiple Hypotheses and Single Evidence

Consider a set of hypotheses {H1,H2,…,Hk} and evidence E

$$P(Hi \mid E) = \frac{P(E \mid Hi) * P(Hi)}{P(E)}$$

Another definition is :

$$P(E) = \sum_{j=1}^{k} P(E \mid Hj) * P(Hj)$$

$$P(Hi \mid E) = \frac{P(E \mid Hi) * P(Hi)}{\sum_{j=1}^{k} P(E \mid Hj) * P(Hj)}$$

Multiple Hypotheses and Multiple Evidences

Consider a set of hypotheses{H1,H2,…,Hk} and multiple sources of evidence {E1,E2,…,En}

$$P(Hi \mid E1,\ldots,En) = \frac{P(E1,\ldots,En \mid Hi) * P(Hi)}{\sum_{j=1}^{k} P(E1,\ldots,En \mid Hj) * P(Hj)}$$

Probabilities in Rules and Facts of Rule-Based System

- A fact ‘battery in a randomly picked computer is dead 2% of the time’ can be expressed in Prolog as battery_dead_computer(0.02)
- The rule ‘The probability of the battery being dead is same as the probability of the circuit being faulty’ may be written in Prolog as

  battery_dead_computer(P) :- computer_circuit_faulty(P)

  To ignore weak evidences, we can write as:

  battery_dead_computer(P) :- computer_circuit_faulty(P) , P > 0.1
- The rule ‘If 30% of the time when a computer has a circuit problem, the battery is dead’ can be written as:

  battery_dead_computer(P) :- computer_circuit_faulty(P1) , P1 =0.3

Cumulative Probabilities

- It is very important to combine the probabilities from the facts and successful rules to get a cumulative probability
- Two situations might arise:
  - If sub-goals of a rule are probable, then the probability of the rule to succeed should take care of the probable sub goals
  - If rules with the same conclusion have different probabilities, then the overall probability of the rule has to be found
- Prob(A and B and C and …) = Prob(A)*Prob(B)*Prob(C)* …
- Prob(A or B or C or …) = 1 – [(1-Prob(A))(1-Prob(B))(1-Prob(C)) …]

Prolog Programs for Computing Cumulative Probabilities

AND-combination: A list of all the probabilities is passed as an argument and the product of all these probabilities is computed to obtain **and-combination** effect using Prolog rules as follows:

Prob(A and B and C and …) = Prob(A)*Prob(B)*Prob(C)* …

```
and_combination([P],P)
and_combination([H|T],P) :- and_combination(T,P1), P is P1*H
```

The rule 'when a computer has a circuit problem and the battery is old , then the battery is dead'

```
battery_dead_computer(P):-computer_circuit_faulty(P1),battery_old(P2), P is P1*P2
battery_dead_computer(P):-computer_circuit_faulty(P1), battery_old(P2),
                                    and_combination([P1,P2],P)
```

Prolog Programs for Computing Cumulative Probabilities

OR-combination:

Prob(A or B or C or …) = 1- [(1-Prob(A))(1-Prob(B))(1-Prob(C))…]

or_combination([P],P)

or_combination([H|T],P) :- or_combination(T,P1), P is 1-((1-H)*(1-P1))

overall(P) :- findall(Prob, pred(Prob),L), or_combination(L,P) where L is List containing all the probabilities for which predicate is satisfied

<u>Example</u>: compute(0.4) :- a
compute(0.5) :- b, c
compute(0.2) :- a, d

overall_compute(P) :- findall(P1,compute(P1),L),or_combination(L,P)

L=[0.4,0.5,0.2] , P =(1-0.6*0.5*0.8) = 0.760

Handling Negative Probabilities

- Probable sub goals in rules which are not true can cause problems if not handled properly
- For example the rule: g(P) :- a(P1),b(P2),not(c(P3)) where a,b,c are all certain subgoals
- If the rule is 80% probable, then cumulative probabilities cannot be calculated using and_combination
- Hence the rule can be rewritten as:

```
g(P) :- a(P1),b(P2),not(c(P3)), inverse(P3,P4),
            and_combination([P1,P2, P4,0.8],P)
inverse(P,Q) :- Q is 1-P
```

Rule-Based System Using Probability: Example

- A simple rule-based system for the diagnosis of malfunctioning of some equipment, for eg. a landline telephone
- Following could be the reasons for the malfunctioning of telephones:
  - ➢ Faulty instrument
  - ➢ Problem in Exchange
  - ➢ Broken Cable

Rules regarding Faulty Telephone Instrument

Rule 1: If the telephone instrument is old and has been repaired several times in the past then it is 40% sure that the fault lies with the instrument. In Prolog:

```
telephone_not_working(0.4)  :- ask(tele_history)
```

Rule-Based System Using Probability: Example

Rules regarding Faulty Telephone Instrument

Rule 2: If the instrument has fallen on the ground and broken, then it is 80% sure that the fault lies with the instrument. In Prolog:

```
telephone_not_working(0.8)  :- ask(telephone_broken)
```

Rule 3: If there are children in the house who play with the keypad of the telephone, with some probability, then it is 80% sure that the instrument is faulty because of excessive and unusual usage. In Prolog:

```
telephone_not_working(P):-ask(children_present,P1), ask(children_keypad,P2),
                                    and_combination([P1,P2,0.8],P)
```

Cumulative probability for telephone not working because of fault of instrument, the rule is :

```
instrument_faulty(P) :- findall(X,telephone_not_working(X),L),
                              or_combination(L,P)
```

Rule-Based System Using Probability: Example

Rules for diagnosis of Instrument fault

The rule for diagnosis based on instrument fault which we assume to be 70% sure can be written as:

```
diagnosis(‘Instrument is faulty’, P) :- instrument_faulty(P1),
                                        and_combination([P1,0.7],P)
```

Bayesian Method: Advantages and Disadvantages

Advantages

- Bayesian method is based on a strong theoretical foundation in probability theory, it is currently the most advanced of all certainty methods
- This method has well-defined semantics for decision making

Disadvantages

- The system using Bayesian approach needs quite a large amount of probability data to construct a knowledge base
- Since conditional probabilities are based on statistical data, the sample sizes must be sufficient to ensure that the probabilities obtained from them are accurate
- If conditional probabilities are based on human experts, then the question of values being consistent and comprehensive arises

Bayesian Method: Advantages and Disadvantages

Disadvantages

- Since associations between the hypotheses and evidences are reduced to numbers, it eliminates the actual knowledge embedded within the data. Therefore, a system’s ability to explain its reasoning and browsing through the hierarchy of evidences to hypothesis to a user is lost

In spite of these problems, Bayesian theory provides an attractive basis for an uncertain reasoning system. Other mechanisms developed to utilize the power of this theory and ease of its implementation are:

- Bayesian Belief Networks
- Certainty Factor Theory
- Dempster-Shafer Theory

Bayesian Belief Networks(BBN)

- Joint probability requires 2^n entries for n variables to store the distribution explicitly
- The time and storage requirements for such computations become impractical as the value of n increases
- Bayesian Belief Network is a probabilistic graphical model that encodes probabilistic relationships among a set of variables with their probabilistic dependencies
- It is an efficient structure for storing joint probability distribution
- Only one branch of the tree needs to be traversed in the network
- It may be formally defined as an acyclic directed graph where the nodes of the graph represent evidence or hypotheses and an arc connecting two nodes represents dependence between them

Bayesian Belief Networks(BBN)

- Joint probability of n variables (dependent or independent) may be computed as

  $P(X1,\ldots,Xn) = P(Xn|X1,\ldots Xn\text{-}1) * P(X1,\ldots,Xn\text{-}1)$ or

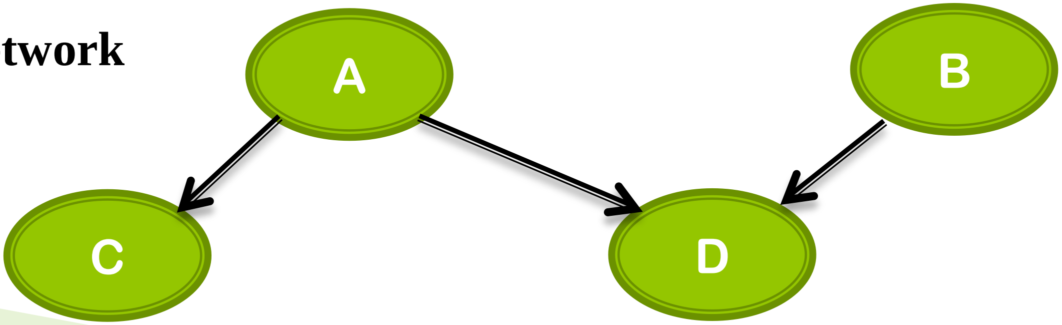
  $P(X1,\ldots,Xn) = P(Xn|X1,\ldots Xn\text{-}1) * P(Xn\text{-}1|X1,\ldots,Xn\text{-}2) * \ldots * P(X2|X1) * P(X1)$

- In BBN, Joint Probability Distribution can be written as :

  $P(X1,\ldots,Xn) = \prod_{i=1}^{n} P(Xi \mid parent\_nodes(Xi))$

- The nodes having parents are called conditional nodes, if the value of a node is observed , then the node is said to be an evidence node
- Nodes with no children are termed as hypotheses nodes, while nodes with no parents are called independent and unconditional nodes

Bayesian Belief Network



Bayesian Belief Networks(BBN)

For Example:

P(A) =0.3

P(B) =0.6

P(C|A) = 0.4

P(C|~A) = 0.2

P(D|A,B) = 0.7

P(D|A,~B) = 0.4

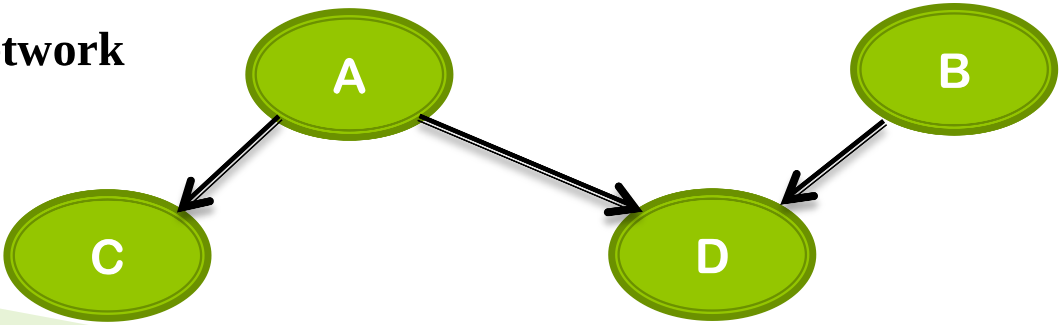
P(D|~A,B) = 0.2

P(D|~A,~B)= 0.01

Conditional Probability Table

P(A)	**P(B)**	**A**	**P(C)**	**A**	**B**	**P(D)**
0.3	0.6	T	0.4	T	T	0.7
		F	0.2	T	F	0.4
				F	T	0.2
				F	F	0.01

Bayesian Belief Network



Bayesian Belief Networks(BBN)

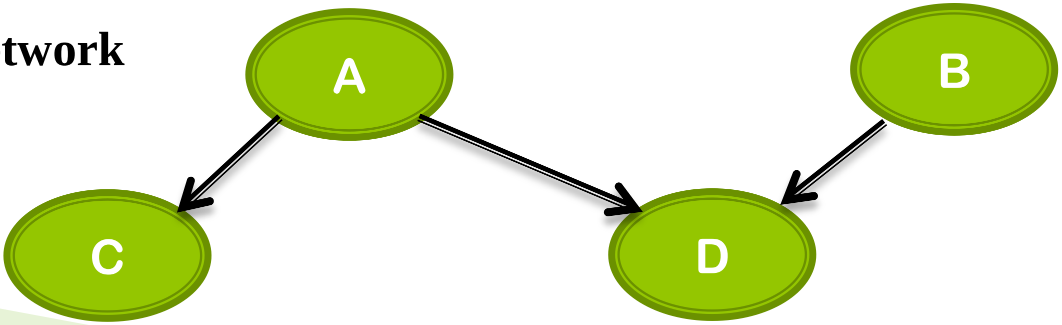
P(A,B,C,D)= P(D|A,B,C) *P(C|A,B) * P(B|A)*P(A)

P(A,B,C,D)= P(D|A,B) *P(C|A) * P(B)*P(A)=0.7*0.4*0.6*0.3=0.0504

Above result can be obtained by using BBN formula for joint probability:

$P(X1,\ldots,Xn) = \prod_{i=1}^{n} P(Xi \mid parent\_nodes(Xi))$

Bayesian Belief Network



Inference using Bayesian Belief Networks(BBN)

What is the likelihood or probability that the hypothesis is A given C?

$P(A|C) = P(A,C) / P(C)$

where

$P(A,C) = \sum_{B,D \in \{T,F\}} P(A, B, C, D)$

$P(C) = \sum_{A,B,D \in \{T,F\}} P(A, B, C, D)$

$$P(A|C) = \frac{\sum_{B,D \in \{T,F\}} P(A, B, C, D)}{\sum_{A,B,D \in \{T,F\}} P(A, B, C, D)}$$

$\sum_{B,D \in \{T,F\}} P(A, B, C, D) = P(A,B,C,D)+P(A,\sim B,C,D)+P(A,B,C,\sim D)+ P(A,\sim B,C,\sim D)$

$\sum_{A,B,D \in \{T,F\}} P(A, B, C, D) = P(A,B,C,D)+P(A, B,C, \sim D)+P(A, \sim B,C,D)+P(A,\sim B,C,\sim D)$

$+P(\sim A,B,C,D)+P(\sim A,\sim B,C,D)+P(\sim A, B,C, \sim D)+P(\sim A,\sim B,C,\sim D)$

Inference using Bayesian Belief Networks(BBN)

P(A,B,C,D) = P(D|A,B) *P(C|A) * P(B)*P(A)=0.7*0.4*0.6*0.3=0.0504

P(A,~B,C,D)=P(D|A,~B) *P(C|A) * P(~B)*P(A)=0.4*0.4*0.4*0.3=0.0192

P(A,B,C,~D)= P(~D|A,B) *P(C|A) * P(B)*P(A)=0.3*0.4*0.6*0.3=0.0216

P(A,~B,C,~D)=P(~D|A,~B) *P(C|A) * P(~B)*P(A)=0.6*0.4*0.6*0.3=0.0288

Numerator=0.0504+0.0192+0.0216+0.0288=0.12

P(~A,B,C,D)=P(D|~A,B) *P(C|~A) * P(B)*P(~A)=0.7*0.4*0.6*0.7=0. 1176

P(~A,~ B,C,D)=P(D|~A,~B) *P(C|~A) * P(~B)*P(~A)=0.01*0.2*0.4*0.7=0.00056

P(~A, B,C, ~D)=P(~D|~A,B) *P(C|~A) * P(B)*P(~A)=0.8*0.2*0.6*0.7=0.0672

P(~A,~B,C,~D)=P(~D|~A,~B) *P(C|~A) * P(~B)*P(~A)=0.99*0.2*0.4*0.7=0.05544

Denominator=0.12+ 0.1176+0.00056+0.0672+0.05544=0.3608

Hence P(A|C) = 0.12/0.3608 = 0.33(approx.)

Bayesian Belief Network : Advantages and Disadvantages

Advantages

- Since this model encodes dependencies among all variables, it can easily handle situations where some data entries are missing
- It is intuitively easier for a human to understand direct dependencies and local distributions than complete joint distributions
- It can be used to learn causal relationships and hence can be used to gain understanding about a problem domain and to predict the consequences of intervention
- It is an ideal representation for combining prior knowledge and data because the model has both causal and probabilistic semantics
- Bayesian statistical methods in conjunction with Bayesian Networks offer an efficient and principled approach for avoiding over-fitting of data

Bayesian Belief Network : Advantages and Disadvantages

Disadvantages

- The probabilities are described as a single numeric point value. This can be a distortion of the precision that is actually available for supporting evidence
- There is no way to differentiate between ignorance and uncertainty. These are distinct two different concepts and be treated as such
- There exists the computational difficulty of exploring a previously unknown network
- All branches must be calculated to find the probability of any branch of the network, which is a NP-hard task. It might either be too costly to perform or impossible, given the number and combination of variables
- The quality and extent of the prior beliefs used in Bayesian inference processing are major shortcomings

Bayesian Belief Network : Advantages and Disadvantages

Disadvantages

- Reliability of Bayesian Network depends on the reliability of prior knowledge
- Selecting the proper distribution model to describe the data has a notable effect on the quality of the resulting network. Therefore, selection of the statistical distributions for modeling the data is very important

Certainty Factor Theory

- Certainty Factor Theory provides another way of measuring uncertainty by describing a practical way of compromising on pure Bayesian system
- Certainty Factor(CF) is based on a number of observations
- The certainty factor is based on some simplifying assumptions for creating confidence measures and for combining these confidences
- These assumptions involve representing **confidence for** and **confidence against** separately by measures of belief (MB)and the measure of disbelief (MD) respectively
- MB[H,E] is a measure of belief in the range [0,1] in the hypothesis H given the evidence E. If evidence supports it fully then MB[H,E]=1 otherwise 0
- MD[H,E] is a measure of disbelief in the range [0,1] in the hypothesis H given the evidence E. It measures the extent to which the evidence E supports the negation of the hypothesis H. MD is not complement of MB

Certainty Factor Theory

- The measure of belief calculates the relative decrement of disbelief in a given hypothesis H due to some evidence E

$$MB[H,E] = \frac{(1-P(H)) - (1-P(H|E))}{(1-P(H))} = \frac{P(H|E) - P(H)}{(1-P(H))}$$

- In order to avoid getting a negative value of belief, MB may be written as:

$$MB[H,E] = \begin{cases} 1, & P(H)=1 \\ \dfrac{\mathrm{Max}\{P(H|E), P(H)\} - P(H),}{1 - P(H)} & \text{otherwise} \end{cases}$$

- Similarly the measure of disbelief(MD) is similarly defined as the relative decrement of belief in a given hypothesis H due to some evidence E

$$MD[H,E] = \frac{P(H) - P(H|E)}{P(H)}$$

$$MD[H,E] = \begin{cases} 1, & P(H)=0 \\ \dfrac{P(H) - \mathrm{Min}\{P(H|E), P(H)\},}{P(H)} & \text{otherwise} \end{cases}$$

Certainty Factor Theory

Certainty Factor, CF[H,E] =MB[H,E] – MD[H,E] where -1≤ CF[H,E] ≤1

- Positive certainty factor indicates evidence for the validity of the hypothesis
- Belief in the hypothesis is directly proportional to the value of CF
- If CF=1, then hypothesis is said to be true, while if CF= -1, the hypothesis is considered to be false. If CF=0, then there is no evidence regarding whether the hypothesis is true or false
- These measures satisfy the following properties:
    - ➢ If hypothesis H is true assuming E, then MB[H,E]=1,MD[H,E]=0 and CF[H,E]=1
    - ➢ If hypothesis H is not true assuming E, then MB[H,E]=0,MD[H,E]=1 and CF[H,E]=-1
    - ➢ The sum of CF of belief and disbelief of hypothesis H, given evidence E is 0, i.e. CF[H,E]+CF[~H,E]=0

Certainty Factor Theory

Method for computing CF in three cases:

Case 1 : Incrementally Acquired Evidence

(Two evidences supporting hypothesis H)

$$\text{MB[H,E1 and E2]=} \begin{cases} 0, & \text{if MD[H, E1 and E2]=1} \\ \text{MB[H,E1]+MB[H,E2]*(1-MB[H,E1]),} & \text{otherwise} \end{cases}$$

Similarly,

$$\text{MD[H,E1 and E2]=} \begin{cases} 0, & \text{if MB[H, E1 and E2]=1} \\ \text{MD[H,E1]+MD[H,E2]*(1-MD[H,E1]),} & \text{otherwise} \end{cases}$$

CF can be written as

CF[H,E1 and E2]=MB[H,E1 and E2] - MD[H,E1 and E2]

Certainty Factor Theory

Method for computing CF in three cases:

Case 2 : Combination of two hypotheses based on same evidence

(Two hypotheses H1 and H2 based on the same evidence E)

Measures of belief and disbelief for conjunction of hypotheses are defined as:

MB[H1 and H2,E]= Min(MB[H1,E],MB[H2,E])

MD[H1 and H2,E]= Max(MD[H1,E],MD[H2,E])

Then, CF[H1 and H2,E]=MB[H1 and H2,E] - MD[H1 and H2,E]

Measures of belief and disbelief for disjunction of hypotheses are defined as:

MB[H1 or H2,E]= Max(MB[H1,E],MB[H2,E])

MD[H1 or H2,E]= Min(MD[H1,E],MD[H2,E])

Then, CF[H1 or H2,E]=MB[H1 or H2,E] - MD[H1 or H2,E]

Certainty Factor Theory

Method for computing CF in three cases:

Case 3 : Chained Rule

In chained rule, the rules are chained together with the result that the outcome of one rule is the input of another rule. For eg. E1 → E2 → H

Let E1 be the evidence that made us to believe E2. CF[E2,E1] be the confidence factor of E2 given evidence E1

If we are sure about the validity of E2, then

$$MB[H,E2] = MB1[H,E2]*Max\{0,CF[E2,E1]\}$$

Similarly,

$$MD[H,E2] = MD1[H,E2]*Min\{1,CF[E2,E1]\}$$

Then, $CF[H,E2] = MB[H,E2] – MD[H,E2]$

Use of CF in MYCIN

Certainty Factor of Conjunction and Disjunction of Hypotheses

CF[H1 and H2] = Min(CF[H1],CF[H2])

CF[H1 or H2] = Max(CF[H1],CF[H2])

Combining Two Certainty Factors

$$\text{Combine}(CF[H1],CF[H2]) = \begin{cases} \{CF[H1]+CF[H2]-(CF[H1]*CF[H2])\}, \\ \qquad \text{if}\{CF[H1]>0 \text{ and } CF[H2]>0\} \\ \{CF[H1]+CF[H2]+(CF[H1]*CF[H2])\}, \\ \qquad \text{if}\{CF[H1]<0 \text{ and } CF[H2]<0\} \\ \{CF[H1]+CF[H2]\}/1-\text{Min}(|CF[H1]|,|CF[H2]|), \text{ otherwise} \end{cases}$$

Dempster-Shafer Theory

- The Dempster-Shafer Theory (or the D-S theory) was developed by AP Dempster in 1968 and was extended by G Shafer in 1976
- It is a mathematical theory of belief functions which is basically a generalization of the Bayesian theory of probability
- Belief functions allow us to base degrees of belief or confidence for one event on probabilities of related events, whereas Bayesian Theory requires probabilities for each event
- The degrees of belief may or may not have the mathematical properties of probabilities
- Dempster-Shafer degrees of belief resemble certainty factors
- D-S theory is more attractive because it is relatively flexible and is based on two ideas, namely, obtaining degrees of belief for one event from probabilities of related events and obtaining a rule for combining such degrees of belief when they are based on independent items of evidence

Dempster Theory Formalism

- Let U be the universal set of all hypotheses, propositions or statements under consideration
- The power set P(U), is the set of all possible subsets of U, including the empty set ϕ
- The theory of evidence assigns a belief mass to each subset of the power set
- A function m: P(U) → [0,1] is called a basic belief assignment (BBA) function. It satisfies the following two axioms:
  - **m(ϕ)=0**
  - Σm(A)=1, ∀A ∈ P(U)
- Sometimes, m is also known as a basic probability assignment
- m(A) is a measure of that portion of the total belief committed exactly to A but to no particular subset of A
- The value of m(A) is called mass assigned to A on the unit interval and pertains only to the set A and makes no additional claims about any subsets of A

Dempster’s Rule of Combination

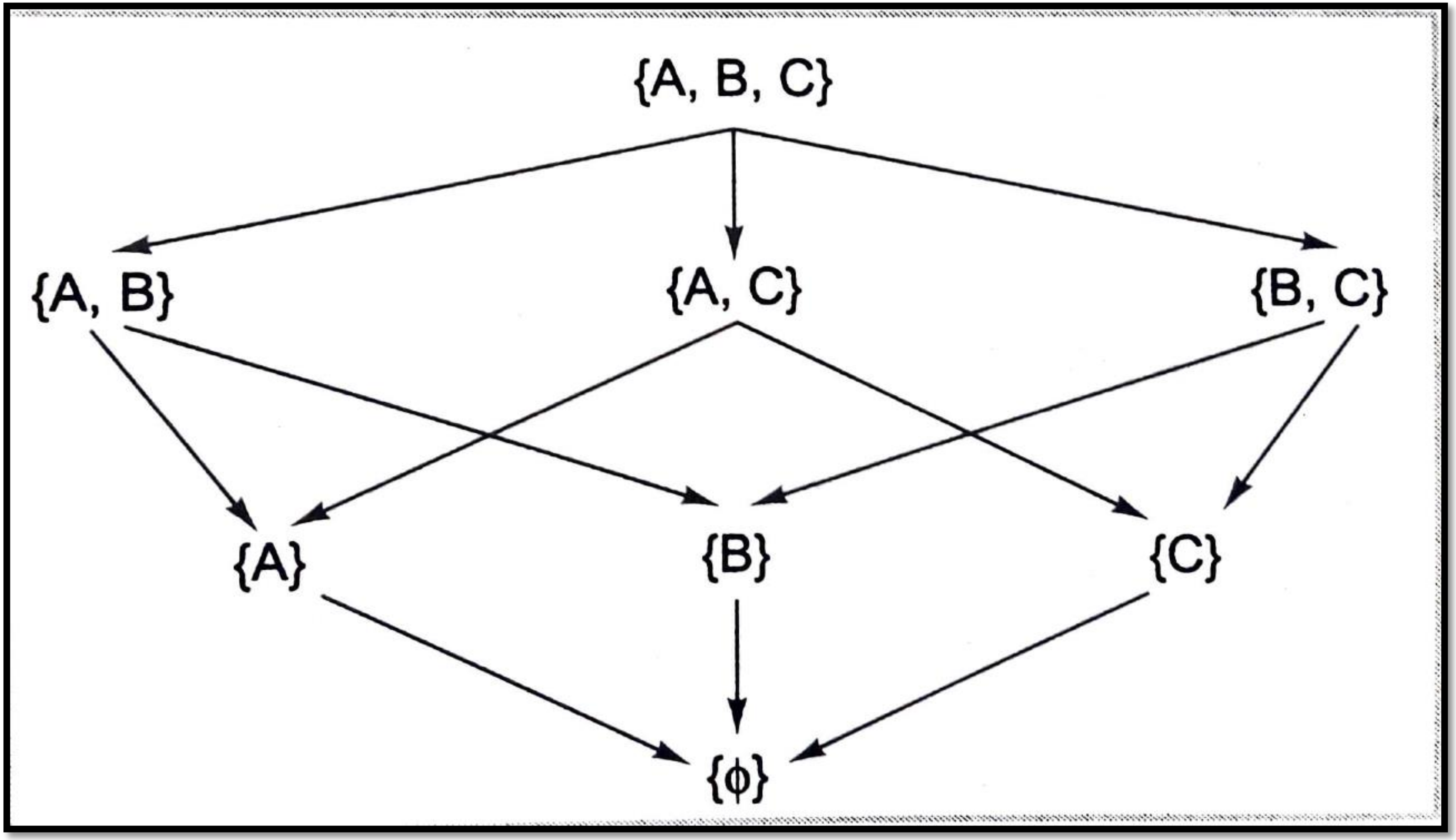
- It is a generalization of Bayes’ rule. It strongly emphasizes the agreement between multiple sources and ignores all the conflicting evidence through a normalization factor
- Assume that m1 and m2 are two belief functions used for representing multiple sources of evidences for two different hypotheses
- Let A,B $\subseteq$ U, such that $m1(A) \neq 0$, $m2(B) \neq 0$. The Dempster’s rule for combining two beliefs functions to generate a function m3 may be defined as

$$\mathbf{m3(\phi)=0}$$

$$\mathbf{m3(C)} = \frac{\sum_{\mathbf{A \cap B = C}} (\mathbf{m1}(\mathbf{A}) * \mathbf{m2}(\mathbf{B}))}{\mathbf{1} - \sum_{\mathbf{A \cap B = \phi}} (\mathbf{m1}(\mathbf{A}) * \mathbf{m2}(\mathbf{B}))}$$

- This belief function gives us a new value when applied on the set C, where C = A ⋂ B
- The combination of two belief functions is called joint mass
- m3 can also be written as (m1 o m2) and $[\sum_{A \cap B = \phi} (m1(A) * m2(B))\ ]$ known as normalization factor, is a measure of the amount of conflict between the two mass sets

Dempster’s Rule of Combination



Lattice Subset of U = { A, B, C}

Dempster's Rule of Combination

Combination of m1 and m2	**m2({A,C})=0.6**	**m2({B,C})=0.8**	**m2(U)=0.4**
m1({A,B})=0.4	m3({A})=0.24	m3({B})=0.32	m3({A,B})=0.16
m1(U)=0.2	m3({A,C})=0.12	m3({B,C})=0.16	m3({U})=0.08

Joint Belief of m1 and m2

Example of a diagnostic system

Suppose we have mutually exclusive hypotheses represented by U={flu,measles,cold,cough} , Assume m(U)=1

Suppose we acquire evidence(say fever) that supports the correct diagnosis in the set {flu, measles} with its corresponding m value as 0.8 Then we get m({flu, measles})=0.8 and m(U)=0.2

Dempster’s Rule of Combination

Example of a diagnostic system

Two belief functions m1 and m2 based on the evidence of fever and headache

m1({flu, measles})=0.8

m1(U)=0.2

m2({flu, cold})=0.6

m2(U)=0.4

Value of m3

Combination of m1 and m2	**m2({flu, cold})=0.6**	**m2(U)=0.4**
m1({flu, measles})=0.8	m3({flu})=0.48	m3({flu, measles})=0.32
m1(U)=0.2	m3({flu,cold})=0.12	m3({U})=0.08

Dempster's Rule of Combination

Example of a diagnostic system

Further if we have another evidence function m4 of sneezing with the following belief values:Two belief functions m1 and m2 based on the evidence of fever and headache

m4({cold, cough})=0.7

m4(U)=0.3

Value of m5

Combination of m3 and m4	**m4({cold, cough})=0.7**	**m4(U)=0.3**
m3({flu})=0.48	m5(ϕ)=0.336	m5({flu})=0.144
m3({flu,cold})=0.12	m5({cold})=0.084	m5({flu,cold})=0.036
m3({flu,measles})=0.32	m5(ϕ)=0.224	m5({flu,measles})=0.096
m3({U})=0.08	m5({cold,cough})=0.056	m5(U)=0.024

Dempster’s Rule of Combination

Example of a diagnostic system

Total belief for empty set (ϕ) is 0.56

We have to scale down the remaining values of non-empty sets by dividing a factor (1-0.56=0.44) Hence, the final belief values may be written as:

m5({flu})	=(0.144/0.44)=0.327
m5({cold})	=(0.084/0.44)=0.191
m5({flu,cold})	=(0.036/0.44)=0.082
m5({flu,measles})	=(0.096/0.44)=0.218
m5({cold,cough})	=(0.056/0.44)=0.127
m5(X)	=(0.024/0.44)=0.055

Therefore we have seen that degree of belief to a set will keep on changing if we get more evidences supporting or opposing it

Dempster’s Rule of Combination

Handling different situations

- If we get an empty set (ϕ) by intersection operation, then we have to redistribute any belief that is assigned to ϕ sets proportionately across non-empty sets using the value $\mathbf{1 - \sum_{A \cap B = \phi} (m1(A) * m2(B))}$ in the denominator of belief values for non-empty sets
- While computing a new belief, we may obtain same subset generated from different intersection process. The value m for this set is computed by summing all such values

Belief and Plausibility

- The term belief (or support) of a set A denoted by bel(A) may be defined formally as follows:
- **Definition: Belief is the sum of all masses of subsets of set A of interest and may be expressed as:**

  $$\mathbf{bel(A)= \sum m(B)\ , \forall B \subseteq A}$$

- For example, if X = {A,B,C}, then

  $$\mathrm{bel}(X) = m(A)+m(B)+m(C) +m(\{A,B\}) +m(\{A,C\}) +m(\{B,C\}) +m(\{A,B,C\})$$

- In Dempster’s theory, a belief interval can also be defined for a subset A. It is represented as sub-interval [bel(A),pl(A)] of [0,1], where pl(A) is called plausibility of A
- **Definition: The plausibility may be formally defined as the sum of all masses of the sets B that intersect the set of interest A. It may be expressed as:**
- $\mathbf{pl(A) = \sum m(B)\ , \forall B}$ **such that** $\mathbf{B \cap A \neq \phi}$
- Precise probability of a set of interest, say P(A) : $\mathrm{bel}(A) \leq P(A) \leq \mathrm{pl}(A)$

Belief and Plausibility

- Some important results are:
  - $pl(A) \geq bel(A)$
  - $pl(\phi) = bel(\phi) = 0$
  - $bel(A) + bel(\sim A) \leq 1$
  - $pl(A) + pl(\sim A) \geq 1$
  - $bel(U) = pl(U) = 1$
  - $pl(A) + bel(\sim A) = 1$ $[pl(A) = 1 - bel(\sim A)]$
- Dempster and Shafer suggested use of the interval$[bel(A), pl(A)]$ to measure uncertainty of A

Belief and Plausibility

- For example: Suppose for the proposition **The Car is Stolen**, we have belief of 0.6 and a plausibility of 0.8, that is we have evidence that the proposition is true with a confidence of 0.6. The evidence of **Car is not stolen** is 0.2 (bel(~A) = 1 – pl(A))
- Then remaining mass of 0.2 (the gap between supporting evidence and contrary evidence i.e. 1-(0.6+0.2)= 0.2 is indeterminate implying that the car may or may not be stolen

Hypothesis	Mass	Belief	Plausibility
Null (neither stolen or not stolen)	0	0	0
Car stolen	0.6	0.6	0.8
Car not stolen	0.2	0.2	0.4
Either(stolen or not)	0.2	1.0	1.0

Mass, Belief and Plausibility values of Hypothesis

Belief and Plausibility

- The null hypothesis is set to zero by definition as it corresponds to no solution
- The hypotheses stolen and not stolen have degree of belief (mass) of 0.6 and 0.2, respectively
- Finally **Either** hypothesis gets value of 0.2 so that the sum of the masses is 1
- The belief for the stolen and not stolen hypotheses matches their corresponding masses because they have no subsets
- Belief for **Either** consists of the sum of all three masses (Either, stolen and not stolen) = 1.0 because stolen and not stolen are each subsets of **Either**
- The plausibility for stolen is 1-m(not stolen) = 1.0-0.2 =0.8 and for ‘not stolen’ is 1-m(stolen)=1.0-0.6=0.4
- Plausibility for **Either** is the sum of m(stolen)+m(not stolen)+m(Either)=1.0
- The universal hypothesis (Either) will always have 100% support and plausibility

Advantages of Dempster Shafer Theory

- Prior and conditional probabilities need not be specified in Dempster –Shafer framework
- Any information contained in the missing prior and conditional probabilities is not used in the Dempster –Shafer framework unless it can be obtained indirectly. D-S Theory allows one to specify a degree of ignorance in this situation instead of being forced to supply prior probabilities which add to unity

Textbook:

1. Artificial Intelligence, Saroj Kaushik, 1st Edition, Cengage Learning

Reference Books:

1. Artificial Intelligence, Elaine Rich, Kevin Knight, Shivashankar B Nair, 3rd Edition, Tata McGraw Hill Education Private Limited., 2009
2. Artificial Intelligence- A modern Approach, 3rd Edition, Stuart Russel, Peter Norvig, Pearson Education