UNIT-4
Statistical reasoning

By: Mohit Goel

Reasoning

- The reasoning is the mental process of deriving logical conclusion and making predictions from available knowledge, facts, and beliefs.
- "**Reasoning is a way to infer facts from existing data.**"
- It is a general process of thinking rationally, to find valid conclusions.
- In artificial intelligence, the reasoning is essential so that the machine can also think rationally as a human brain, and can perform like a human.

Symbolic versus statistical reasoning

The (Symbolic) methods basically represent uncertainty belief as being

- **True,**
- **False, *or***
- **Neither True nor False.**

Some methods also had problems with

- **Incomplete Knowledge**
- **Contradictions in the knowledge.**

Statistical methods provide a method for representing beliefs that are uncertain but for which there may be some supporting (or contradictory) evidence.

- This is useful for dealing with problems where there is randomness and unpredictability (such as in games of chance).
- To do all this in a principled way requires techniques for probabilistic reasoning.

Probabilistic reasoning

- Probabilistic reasoning is a way of knowledge representation where we **apply the concept of probability to indicate the uncertainty in knowledge.**
- In the real world, there are lots of scenarios, where the certainty of something is not confirmed, such as
  - ***"It will rain today,"***
  - ***"behavior of someone for some situations,"***
  - ***"A match between two teams or two players."***
- These are probable sentences for which we can assume that **it will happen but not sure about it**, so here use probabilistic reasoning.

Bayes Theorem

- **The notion of conditional probability : P(H\E)**

Let :

$P(H_i \backslash E) =$ the probability that hypothesis H_i is true given evidence E
$P(E \backslash H_i) =$ the probability that we will observe evidence E given that hypothesis i is true
$P(H_i) =$ the *a priori* probability that hypothesis i is true in the absence of any specific evidence. These probabilities are called prior probabilities or *prlors*.
$k =$ the number of possible hypotheses

Bayes' theorem then states that

$$P(H_i \backslash E) = \frac{P(E \mid H_i) \cdot P(H_i)}{\sum_{n=1}^{k} P(E \mid H_n) \cdot P(H_n)}$$

- **Bayes' theorem then states that**

$$P(H_i \backslash E) = \frac{P(E \mid H_i) \cdot P(H_i)}{\sum_{n=1}^{k} P(E \mid H_n) \cdot P(H_n)}$$

- **Further, if we add a new piece of evidence, e, then**

$$P(H \backslash E, e) = P(H \mid E) \cdot \frac{P(e \mid E, H)}{P(e \mid E)}$$

Bayes Theorem

- **The notion of conditional probability : P(H\E)**

Let :

$P(H_i \backslash E)$ = the probability that hypothesis H_i is true given evidence E
$P(E \backslash H_i)$ = the probability that we will observe evidence E given that hypothesis i is true
$P(H_i)$ = the *a priori* probability that hypothesis i is true in the absence of any specific evidence. These probabilities are called prior probabilities or *prlors*.
k = the number of possible hypotheses

Bayes' theorem then states that

$$P(H_i \backslash E) = \frac{P(E \mid H_i) \cdot P(H_i)}{\sum_{n=1}^{k} P(E \mid H_n) \cdot P(H_n)}$$

- **Bayes' theorem then states that**

$$P(H_i \backslash E) = \frac{P(E \mid H_i) \cdot P(H_i)}{\sum_{n=1}^{k} P(E \mid H_n) \cdot P(H_n)}$$

- **Further, if we add a new piece of evidence, e, then**

$$P(H \backslash E, e) = P(H \mid E) \cdot \frac{P(e \mid E, H)}{P(e \mid E)}$$

The Bayes' theorem is **intractable for several reasons** :

✓ The knowledge acquisition problem is insurmountable; **too many probabilities have to be provided.** And there are substantial evidence that people are very bad probability estimators.

✓ The **space** that would be required to **store all the probabilities is too large.**

✓ The **time required** to compute the probabilities is too large.

Even if we have the above mentioned problems the theorem is still very useful for the uncertain system. Various mechanisms to use it have been developed.

Some of them are:

1. Adding certainty factors to rules
2. Bayesian networks
3. Dempster-Shafer-Theory

Certainty Factors and Rule-based systems

- Probability-based reasoning adopted Bayes Theorem for handling uncertainty.
- But to apply Bayes theorem, there is a need to estimate a **priori and conditional** probabilities which are difficult to be calculated in many domains.
- For this one practical way of compromising on a pure Bayesian system is to adopt **certainty factors**.
- Certainty factors theory is a popular alternative to Bayesian reasoning.

Certainty factors theory and evidential reasoning

- A **certainty factor** (*cf*), a number to measure the expert's belief.
  - maximum value of the certainty factor is, say, +1.0 (definitely true)
  - minimum –1.0 (definitely false).
- For example, if the expert states that some evidence is almost certainly true, a *cf* value of 0.8 would be assigned to this evidence.

Certainty factor & Rule based systems

- Here we describe one practical approach of compromising a pure Bayesian system. The approach we discuss here is founded in "MYCIN" system, which attempts to recommend appropriate therapies for patients with bacterial infections.
- It interacts with physician to acquire clinical data it needs.
- It is an expert system, since it performs a task normally done by human expert.

Cont'd…

- MYCIN represents most of its diagnostic knowledge as a set of rules. Each rule is associated with it a "certainty factor".
- To understand how MYCIN exploits uncertain information, we need to answer two quesrions.

1. What do certainty factor means?
2. How does "MYCIN" combines estimates of certainty to produce final estimate of certainty of its conclusion?

- Associated with each rule is a certainty factor.
- **CF – a measure of the belief or disbelief on a conclusion in the presence of a set of evidence.**

Measures of Belief

- The values of MB(H, E) and MD(H, E) range between **0 and 1**.
- The strength of belief or disbelief in hypothesis H depends on the kind of evidence E observed.
- Some facts may increase the strength of belief, but some increase the strength of disbelief.

- $MB[h, e]$—a measure (between 0 and 1) of belief in hypothesis h given the evidence e. MB measures the extent to which the evidence supports the hypothesis. It is zero if the evidence fails to support the hypothesis.

- $MD[h,e]$—a measure (between 0 and 1) of disbelief in hypothesis h given the evidence e. MD measures the extent to which the evidence supports the negation of the hypothesis. It is zero if the evidence supports the hypothesis.

$$CF[h, e] = MB[h, e] - MD[h, e]$$

- Two important things, **one is evidence e and another the hypothesis h**
  - i.e. it gives a conclusion.
  - It is possible for related evidences giving rise to a hypothesis or conclusion.
- The values of the CF are determined the **domain expert** who creates the knowledge base.
- In expert system that involving multiple rules relating to the same conclusion, the rules must be structured in a way helping the user to enhance either the belief or disbelief.

Combining Uncertain Rules

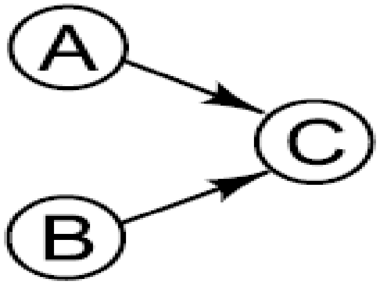
Goals for combining rules :

- Since the order in which evidence is collected is arbitrary, the combining functions should be commutative and associative.
- Until certainty is reached, additional confirming evidence should increase MB (and similarly for disconfirming evidence and MD).
- If uncertain inferences are chained together, then the result should be less certain than either of the inferences alone.

Combining Uncertain Rules

(a) (b) (c)

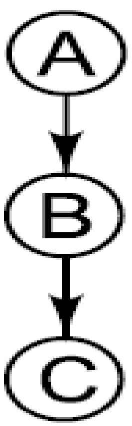
Combining Uncertain Rules



->several rules provide evidence that relates to a hypothesis

$A \wedge B$

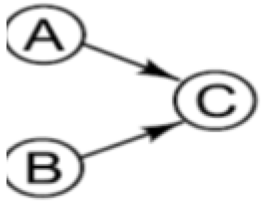
->Belief is a collection of several proportions taken together



->output of one rule provides input to another

Combining Two Pieces of Evidence

1:



$$MB[h, s_1 \wedge s_2] = \begin{cases} 0 & \text{if } MD[h, s_1 \wedge s_2] = 1 \\ MB[h, s_1] + MB[h, s_2] \cdot (1 - MB[h, s_1]) & \text{otherwise} \end{cases}$$

$$MD[h, s_1 \wedge s_2] = \begin{cases} 0 & \text{if } MB[h, s_1 \wedge s_2] = 1 \\ MD[h, s_1] + MD[h, s_2] \cdot (1 - MD[h, s_1]) & \text{otherwise} \end{cases}$$

Combining Two Pieces of Evidence

- **Interpretation of formula:**
- The measure of belief in *h* is 0 if *h* is disbelieved certainly. Otherwise, the measure of belief in *h* given two observations *s1* and *s2* is the measure of belief for given one observation plus increment for second observation.
- Suppose we make an initial observation that confirms our belief in *h* with MB=0.3. Then, MD[h,s1]=0 and CF[h,s1]=0.3. For second observation, MB[h,s2]=0.2.
- So, $MB[h,s1 \wedge s2]=0.3+(0.2*0.7)=0.44$

2: Compute certainty factor of combination of hypothesis

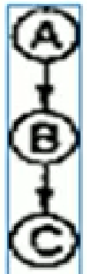
MB[h1^h2,e]=min(MB[h,e],MB[h2,e])

MB[h1vh2,e]=max(MB[h1,e],MB[h2,e])

MD[h1^h2,e]=max(MD[h1,e],MD[h2,e])

MD[h1vh2,e]=min(MD[h1,e],MD[h2,e])

3:



(c)

=>Next, consider the scenario in Fig©

-Rules are chained together with the result that the uncertain outcome of one rule must provide the input to another.

-It is solved institutions where the evidence is the outcome of an experiment or a laboratory test whose results are not completely accurate.

-The CF of the hypothesis must take both the strength with which evidence suggests the hypothesis and the level of confidence in the evidence.

-MYCIN provides a chaining rule.

MB[h,s]=MB`[h,s].max(0,CF[s,e]) where,

MB`[h,s]-> measure of belief in h given that it is absolutely sure of the validity of s.

e -> evidence that leads to believe in s.

Example:

IF **there is enough fuel in the vehicle**

AND **the ignition system is working correctly** **AND** **the vehicle doesn't start**

THEN **fault lies in the fuel flow CF=0.75**

⇒ The CF value of 0.75 is assigned by the domain expert on the assumption that every IF part is known with 100% certainty.

⇒ But if the IF part is not known with 100% certainty, then the above formula should be used.

⇒ The above methods are used by MYCIN for certainty factor calculations.

An Example of Combining Two Observations

$$MB[h, s_1] = 0.3$$

$$MD[h, s_1] = 0.0$$

$$CF[h, s_1] = 0.3$$

$$MB[h, s_2] = 0.2$$

$$MB[h, s_1 \wedge s_2] = 0.3 + 0.2 \cdot 0.7$$
$$= 0.44$$

$$MD[h, s_1 \wedge s_2] = 0.0$$

$$CF[h, s_1 \wedge s_2] = 0.44$$

Which of the following statements define the certainty factor accurately?

A. The certainty factor is same as the probability of any event

B. The Certainty Factor (CF) is a numeric value that tells us about how likely an event or a statement is supposed to be tru*e*

C. The Certainty Factor (CF) is a numeric value that tells us about how certain we are about performing a particular task

D. None of the above

"The Certainty Factor (CF) is a numeric value which tells us about how likely an event or a statement is supposed to be true."

What is the range of this numeric value, i.e. Certainty Factor?

A. Between 0 to 1 (Both inclusive)

B. Between 0 to 1 (Both exclusive)

C. Between -1 to +1

D. None of the above

- MYCIN works because,

  ->each CF in a MYCIN rule represents the contribution of an individual rule to MYCIN's belief in a hypothesis.

  ->It represents a conditional probability p(HIE).

  ->But in pure Bayesian systems, it is considered only with relevant evidence.

  ->If there are other evidences then joint probabilities need to be considered.

- But in MYCIN all combinations make the assumption that all rules are independent.

  ->The burden of guaranteeing independence is of rule writer.

The Definition of Certainty Factors

- **Original definitions :**

$$MB[h,\, e] = \begin{cases} 1 & \text{if } P(h) = 1 \\ \dfrac{\max[P(h \mid e),\, P(h)] - P(h)}{1 - P(h)} & \text{otherwise} \end{cases}$$

- **Similarly, the MD is the proportionate decrease in belief in h as a result of e:**

$$MD[h,\, e] = \begin{cases} 1 & \text{if } P(h) = 0 \\ \dfrac{\min[P(h \mid e),\, P(h)] - P(h)}{-P(h)} & \text{otherwise} \end{cases}$$

- **But this definition is incompatible with Bayesian conditional probability. The following, slightly revised one is not :**

$$MB[h,\, e] = \begin{cases} 1 & \text{if } P(h) = 1 \\ \dfrac{\max[P(h \mid e),\, P(h)] - P(h)}{(1 - P(h) \cdot P(h \mid e)} & \text{otherwise} \end{cases}$$

What if the Observations are not Independent

But in MYCIN all combinations make the assumption that all rules are independent.

- **Scenario (a) :**

Reconsider a rule with three antecedents and a *CF* of 0.7. Suppose that if there were three separate rules, each would have had a *CF* of 0.6 In other words, they are not independent. Then, using the combining rules, the total would be:

$$\begin{aligned} MB[h, s \wedge s_2] &= 0.6 + (0.6 \cdot 0.4) \\ &= 0.84 \\ MB[h, (s_1 \wedge s_2) \wedge s_3] &= 0.84 + (0.6 \cdot 0.16) \\ &= 0.936 \end{aligned}$$

- **This is very different than 0.7.**

What if the Observations are not Independent

- **Scenario (c) :**

  Events :

  S: sprinkler was on last night
  W: grass is wet
  R: it rained last night

We can write MYCIN-style rules that describe predictive relationships among these three events:

```
If:  the sprinkler was on last night
then there is suggestive evidence (0.9) that
     the grass will be wet this morning
```

Taken alone, this rule may accurately describe the world. But now consider a second rule:

```
If:  the grass is wet this morning
then there is suggestive evidence (0.8) that
     it rained last night
```

Taken alone, this rule makes sense when rain is the most common source of water on the grass. But if the two rules are applied together, using MYCIN's rule for chaining, we get

$MB[W,S] = 0.9$ {sprinkler suggests wet}
$MB[R,W] = 0.8 \cdot 0.9 = 0.72$ {wet suggests rains}

- **So Sprinkler made us believe rain.**

It define probabilistic independencies and dependencies among variable in the network.

It is a probabilistic graphical model which represent a set of variables and their conditional dependencies using a **Directed Acyclic Graph (DAG)**

It consist of :

(a)DAG

(b)Table of conditional probabilities

Node: Corresponds to a random variable

Arc/Directed Arrows: represent casual relationship or conditional probabilities among random variables.

Each node corresponds to the random variables, and a variable can be continuous or discrete.

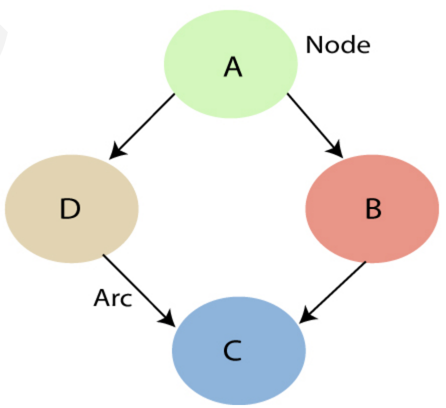
Arc or directed arrows represent the causal relationship or conditional probabilities between random variables. These directed links or arrows connect the pair of nodes in the graph.

These links represent that one node directly influence the other node, and if there is no directed link that means that nodes are independent with each other

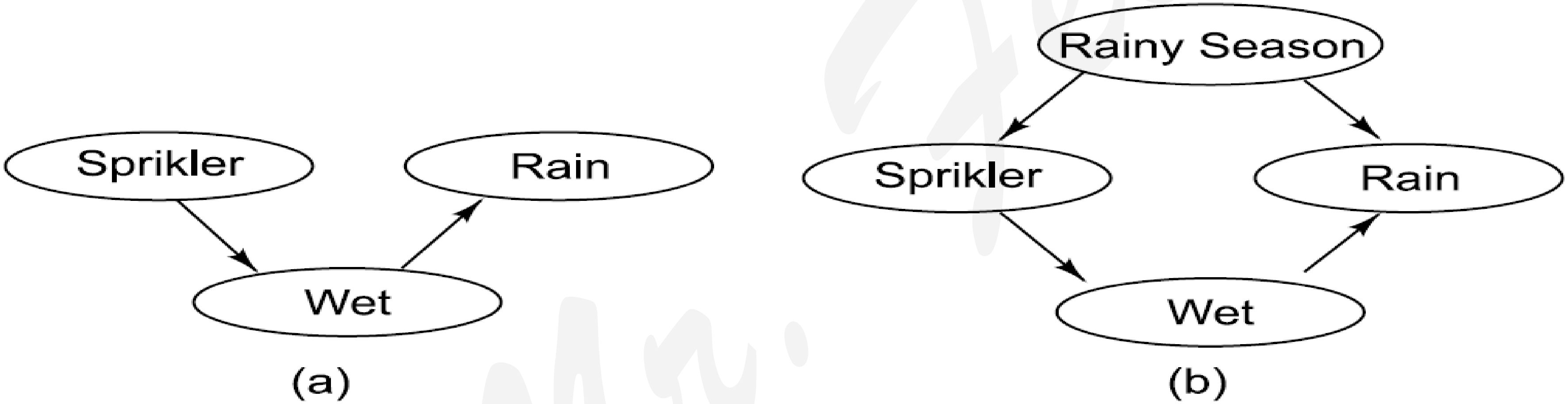
In the above diagram, A, B, C, and D are random variables represented by the nodes of the network graph.

If we are considering node B, which is connected with node A by a directed arrow, then node A is called the parent of Node B.

Node C is independent of node A.



Bayesian Networks : Representing Causality Uniformly



Conditional Probabilities for Bayesian Network

Attribute	*Probability*
$p(Wet \backslash Sprinkler, Rain)$	0.95
$P(Wet \backslash Sprinkler, \neg Rain)$	0.9
$p\{Wet \backslash \neg Sprinkler, Rain)$	0.8
$p(Wet \backslash \neg Sprinkler, \neg Rain)$	0.1
$p(Sprinkler \backslash RainySeason)$	0.0
$p(Sprinkler \backslash \neg RainySeason)$	1.0
$p(Rain \backslash RainySeason)$	0.9
$p(Rain \backslash \neg RainySeason)$	0.1
$p(RainySeason)$	0.5

Joint probability distribution:

If we have variables x1, x2, x3,....., xn, then the probabilities of a different combination of x1, x2, x3.. xn, are known as Joint probability distribution.

$\mathbf{P[x_1, x_2, x_3, \ldots, x_n]}$, it can be written as the following way in terms of the joint probability distribution.

$\mathbf{= P[x_1 | x_2, x_3, \ldots, x_n] P[x_2, x_3, \ldots, x_n]}$

$\mathbf{= P[x_1 | x_2, x_3, \ldots, x_n] P[x_2 | x_3, \ldots, x_n] \ldots P[x_{n-1} | x_n] P[x_n].}$

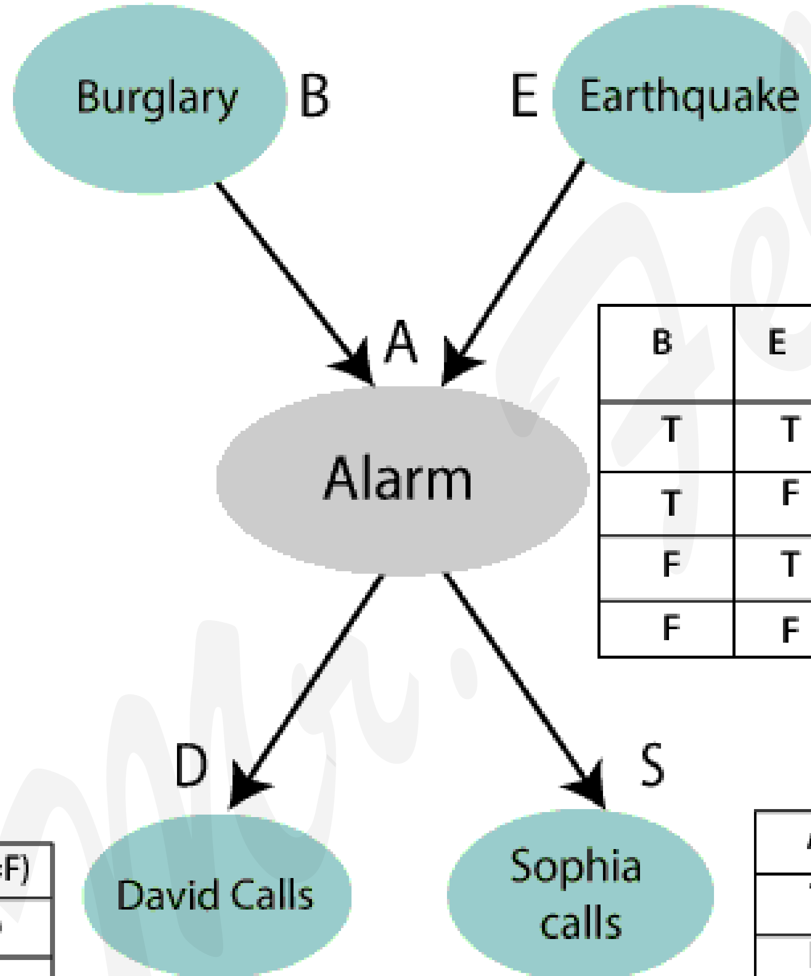
In general for each variable Xi, we can write the equation as:

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

Example: Harry installed a new burglar alarm at his home to detect burglary. The alarm reliably responds at detecting a burglary but also responds for minor earthquakes. Harry has two neighbors David and Sophia, who have taken a responsibility to inform Harry at work when they hear the alarm. David always calls Harry when he hears the alarm, but sometimes he got confused with the phone ringing and calls at that time too. On the other hand, Sophia likes to listen to high music, so sometimes she misses to hear the alarm. Here we would like to compute the probability of Burglary Alarm.

Problem:

Calculate the probability that alarm has sounded, but there is neither a burglary, nor an earthquake occurred, and David and Sophia both called the Harry.



T	0.002
F	0.998

T	0.001
F	0.999

B	E	P(A=T)	P(A=F)
T	T	0.94	0.06
T	F	0.95	0.04
F	T	0.69	0.69
F	F	0.999	0.999

A	P (D=T)	P (D=F)
T	0.91	0.09
F	0.05	0.95

A	P (S=T)	P (S=F)
T	0.75	0.25
F	0.02	0.98

List of all events occurring in this network:

Burglary (B) **Earthquake(E)** **Alarm(A)**

David Calls(D) **Sophia calls(S)**

We can write the events of problem statement in the form of probability: **P[D, S, A, B, E]**, can rewrite the above probability statement using joint probability distribution:

P[D, S, A, B, E]= P[D | S, A, B, E]. P[S, A, B, E]

=P[D | S, A, B, E]. P[S | A, B, E]. P[A, B, E]

= P [D| A]. P [S| A, B, E]. P[A, B, E]

= P[D | A]. P[S | A]. P[A| B, E]. P[B, E]

= P[D | A]. P[S | A]. P[A| B, E]. P[B |E]. P[E]

Problem:
Calculate the probability that alarm has sounded, but there is neither a burglary, nor an earthquake occurred, and David and Sophia both called the Harry.

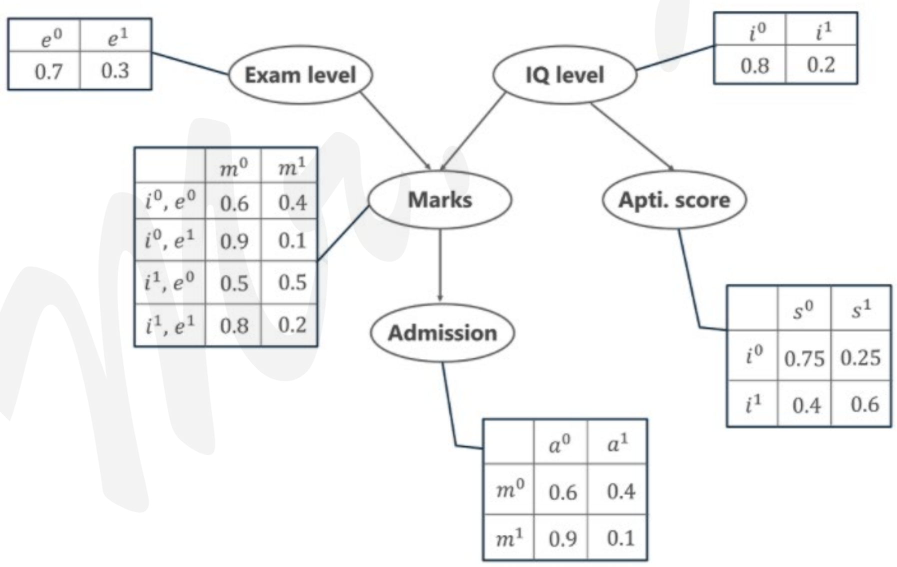
P(S, D, A, ¬B, ¬E) = P (S|A) *P (D|A)*P (A|¬B ^ ¬E) *P (¬B) *P (¬E).

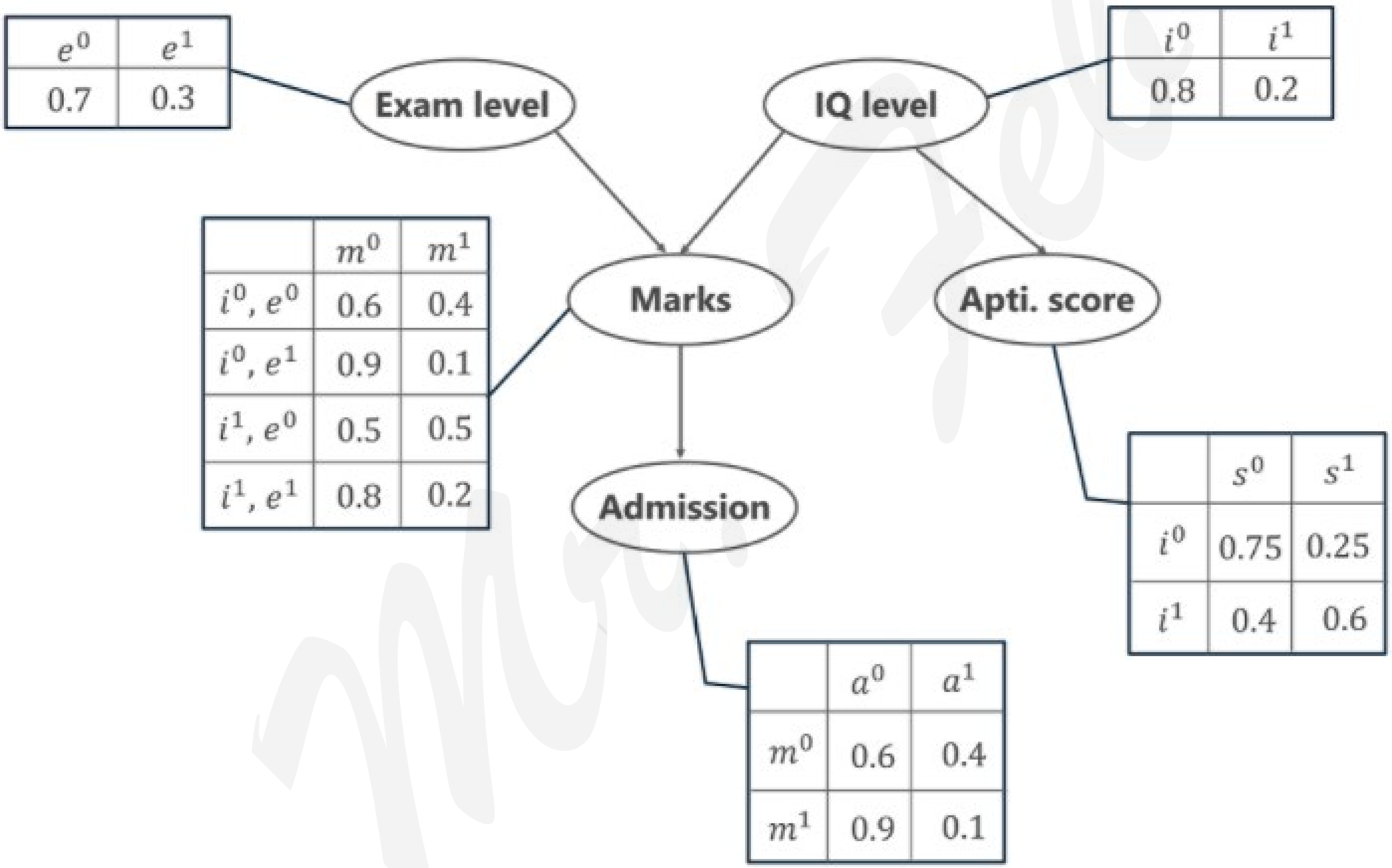
= 0.75* 0.91* 0.001* 0.998*0.999

= 0.00068045.

Exam Level (***e***)– This discrete variable denotes the difficulty of the exam and has two values (0 for easy and 1 for difficult)
IQ Level (***i***) – This represents the Intelligence Quotient level of the student and is also discrete in nature having two values (0 for low and 1 for high)

Additionally, the IQ level of the student also leads us to another variable, which is the Aptitude Score of the student (***s***). Now, with marks the student has scored, he can secure admission to a particular university. The probability distribution for getting admitted (***a***) to a university is also given below.





e^0	e^1
0.7	0.3

i^0	i^1
0.8	0.2

	m^0	m^1
i^0, e^0	0.6	0.4
i^0, e^1	0.9	0.1
i^1, e^0	0.5	0.5
i^1, e^1	0.8	0.2

	s^0	s^1
i^0	0.75	0.25
i^1	0.4	0.6

	a^0	a^1
m^0	0.6	0.4
m^1	0.9	0.1

Exam Level (e)
IQ Level (i)
Aptitude Score (s)
Marks (m)
Admission (a)

These five variables are represented in the form of a Directed Acyclic Graph (DAG) in a Bayesian Network format with their Conditional Probability tables. Now, to calculate the Joint Probability Distribution of the 5 variables the formula is given by,

P[a, m, i, e, s]= P(a | m) . P(m | i, e) . P(i) . P(e) . P(s | i)

Calculate the probability that in spite of the exam level being difficult, the student having a low IQ level and a low Aptitude Score, manages to pass the exam and secure admission to the university

From the above word problem statement, the Joint Probability Distribution can be written as below,

P[a=1, m=1, i=0, e=1, s=0]

From the above Conditional Probability tables, the values for the given conditions are fed to the formula and is calculated as below.

P[a=1, m=1, i=0, e=0, s=0] = P(a=1 | m=1) . P(m=1 | i=0, e=1) . P(i=0) . P(e=1) . P(s=0 | i=0)

= 0.1 * 0.1 * 0.8 * 0.3 * 0.75

= **0.0018**

In another case, calculate the probability that the student has a High IQ level and Aptitude Score, the exam being easy yet fails to pass and does not secure admission to the university.

The formula for the JPD is given by

P[a=0, m=0, i=1, e=0, s=1]

Thus,

P[a=0, m=0, i=1, e=0, s=1]= P(a=0 | m=0) . P(m=0 | i=1, e=0) . P(i=1) . P(e=0) . P(s=1 | i=1)

= 0.6 * 0.5 * 0.2 * 0.7 * 0.6

= **0.0252**

Where does the bayes rule can be used?

a) Solving queries

b) Increasing complexity

c) Decreasing complexity

d) Answering probabilistic que*ry*

What does the bayesian network provides?

a) Complete description of the *domain*

b) Partial description of the domain

c) Complete description of the problem

d) None of the mentioned

How the bayesian network can be used to answer any query?

a) Full distribution

b) Joint distributio*n*

c) Partial distribution

d) All of the mentioned

Bayesian Belief Network is also known as ?

A. belief network
B. decision network
C. Bayesian model
D. All of the above

The Bayesian network graph does not contain any cyclic graph. Hence, it is known as a

A. DCG
B. DAG
C. CAG
D. SAG

In a Bayesian network variable is?

A. continuous
B. discrete
C. Both A and B
D. None of the above

If we have variables x1, x2, x3,....., xn, then the probabilities of a different combination of x1, x2, x3.. xn, are known as?

A. Table of conditional probabilities
B. Causal Component
C. Actual numbers
D. Joint probability distributio*n*

Bayesian Network consist of ?	
a.	2 components
b.	3 components
c.	4 components
d.	5 components

The nodes and links form the structure of the Bayesian network, and we call this the ?	
a.	structural specificati*on*
b.	multi-variable nodes
c.	Conditional Linear Gaussian distributions
d.	None of the above

Dempster – Shafer Theory (DST)

- DST is a mathematical theory of evidence based on belief functions and plausible reasoning. It is used to combine separate pieces of information (evidence) to calculate the probability of an event.
- DST offers an alternative to traditional probabilistic theory for the mathematical representation of uncertainty.

Dempster – Shafer Theory (DST)

Example :

- Let A represent the proposition "Moore is attractive". Then the axioms of probability insist that $P(A) + P(\neg A) = 1$.
- Now suppose that Andrew does not even know who "Moore" is, then
- We cannot say that Andrew believes the proposition if he has no idea what it means.
- Also, it is not fair to say that he disbelieves the proposition.
- It would therefore be meaningful to denote Andrew's belief B of $B(A)$ and $B(\neg A)$ as both being 0.
- Certainty factors do not allow this.

Dempster-Shafer Model

The idea is to allocate a number between 0 and 1 to indicate a degree of belief on a proposal as in the probability framework.

However, it is not considered a probability but a belief mass. The distribution of masses is called basic belief assignment.

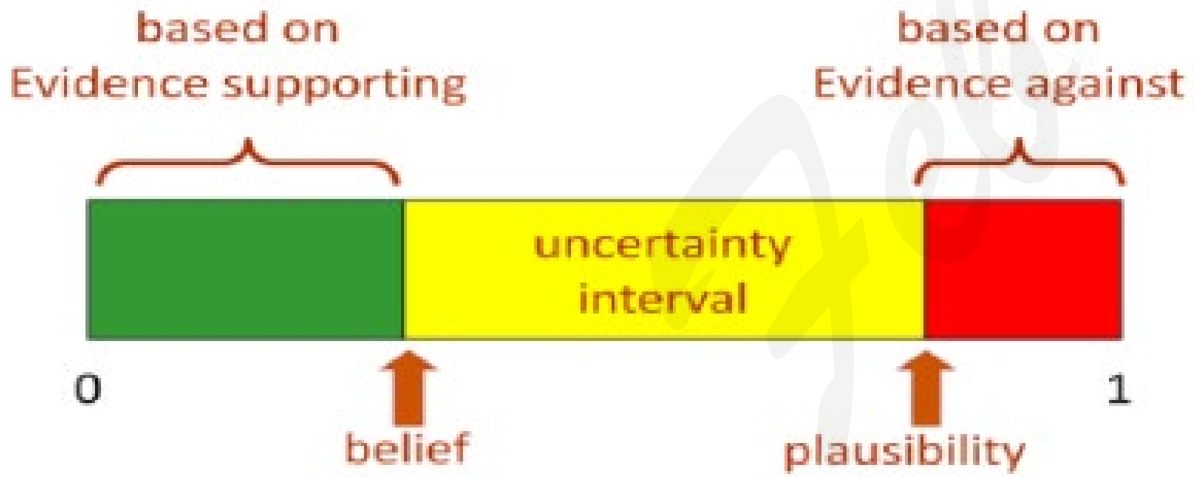
In other words, in this formalism a degree of belief (referred as mass) is represented as a belief function rather than a Bayesian probability distribution.

Belief and Plausibility

Shafer's framework allows for belief about propositions to be represented as intervals, bounded by two values, belief (or support) and plausibility:

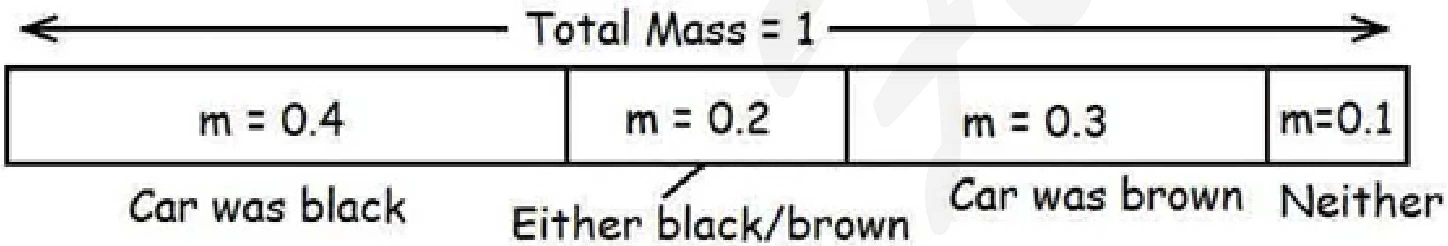
belief $\leq$ plausibility

Belief in a hypothesis is constituted by the sum of the masses of all sets enclosed by it (i.e. the sum of the masses of all subsets of the hypothesis). It is the amount of belief that directly supports a given hypothesis at least in part, forming a lower bound.
Plausibility is 1 minus the sum of the masses of all sets whose intersection with the hypothesis is empty. It is an upper bound on the possibility that the hypothesis could possibly happen, up to that value, because there is only so much evidence that contradicts that hypothesis.


based on
Evidence supporting
based on
Evidence against
uncertainty
interval
0
1
belief
plausibility

Belief and Plausibility

Witness: "I'm fairly sure that the car was either brown or black. Probably black, though it could have been brown. I could be wrong though."



Belief: Bel(Black) = 0.4 Bel(Brown) = 0.3

Bel(not Black) = 0.3 + 0.1 = 0.4 Bel(not Brown) = 0.5

Plausibility: Pl(Black) = 0.4 + 0.2 = 0.6 = 1 - Bel(not Black)

Pl(Brown) = 0.2 + 0.3 = 0.5 = 1 - Bel(not Brown)

In general: $Pl(A) = 1 - Bel(\text{not } A)$

Belief in A is the sum of the mass values that form subsets of A.

The car was black:	0.25
The car was black with chrome trims:	0.05
The car was black with a powerful engine:	0.1
The car was brown:	0.3
The car was either black or brown:	0.2
The car was some other colour:	0.1

Subset of "black car"

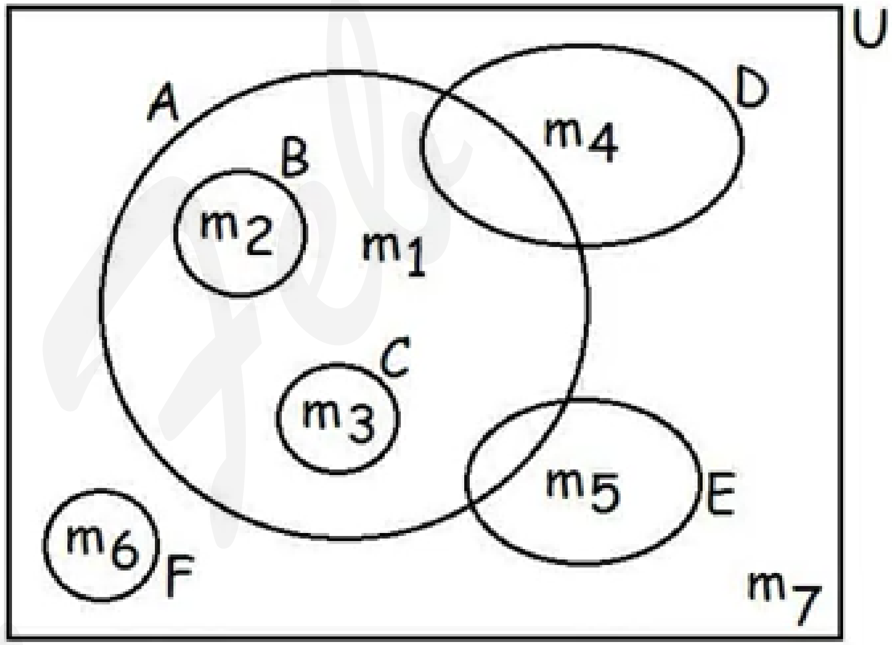
Compatible with black car (plausibility)

$$Bel(A) = \sum_{X \subseteq A} m_X$$

$$= m_1 + m_2 + m_3$$

$$Pl(A) = \sum_{X \cap A \neq \emptyset} m_X$$

$$= m_1 + m_2 + m_3 + m_4 + m_5$$



$Pl(A) \geq Bel(A)$

All masses add to 1.

Hypothesis	Mass	Belief	Plausibility
Null (neither real nor fake)	0	0	0
Real	0.2	0.2	0.5
Fake	0.5	0.5	0.8
Either (real or fake)	0.3	1.0	1.0

Example: Belief assignment

Suppose a system has five members, say five independent states, and exactly one of which is actual. If the original set is called S, | S | = 5, then the set of all subsets (the power set) is called 2^S.

If each possible subset as a binary vector (describing any member is present or not by writing 1 or 0), then 32 subsets are possible, ranging from the empty subset (0, 0, 0, 0, 0) to the "everything" subset (1, 1, 1, 1, 1).

The "empty" subset represents a "contradiction", which is not true in any state, and is thus assigned a mass of one ;

The remaining masses are normalized so that their total is 1.

Frames of discernment

Given a set of possible elements, called **environment**,

$$\Theta = \{\theta_1, \theta_2, \ldots, \theta_n\}$$

that are mutually exclusive and exhaustive.

- The environment is the set of objects that are of interest to us.
- For example,

$$\Theta = \{\text{airline, bomber, fighter}\}$$
$$\Theta = \{\text{red, green, blue, orange, yellow}\}$$

One way of thinking about Θ is in terms of questions and answers. Suppose

$$\Theta = \{\text{airline, bomber, fighter}\}$$

and the questions is, “what are the military aircraft?”. The answer is the subset of Θ

$$\{\theta_2, \theta_3\} = \{\text{bomber, fighter}\}$$

Dempster -Shafer Theory

- **We consider the interval :**

$$[Belief,\ Plausibility]$$

- **Plausibility (Pl) is defined to be :**

$$Pl(s) = 1 - Bel(\neg s)$$

- **Let the frame of discernment be anΘ, austive, mutually exclusive set of hypothesis.**

- **Let m be a probability density function.**

- **We define the combination m_3 of m_1 and m_2 to be**

$$m_3(Z) = \frac{\sum_{X \cap Y = Z} m_1(X) \cdot m_2(Y)}{1 - \sum_{X \cap Y = \phi} m_1(X) \cdot m_2(Y)}$$

- Belief Interval [Bel(A),Pl(A)], confidence in A
  - **Interval width is good aid in deciding when you need more evidence**
- [0,1] no belief in support of proposition
  - **total ignorance**
- [0,0] belief the proposition is false
- [1,1] belief the proposition is true
- [.3,1] partial belief in the proposition is true
- [0,.8] partial disbelief in the proposition is true
- [.2,.7] belief from evidence both for and against propostion

Example: While assessing the grades of the class of 100 students, two of the class teachers responded the overall result as follow. First teacher assessed that 40 students will get A and 20 students will get B grade amongst the total 60 students he interviewed. Whereas second teacher stated that 30 students will get A grade and 30 students will get either A or B amongst the 60 students he took the interview. Combining both evidences to find the resultant evidence, we will do following calculations. Here frame of discernment $\theta = \{A, B\}$ and Power set $2^{\theta} = \{\emptyset, A, B, (A, B)\}$,

Evidence (1) = Ev_1	Evidence (2) = Ev_2
$m_1(A) = 0.4$	$m_2(A) = 0.3$
$m_1(B) = 0.2$	$m_2(A, B) = 0.3$
$m_1(\theta) = 0.4$	$m_2(\theta) = 0.4$

Belief function (bel):

$Bel_1(A) = m_1(A) = 0.4$	$Bel_2(A) = m_2(A) = 0.3$
$Bel_1(B) = m_1(B) = 0.2$	$Bel_2(A,B) = m_2(A) + m_2(B) + m_2(A,B) = 0.3 + 0 + 0.3 = 0.6$
$Bel_1(\theta) = m_1(A) + m_1(B) + m_1(\Theta) = 0.4 + 0.2 + 0.4 = 1.0$	$Bel_2(\theta) = m_2(A) + m_2(B) + m_2(A,B) + m_2(\Theta) = 0.3 + 0 + 0.3 + 0.4 = 1.0$

$Pl_1(A) = m_1(A) + m_1(\theta) = 0.4 + 0.4$ $= 0.8$	$Pl_2(A) = m_2(A) + m_2(\theta) = 0.3 + 0.4$ $= 0.7$
$Pl_1(B) = m_1(B) + m_1(\theta) = 0.2 + 0.4$ $= 0.6$	$Pl_1(A, B) = m_2(A) + m_2(A, B)$ $+ m_2(\theta)$ $= 0.3 + 0.3 + +0.4 = 1.0$
$Pl_1(\theta) = m_1(A) + m_1(B) + m_1(\theta)$ $= 0.4 + 0.2 + 0.4 = 1.0$	$Pl_2(\theta) = m_2(A) + m_2(A, B) + m_2(\theta)$ $= 0.3 + 0.3 + 0.4 = 1.0$

Evidences	$m_1(A)=0.4$	$m_1(B)=0.2$	$m_1(\theta)=0.4$
$m_2(A)=0.3$	$m_{1\text{-}2}(A)$ 0.12	$m_{1\text{-}2}(\emptyset)$ 0.06	$m_{1\text{-}2}(A)$ 0.12
$m_2(A,B)=0.3$	$m_{1\text{-}2}(A)$ 0.12	$m_{1\text{-}2}(B)$ 0.06	$m_{1\text{-}2}(A,B)$ 0.12
$m_2(\theta)=0.4$	$m_{1\text{-}2}(A)$ 0.16	$m_{1\text{-}2}(B)$ 0.08	$m_{1\text{-}2}(\theta)$ 0.16

$k = 0.06$ and $1 - k = 0.94$

$$m_{1-2}(A) = \frac{0.12 + 0.12 + 0.12 + 0.16}{0.94} = 0.553$$

$$m_{1-2}(A, B) = \frac{0.12}{0.94} = 0.128$$

$$m_{1-2}(B) = \frac{0.06 + 0.08}{0.94} = 0.149$$

$$m_{1-2}(\theta) = \frac{0.16}{0.94} = 0.170$$

$\text{Bel}_{1\text{-}2}(A) = m_{1\text{-}2}(A) = 0.553$

$\text{Bel}_{1\text{-}2}(B) = m_{1\text{-}2}(B) = 0.149$

$\text{Bel}_{1-2}(A, B) = m_{1-2}(A) + m_{1-2}(B) + m_{1-2}(A, B) = 0.553 + 0.149 + 0.128 = 0.83$

$\text{Bel}_{1-2}(\theta) = m_{1-2}(A) + m_{1-2}(B) + m_{1-2}(A, B) + m_{1-2}(\theta) = 0.553 + 0.149 + 0.128 + 0.170 = 1$

Dempster - Shafer Example

- Let Θ be :

  All : allergy

  Flu : flu

  Cold : cold

  Pneu : pneumonia

- When we begin, with no information m is :

  $\{\Theta\}$ (1.0)

- Suppose m_1 corresponds to our belief after observing fever.

  $\{Flu, Cold, Pneu\}$ (0.6)

  $\{\Theta\}$ (0.4)

- Suppose m_2 corresponds to our belief after observing a runny nose.

  $\{All,\ Flu, Cold\}$ (0.8)

  Θ (0.2)

	$\{A, F, C\}$ (0.8) Θ (0.2)
$\{F, C, P\}$ (0.6) Θ (0.4)	

Dempster - Shafer Example (Cont'd)

- Then we can combine m_1 and m_2 :

	$\{A, F, C\}$	(0.8)	Θ	(0.2)
$\{F, C, P\}$ (0.6)	$\{F, C\}$	(0.48)	$\{F, C, P\}$	(0.12)
Θ (0.4)	$\{A, F, C\}$	(0.32)	Θ	(0.08)

- So we produce a new, combined m_3 :

$\{Flu, Cold\}$	(0.48)
$\{All, Flu, Cold\}$	(0.32)
$\{Flu, Cold, Pneu\}$	(0.12)
Θ	(0.08)

- Suppose m_4 corresponds to our belief after that the problem goes away on trips :

$\{All\}$	(0.9)
Θ	(0.1)

So we produce a new, combined m_3 :

$\{Flu, Cold\}$	(0.48)
$\{All, Flu, Cold\}$	(0.32)
$\{Flu, Cold, Pneu\}$	(0.12)
Θ	(0.08)

Suppose m_4 corresponds to our belief after that the problem goes away on trips :

$\{All\}$	(0.9)
Θ	(0.1)

	$\{A\}$	(0.9)	Θ	(0.1)
$\{F, C\}$				
$\{A, F, C\}$				
$\{F, C, P\}$				
Θ				

Dempster - Shafer Example (Cont'd)

		$\{A\}$	(0.9)	Θ	(0.1)
$\{F, C\}$	(0.48)	ϕ	(0.432)	$\{F, C\}$	(0.048)
$\{A, F, C\}$	(0.32)	$\{A\}$	(0.288)	$\{A, F, C\}$	(0.032)
$\{F, C, P\}$	(0.12)	ϕ	(0.108)	$\{F, C, P\}$	(0.012)
Θ	(0.08)	$\{A\}$	(0.072)	Θ	(0.008)

ϕ

Dempster - Shafer Example (Cont'd)

- Then we can combine m_1 and m_2 :

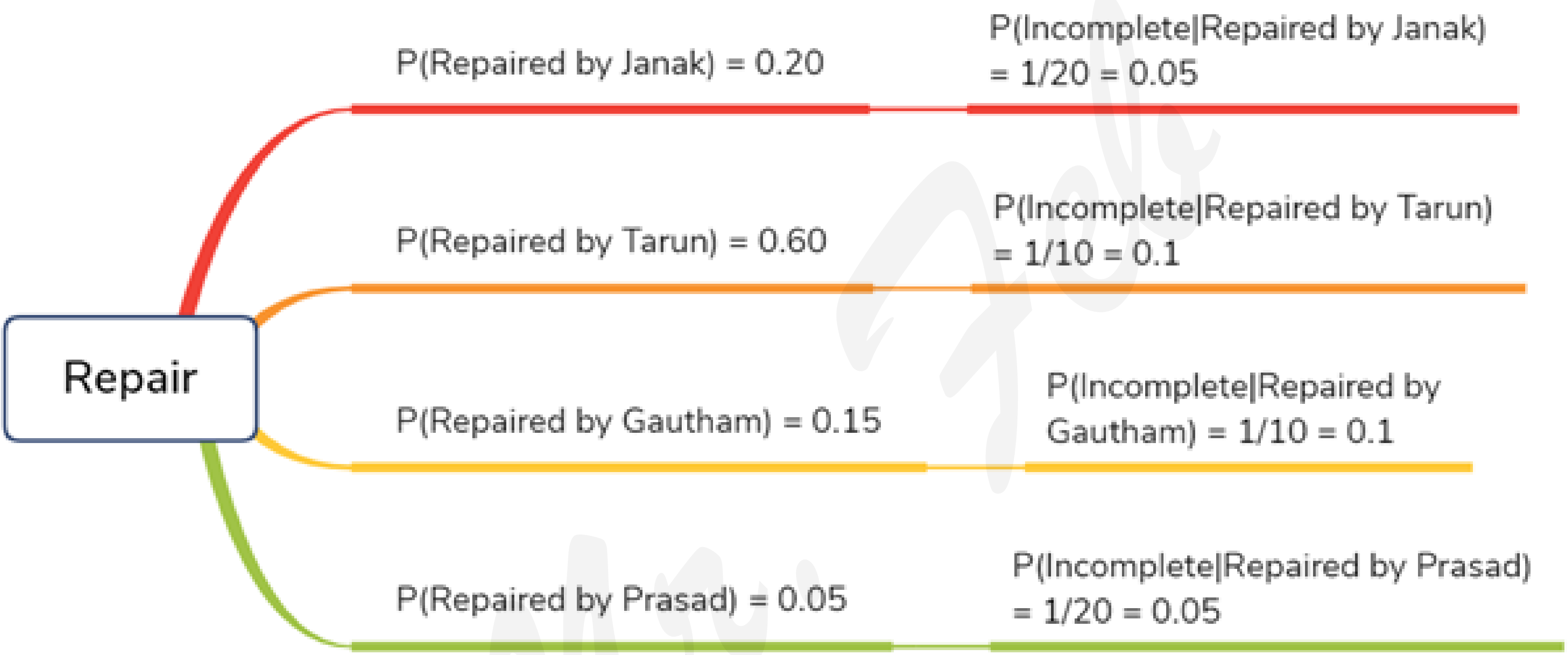
		$\{A\}$	(0.9)	Θ	(0.1)
$\{F, C\}$	(0.48)	ϕ	(0.432)	$\{F, C\}$	(0.048)
$\{A, F, C\}$	(0.32)	$\{A\}$	(0.288)	$\{A, F, C\}$	(0.032)
$\{F, C, P\}$	(0.12)	ϕ	(0.108)	$\{F, C, P\}$	(0.012)
Θ	(0.08)	$\{A\}$	(0.072)	Θ	(0.008)

- Normalizing to get rid of the belief of 0.54 associated with ϕ gives m_5 :

$\{Flu, Cold\}$	(0.104)
$\{All, Flu, Cold\}$	(0.0696)
$\{Flu, Cold, Pneu\}$	(0.026)
$\{All\}$	(0.782)
Θ	(0.017)

Example of Bayes theorem

Technicians regularly make repairs when breakdowns occur on an automated production line. Janak, who services 20% of the breakdowns, makes an incomplete repair 1 time in 20. Tarun, who services 60% of the breakdowns, makes an incomplete repair 1 time in 10. Gautham, who services 15% of the breakdowns, makes an incomplete repair 1 time in 10 and Prasad, who services 5% of the breakdowns, makes an incomplete repair 1 time in 20. For the next problem with the production line diagnosed as being due to an initial repair that was incomplete, what is the probability that this initial repair was made by Janak?


Repair
P(Repaired by Janak) = 0.20
P(Incomplete|Repaired by Janak) = 1/20 = 0.05
P(Repaired by Tarun) = 0.60
P(Incomplete|Repaired by Tarun) = 1/10 = 0.1
P(Repaired by Gautham) = 0.15
P(Incomplete|Repaired by Gautham) = 1/10 = 0.1
P(Repaired by Prasad) = 0.05
P(Incomplete|Repaired by Prasad) = 1/20 = 0.05

Reverse Probability
P(Janak | Incomplete) = ?

Solution:
P(Janak | Incomplete) = P(Repaired by Janak)*P(Incomplete | Repaired by Janak)/P(Incomplete)
P(Janak | Incomplete) = 0.20*0.05/[0.20*0.05 + 0.60*0.1 + 0.15*0.1 + 0.05*0.05]

Example

SpamAssassin works by having users train the system. It looks for patterns in the words in emails marked as spam by the user. For example, it may have learned that the word “free” appears in 20% of the emails marked as spam. Assuming 0.1% of non-spam mail includes the word “free” and 50% of all emails received by the user is spam, find the probability that a mail is a spam if the word “free” appears in it.

Data Given:

- $P(\text{Free} \mid \text{Spam}) = 0.20$
- $P(\text{Free} \mid \text{Non Spam}) = 0.001$
- $P(\text{Spam}) = 0.50 => P(\text{Non Spam}) = 0.50$
- $P(\text{Spam} \mid \text{Free}) = ?$

Data Given:

- $P(\text{Free} \mid \text{Spam}) = 0.20$
- $P(\text{Free} \mid \text{Non Spam}) = 0.001$
- $P(\text{Spam}) = 0.50 => P(\text{Non Spam}) = 0.50$
- $P(\text{Spam} \mid \text{Free}) = ?$

Using Bayes' Theorem:

- $P(\text{Spam} \mid \text{Free}) = P(\text{Spam}) * P(\text{Free} \mid \text{Spam}) / P(\text{Free})$
- $P(\text{Spam} \mid \text{Free}) = 0.50 * 0.20 / (0.50 * 0.20 + 0.50 * 0.001)$
- $P(\text{Spam} \mid \text{Free}) = 0.995$