M T W T F S S
Page No.:
Date:
YOUVA

Unit 5
Probabilistic Reasoning and Bayesian Belief Networks

* Probabilistic Reasoning (Can be asked for 5 marks)

Probability theory is used to discuss events, categories, and hypothesis about which there is no 100% certainity.

Ex: A → B meaning if A is true, B is true. If we are unsure whether A is true, then we cannot make use of this expression.

Consider Example for wheather forecast- "It is Sunny only 10% of the time, and rainy 70% of the time." We can use notation similar to that used for predicate calculus to express such statements:

$P(S) = 0.1$, $P(R) = 0.7$

The first of these statements says that the probability of S ("it is sunny) is 0.1. Second says that probability of R is 0.7.

* Probabilities are always expressed as real numbers between 0 and 1. *

Here, 0 means "definitely not" and probability of 1 means "definitely so". Hence, $P(S) = 1$. means that it is always sunny.

* Operators and notations used in propositional logic can also be used in probabilistic notation.

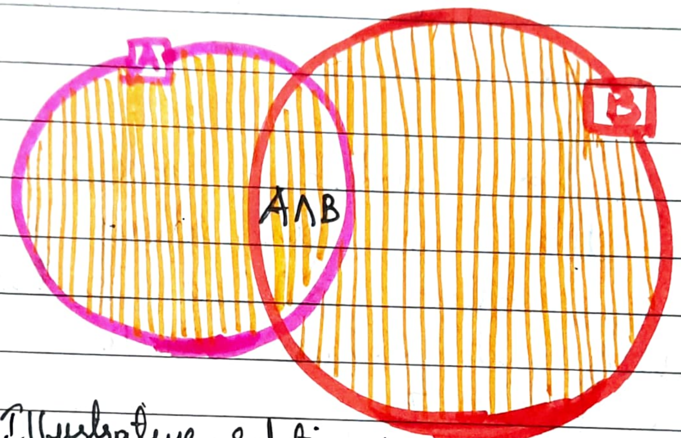
Ex: $P(\neg S)$ means "probability that it is not sunny"; $P(S \wedge R)$ means "the probability that it is both sunny and rainy".

Page No.:
Date:

$P(A \vee B)$ - means "probability that either A is true or B is true" is defined by following rule:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

This rule can be seen to be true by examining the Venn diagram shown below



Illustrating relation b/w $A \wedge B$ and $A \vee B$

The notation $P(B|A)$ can be read as "the probability of B, given A." This is known as Conditional probability - it is conditional on A.

* It states the probability that B is true given that we already know that A is true.

$P(B|A)$ is defined by foll. rule:

$$P(B|A) = \frac{P(B \wedge A)}{P(A)}$$ (this rule cannot be used in cases where $P(A) = 0$)

Ex: let us suppose that the likelihood that it is both sunny and rainy at the same time is 0·01. Then we can calculate the probability that it is rainy, given that it is sunny as follows:

$$P(R|S) = \frac{P(R \wedge S)}{P(S)}$$
$$= \frac{0.01}{0.1} = 0.1.$$

* NOTE : Probability that it is sunny given that it is overcast - $P(S|R)$ - is different from this: $0.01/0.7 = 0.14$; hence $P(A|B) \neq P(B|A)$.

* Joint Probability Distributions - (Can be asked for 5mks)

A joint probability distribution (also known as a joint) can be used to represent the probabilities of combined statements, such as $A \wedge B$. For example, the following table shows a joint probability distribution of two variables, A and B:

	A	¬A
B	0·11	0·09
¬B	0·63	0·17

This shows that $P(A \wedge B) = 0.11$, $P(A \wedge \neg B) = 0.63$. By summing these two values, we can find $P(A) = 0.11 + 0.63 = 0.74$. IIly $P(B) = 0.11 + 0.09 = 0.2$.

It (table) determines the probability of any logical combination of A and B.

Ex, $P(A \vee B) = 0.11 + 0.09 + 0.63 = 0.83$. Also, $P(\neg A \wedge \neg B) = 0.1$[7] and that $P(\neg A \wedge \neg B) = 1 - P(A \vee B) = 1 - 0.17 = 0.83$.

$|||^{ly}$, we can determine conditional probabilities, such as P(B|A) using foll. rule:

$$P(B|A) = \frac{P(B \wedge B)}{P(A)}$$

In this case, $P(B \wedge A) = 0.11$ and $P(A) = 0.11 + 0.63 = 0.74$, so $P(B|A) = 0.11/0.74 = 0.15$.

Bayes Theorem (10 marks with Example)

appeared*

Bayes theorem can be used to calculate the probability that a certain event will occur or that a certain proposition is true, given that we already know a related piece of information.

The theorem is stated as follows:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

P(B) is called the prior probability of B, P(B|A), as well as being called the conditional probability, is also known as the posterior probability of B.

Lets us examine how Bayes theorem is derived.

Consider, $P(A|B) = \frac{P(A \wedge B)}{P(B)}$ — Conditional probability.

We can deduce further –

$P(A \wedge B) = P(A|B) \cdot P(B)$ — ① called product rule

due to commutativity of $\wedge$, we can write:

$P(A \wedge B) = P(B|A) P(A)$ — ②

Hence, from ① & ② we can further deduce-

$$P(B/A)\,P(A) = P(A/B)\,P(B)$$

By Rearranging we get foll.

$$P(B/A) = \frac{P(A/B) \cdot P(B)}{P(A)} \rightarrow \text{Baye's Theorem.}$$

Ex: Medical Diagnosis.

- when one has a cold, one usually has a high temperature - say 80% of the time.
- We can use A to denote "I have a high temperature" and B to denote "I have a cold".

Therefore, we can write this statement of posterior probability $P(A/B)$ as:

$$P(A/B) = 0.8$$

Here, we could use A as a piece of evidence to help us prove or disprove the hypothesis, B, but it could work equally well the other way around.

- Now, we know that 1 in every 10,000 people has a cold, and that 1 in every 1000 people has a high temperature.

We can write these prior probabilities as:

$$P(A) = 0.001, \quad P(B) = 0.0001$$

Now, say if you have a high temperature. what is the likelihood that you have cold?

P.T.O

This can be calculated very easily by using Bayes' Theorem :

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

$$= \frac{0.8 * 0.0001}{0.001}$$

$$= \frac{0.00008}{0.001}$$

$\boxed{P(B|A) = 0.08}$ Hence, shows that having high temperature does not make it very likely that you have a cold.

(You can either use this Ex. or any other valid Example).

* Normalization
It is a process whereby the posterior probabilities of a pair of variables are divided by a fixed value to ensure that they sum to 1.
Consider following two Equations:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A)}$$

$$P(\neg B|A) = \frac{P(A|\neg B) \cdot P(\neg B)}{P(A)}$$

Given, that A is true, B must be either true or false, which means that $P(B/A) + P(\neg B/A) = 1$.

Hence, we can add the two Equations above to give

$$1 = \frac{P(A|B) \cdot P(B)}{P(A)} + \frac{P(A|\neg B) \cdot P(\neg B)}{P(A)}$$

$$\therefore, P(A) = P(A|B) \cdot P(B) + P(A|\neg B) \cdot P(\neg B)$$

We can replace $P(A)$ in the Equation for Bayes theorem, to give

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|\neg B) \cdot P(\neg B)}$$

Hence, it is possible to use Bayes theorem to obtain conditional probability $P(B|A)$ without needing to know or calculate $P(A)$, providing we can obtain $P(A|\neg B)$. [$P(\neg B)$ is simply $1 - P(B)$].

We can write -

$$P(B|A) = \alpha \cdot P(A|B) \cdot P(B)$$

where, α represents normalizing constant:

$$\alpha = \frac{1}{P(A|B) \cdot P(B) + P(A|\neg B) \cdot P(\neg B)}$$

Let us take previous Example of medical diagnosis $P(A) = 0.001$, $P(B) = 0.0001$, $P(A|B) = 0.8$.
Suppose that $P(A|\neg B) = 0.00099$. (Conditional probability that likelihood that a person will have a high temperature if she does not have a cold ($\neg B$).
Lets Calculate $P(B|A)$:

$$P(B|A) = \frac{P(A|B) \cdot P(B)}{P(A|B) \cdot P(B) + P(A|\neg B) \cdot \underbrace{P(\neg B)}_{(1-P(B) = 0.9999}} = \frac{0.8 * 0.0001}{0.8 * 0.001 + 0.00099 * 0.99}$$

$$= \frac{0.00008}{0.001069901} = 0.07$$

|||ly we can calculate $P(\neg B|A)$:

$$P(\neg B|A) = \frac{P(A|\neg B) \cdot P(\neg B)}{P(A|\neg B) \cdot P(\neg B) + P(A|B) \cdot P(B)}$$

$$= \frac{0.00099 * 0.9999}{0.00099 * 0.9999 + 0.8 + 0.0001}$$

$$= \frac{0.000989901}{0.001069901}$$

$P(\neg B|A) = 0.925$.

(* Either derivation or problems may be asked for 5-6 mks).

→ * Net result of $P(B|A) + P(\neg B|A) = 1$ (should be always 1).

$\therefore \boxed{0.075 + 0.925 = 1}$

* Bayesian Belief Networks. (with Example 10 mks)

appeared

Concept of dependence is very important in probability theory. Events A & B, are independent if the likelihood of occurrence of A is entirely unrelated to whether or not B occurs.

→ For Ex: tossing of two coins. Here both coins are independent of each other (in terms of probability) and neither one depends on the other.

→ If A & B are independent, then probability that A and B will both occur can be calculated as:

$$P(A \land B) = P(A) \cdot P(B).$$

We know that this Equation does not hold if A depends on B because we have already seen the foll. Equation:

$$P(B|A) = \frac{P(B \wedge A)}{P(A)}$$

By comparing these two Equations we can see that A and B are independent if $P(B|A) = P(B)$. Means, likelihood of B is unaffected by whether or not A occurs. B is independent of A. If B is dependent on A, then $P(B|A)$ will be different from $P(B)$.

Consider the following belief network.

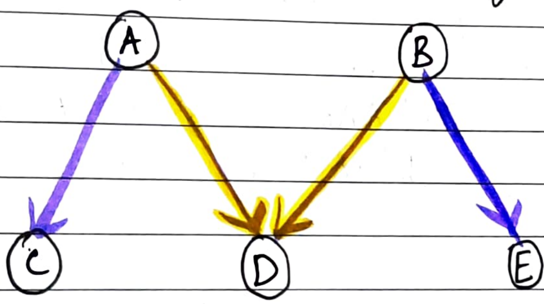


fig. Bay 1

A Bayesian belief network is an acyclic directed graph, where the node in the graph represent evidence or hypotheses, and where an arc that connects two nodes represents a dependence between those two nodes.

In the network above, there are 5 nodes that represent pieces of evidence (A and B) and 3 hypothesis (C, D and E). Arcs between nodes represent the interdependence of the hypotheses. According to this diagram, C and D are both dependent on A, and D and E are both dependent on B. Two nodes that do not have an arc between them are independent of each other. Ex. B is independent of A.

→ Each node in network has a set of probabilities associated with it, based on the values of the nodes on which it is dependent.
Hence, A and B both have just prior probabilities P(C) represents in two cases- A is true and A is false. P(D) also must be represented in four cases, depending on the values of A and B.

appeared Following conditional probabilities might be used in the network (Fig. Bay 1)
$P(A) = 0.1$, $P(B) = 0.7$, $P(C|A) = 0.2$ (you can see how it is represented, since C is dependent on A).
$|||^{ly}$ $P(C|\neg A) = 0.4$, $P(D|A \wedge B) = 0.5$, $P(D|A \wedge \neg B) = 0.4$, $P(D|\neg A \wedge B) = 0.2$, $P(D|\neg A \wedge \neg B) = 0.0001$, $P(E|B) = 0.2$, $P(E|\neg B) = 0.1$.

These probabilities can be also expressed in the form of conditional probability tables, as follows.

P(A)
0.1

P(B)
0.7

A	P(C)
True	0.2
false	0.4

A	B	P(D)
True	True	0.5
True	False	0.4
False	True	0.2
False	False	0.0001

B	P(E)
True	0.2
False	0.1

* Can also represent negative values Ex. $1 - P(A) = P(\neg A)$

* A joint probability can be calculated from Bayesian belief network using the definition of conditional probability.

$$P(B|A) = \frac{P(B \wedge A)}{P(A)}$$

Hence,

$$P(A,B,C,D,E) = P(E|A,B,C,D) * P(A,B,C,D).$$

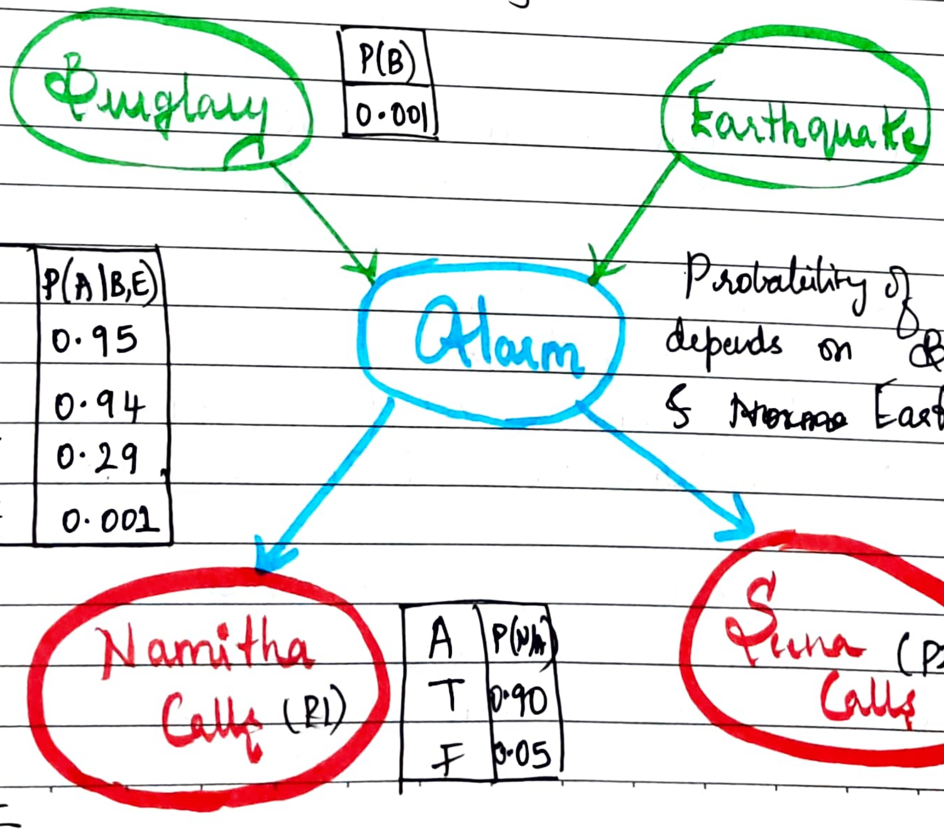
We can apply this rule recursively to obtain

$$P(A,B,C,D,E) = P(E|A,B,C,D) \cdot P(D|A,B,C) \cdot P(C|A,B) \cdot P(B|A) \cdot P(A)$$

Observe E is not dependent on A, C, or D, we can reduce further and write as follows:

$$P(A,B,C,D,E) = P(E|B) \cdot P(D|A,B) \cdot P(C|A) \cdot P(B) \cdot P(A)$$

Ex: Consider the following network:



P(B)
0.001

P(E)
.002

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

Probability of Alarm depends on Burglary & ~~Alarm~~ Earthquake.

A	P(N|A)
T	0.90
F	0.05

A	P(S|A)
T	0.70
F	0.01

You can use → P1 & P2

Consider the problem: Statement

→ New installations of burglar alarm at home.
→ Works fairly well at detecting burglary, but sometimes responds to minor Earthquakes.
→ Two neighbors, Namitha and Seena, who promised to Call you at Work when they hear the alarm.
→ Namitha always calls hearing alarm, but Sometimes Confuses with mobile ringtone and Calls.
→ Seena Keeps her Self busy with music hence Sometimes misses the alarm.
→ Given the Evidence of who has or not Called we would like to Estimate the probability of burglary.
* (you can Express the Statement in your own Words to reduce it, also can use any Other Example as well (Ex: Life at College fromText)

Sol:

Alarm dependence on Burglary and Earthquake

III^ly Namitha and Seena will Call only when Alarm Rings (they are dependent on alarm)

→ Probabilities has been Given to us you can find out for negative probability using foll. Equation: Ex: $P(B) = .001$, No burglary Can be represented as $(1-P(B)) = P(\neg B) = 1-0.001$
$= 0.999$

III^ly Calculate for all others.
Using Bayesian Belief Network you can find joint probabilities b/w any Combination of nodes.

→ 1. So. what is the probability that the alarm has sounded but neither a burglary nor an earthquake has occurred & both Nanditha and Seena call?

Sol:-
$P(N \wedge S \wedge A \wedge \neg B \wedge \neg E) = P(N|A) \cdot P(S|A) \cdot P(A|\neg B, \neg E) \cdot P(\neg B) \cdot P(\neg E)$.

$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998 = 0.00062$.

→ Can also be represented as:

$P(N, S, A, \neg B, \sim E) = P(N|A) \cdot P(S|A) \cdot P(A|\sim B, \sim E) \cdot P(\sim B) \cdot P(\sim E)$

$= 0.00062$ //

4. The Naive Bayes Classifier (Problem — 8-10mks) (theory with problem — 10 mks).

appear – An effective learning system where each piece of data that is to be classified consists of a set of attributes, each of which can take on a number of possible values. The data are then classified into a single classification.

$d_1 \to$ data, posterior probability of each possible classification is calculated: $P(c_i | d_1, \ldots, d_n)$
where c_i is the i^{th} classification, from a set of $|c|$ classifications.

To calculate the posterior probability, we can use Bayes Theorem and rewrite it as:

$$\frac{P(d_1, \ldots, d_n | c_i) \cdot P(c_i)}{P(d_1, \ldots, d_n)}$$

Because, we are trying to find highest probability & because $P(d_1, \ldots, d_n)$ is a constant independent of c_i, we can eliminate it and simply aim to find the classification c_i, for which foll. is maximized

PT

$P(d_1, \ldots, d_n | c_i) \cdot P(c_i)$

The naive Bayes Classifier now assumes that each of the attributes in the data item is independent of the others, in which case $P(d_1, \ldots, d_n | c_i)$ can be rewritten and the foll. value obtained:

$$P(c_i) \cdot \prod_{j=1}^{n} P(d_j | c_i)$$

Thus, Naive Bayes classifier selects a classification for a data set by finding the classification c_i for which the above calculation is a maximum.

Let us now see few examples:

1. Consider the foll. dataset with 14 instances:

Day	Outlook	Temperature	Humidity	Wind	Play Tennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot ~~Mild~~	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

Attributes (Outlook, Temp, Humidity, Wind).
Target Attribute - Play Tennis.

We need to Check for the new instance whether it Can be classified to Yes or No.

New Instance
(Outlook = Sunny, Temperature = cool, Humidity = high, Wind = Strong).

→ Firstly Calculate the prior probabilities and Current Probabilities.

→ Here prior probabilities is p(Yes) and P(No).

P(Yes) = 9/14 → P(Play Tennis = Yes) = 0.64

P(No) = 5/14 → P(PlayTennis = No) 0.36.

Calculate Conditional Probabilities of all attributes.

Outlook	Y	N
Sunny	2/9	3/5
Overcast	4/9	0
Rain	3/9	2/5.

Humidity	Y	N
high	3/9	4/5
normal	6/9	1/5

Temperature	Y	N
hot	2/9	2/5
mild	4/9	2/5
Cool	3/9	1/5

Windy	Y	N
Strong	3/9	3/5
weak	6/9	2/5

(Outlook = Sunny, Temp = Cool, Humidity = high, Wind = Strong) —— ?

$$V_{NB} = P(V_j) * \prod_{j=1}^{n} P(d_j | c_i)$$

$$= \underset{v_j \in \{yes, no\}}{\operatorname{argmax}} P(V_j) \prod_i P(d_j | c_i)$$

$$= \underset{v_j \in \{Yes, No\}}{\operatorname{argmax}} P(V_j) * P(Outlook = Sunny | V_j) \, P(Temp = Cool | V_j) \cdot P(Humidity = high | V_j) \cdot P(Wind = Strong | V_j).$$

$\therefore V_{NB}(Yes) = P(Yes) \cdot P(Sunny | Yes) \cdot P(Cool | Yes) \cdot P(high | Yes)$
$P(Strong | Yes) = 0.0053.$

$V_{NB}(No) = P(No) \cdot P(Sunny | no) \cdot P(Cool | No) \cdot P(high | No).$
$P(Strong | No) = 0.0206.$

Thus, Highest Probability is of No. $\therefore$, New Instance is classified as No.
In Order to Check if your answer is correct, do the following:

$$V_{NB}(Yes) = \frac{V_{NB}(Yes)}{V_{NB}(Yes) + V_{NB}(No)} = \cancel{0.0205} \; 0.205$$

$$V_{NB}(No) = \frac{V_{NB}(No)}{V_{NB}(Yes) + V_{NB}(No)} = \frac{0.0206}{0.0053 + 0.0206} = 0\cancel{.7747}$$

$$= \frac{0.0206}{0.0259} = 0.795$$

$\therefore$, Now add these two
$0.205 + 0.795 = 1.$

Thus your answer is right if you get the sum as 1.

$\boxed{(Outlook = Sunny, Temp = Cool, Humidity = high, Wind = Strong) = No}$

Ex: 2 Given the Dataset, Classify the new instance into M or H.

New Instance:

Sol: (Color = Green, legs = 2, Height = Tall, and Smelly = No) = ?

→ Prior Probabilities $P(M) = 4/8 = 0.5$, $P(H) = \frac{4}{8} = 0.5$.

No.	Color	Legs	Height	Smelly	Species
1	White	3	Short	Yes	M
2	Green	2	Tall	No	M
3	Green	3	Short	Yes	M
4	White	3	Short	Yes	M
5	Green	2	Short	No	H
6	White	2	Tall	No	H
7	White	2	Tall	No	H
8	White	2	Short	Yes	H

← Conditional Probabilities →

Color	M	H
White	2/4	3/4
Green	2/4	1/4

legs	M	H
2	1/4	4/4
3	3/4	0/4

Height	M	H
Tall	1/4	2/4
Short	3/4	2/4

Smelly	M	H
Yes	3/4	1/4
No	1/4	3/4

$P(M \mid \text{New Instance}) = P(M) * P(\text{Color} = \text{Green} \mid M) * P(\text{Legs}=2 \mid M) * P(\text{Height} = \text{tall} \mid M) * P(\text{Smelly} = \text{no} \mid M)$

$P(M \mid \text{New Instance}) = 0.5 * \frac{2}{4} * \frac{1}{4} * \frac{1}{4} * \frac{1}{4} = 0.00390 = \boxed{0.00390}$

$P(H \mid \text{New Instance}) = P(H) * P(\text{Color} = \text{Green} \mid H) * P(\text{Legs} = 2 \mid H) * P(\text{Height} = \text{tall} \mid H) * P(\text{Smelly} = \text{no} \mid H)$

$P(H \mid \text{New Instance}) = 0.5 * \frac{1}{4} * \frac{4}{4} * \frac{2}{4} * \frac{3}{4} = 0.047 \rightarrow (0.0468)$

P.T.O

Now Compare →

$P(H \mid \text{New Instance}) > P(M \mid \text{New Instance})$

$\boxed{\text{Hence new Instance belongs to Speices H.}}$

We Can also Check.

$$\frac{P(H)}{P(H)+P(M)} + \frac{P(M)}{P(H)+P(M)}$$

$$\frac{0.047}{0.047+0.00390} + \frac{0.00390}{0.047+0.00390}$$

$$= \frac{0.047}{0.0509} + \frac{0.00390}{0.0509}$$

$= ~~0.9233 + 0.0764~~$

$= 0.9233 + 0.07662$

$$\boxed{= 0.9999 = 1}$$

Hence our answer is right.

* The Noisy -V- Function — 5mks with Example.

In some situation, it can be possible to use the fact that events in a Bayesian belief Network are related to each other by some kind of mathematical or logical relation.

appeared

Here, fuzzy logic could provide suitable relation. Another useful class of relations is noisy logical relationships.

Let us consider our diagonosis example.

P (A|B) – if one has a cold, then one will also have a high temperature.

IIIy. P (A|C) – if one has the plague, then one will also have a high temperature

We have the following:

$P(A|B) = 0.8$

$P(A|C) = 0.99$

Noisy-V function is based on the assumption that the only possible causes of a high temperature are a cold and the plague (i.e, $P(A|B \vee C) = 1$.

Clearly this is not true in our Ex., but we can fix this by including a leak node in the network which represents all possible causes.

Hence, we include $P(A|D) = 0.9$, where D is the leak node, which represents other causes of a high temperature.

Let us define Noise Parameters for these relationships.

$P(\neg A|B) = 1 - P(A|B) = 0.2$

$P(\neg A|C) = 1 - P(A|C) = 0.01$

$P(\neg A|D) = 1 - P(A|D) = 0.1$

* Noise Parameters are independent of Each Other.

The Noisy-V function for B, C, and D is defined as:

If B, C, and D are all false, then P(A)=0.
Otherwise, P(¬A) is Equal to the product of noise parameters for all Variables that are true.
Ex: If B is true, C and D are false then P(¬A) is Equal to noise parameter for B, and so:

$$P(A) = 1 - 0.2$$
$$= 0.8.$$

If C and D are true, B is false, then P(¬A) is Equal to product of noise parameter for C and D, and so:

$$P(A) = 1 - (0.01 \times 0.1)$$
$$= 0.999.$$

Noisy-V function for our medical diagnosis Example.

B	C	D	P(A)	P(¬A)
false	false	false	0	1
false	false	true	0.9	0.1
false	true	false	0.99	0.01
false	true	true	0.999	$0.01 \times 0.1 = 0.001$
true	false	false	0.8	0.2
true	false	true	0.98	$0.2 \times 0.1 = 0.02$
true	true	false	0.998	$0.2 \times 0.01 = 0.002$
true	true	true	0.9998	$0.2 \times 0.01 \times 0.1 = 0.0002$

* (Write acc. to marks - usually asked for 5 mks).

* Collaborative Filtering (often asked for 5 mks).

It is used for Bayesian reasoning. It is a technique that is increasingly used by online stores (such as Amazon.com, ebay, flipkart ..etc). to provide plausible suggestions to customers based on their previous purchases.

Idea is: If we know that Anne and Bob both like items A, B, and C, and that Anne likes D, then it is reasonable to suppose that Bob would also like D.

Collaborative filtering can be implemented using the Bayesian inference. (which is quite successful method). This involves working with posterior probabilities such as follows:

P (Bob likes Z | Bob likes A, Bob likes B, .., Bob likes Y)

Large amount of data must be collected for this to work accurately. In case of commerce sites, information can be collected on the basis of assuming that if a user buys a book or a CD, then he probably likes it. Data can also be collected by asking the users to rate the products.

Ex Showing working of Collaborative filtering using decision tree.

fig: Coll 1
Decision tree for Collaborative filtering.