UNIT-3

Bayesian concept learning Intoduction

- Why Bayesian methods are important?
- Applications of Bayesian classifiers
- strength of Bayesian classifier
- features of Bayesian learning methods
- advantages of Bayesian classifiers

Bayesian concept learning Intoduction

Bayes Theorem:

- The Bayes Theorem was derived from the work of the 18th century mathematician Thomas Bayes.
- He developed the foundational mathematical principles, known as Bayesian methods, which describe
- the probability of events, and how probabilities should be revised when there is additional information.

WHY BAYESIAN METHODS ARE IMPORTANT?

- Bayesian learning algorithms, like the naive Bayes classifier which calculates the explicit probabilities for hypotheses.
- Bayesian classifiers, the training data are utilized to calculate an observed probability of each class based on feature values.
- These naïve bayes classifier is equally powerful as decision tree and neural network algorithms. Some times it may give even better results.
- It is highly practical approaches to certain types of learning problems

Applications of Bayesian Classifiers

- Text based classification
- Medical diagnosis system
- Network security
- Face recognition
- Weather prediction
- Credit scoring
- Other real time prediction problems.

Strength of Bayesian classifier

- the strengths of Bayesian classifiers is that they utilize all available parameters to the predictions,
- many other algorithms ignore the features that have weak effects.
- In Bayesian classifiers, even if few individual parameters have small effect on the outcome, the collective effect of those parameters could be quite large.
- For such learning tasks, the naive Bayes classifier is most effective.

Features of Bayesian learning methods

- Prior knowledge of the candidate hypothesis is combined with the observed data for arriving at the final probability of a hypothesis.

two important components are

- the prior probability of each candidate hypothesis and
- the probability distribution over the observed data set for each possible hypothesis.

- The Bayesian approach to learning is more flexible than the other approaches, because
  - each observed training pattern can influence the outcome of the hypothesis by increasing or decreasing the estimated probability about the hypothesis,
- whereas most of the other algorithms tend to eliminate a hypothesis if that is inconsistent with the single training pattern.

Features of Bayesian learning methods...

- Bayesian methods can perform better than the other methods while validating the hypotheses that make probabilistic predictions.
- For example, when starting a new software project, on the basis of the demographics of the project, we can predict the probability of encountering challenges during execution of the project.
- In Bayesian methods, it is possible to classify new instances by combining the predictions of multiple hypotheses, weighted by their respective probabilities.

Features of Bayesian learning methods...

- Features of Bayesian learning methods...
- the Bayesian method largely depends on the availability of initial knowledge about the probabilities of the hypothesis set.
- if these probabilities are not known to us in advance, we have to use
  - some background knowledge,
  - previous data or assumptions about the data set, and
  - the related probability distribution functions to apply this method.
- Moreover, it normally involves high computational cost to arrive at the optimal Bayes hypothesis.

Advantages of Bayes Classifier

- It doesn't require as much training data hence it is simple and easy to implement
- It handle both classification and regression problems I.e. it can use both continuous and discrete data
- It is highly scalable with the number of predictors and data points
- It is fast and can be used to make real-time predictions. • It is not sensitive to irrelevant features

concept learning.

- child starts to learn meaning of new words, e.g. ‘ball’.
- The child is provided with positive examples of ‘objects’ which are ‘ball’. At first, the child may be confused with many different colours, shapes and sizes of the balls and may also get confused with some objects which look similar to ball, like a balloon or a globe. The child’s parent continuously feeds her positive examples like ‘that is a ball’, ‘this is a green ball’, ‘bring me that small ball’, etc.

concept learning.

- there are negative examples used for such concept teaching, like ‘this is a non-ball’, but the parent may clear the confusion of the child when it points to a balloon and says it is a ball by saying ‘that is not a ball’.
- But it is observed that the learning is most influenced through positive examples rather than through negative examples, and the expectation is that the child will be able to identify the object ‘ball’ from a wide variety of objects and different types of balls kept together once the concept of a ball is clear to her. We can extend this example to explain how we can expect machines to learn through the feeding of positive examples, which forms the basis for concept learning

BAYES’ THEOREM

- The learning concept with the mathematical model of Bayes,
- we can correlate the learning process of ‘meaning of a word’ as equivalent to learning, a concept using binary classification.
- a concept set C and a corresponding function f(k).
- f(k) = 1, when k is within the set C and f(k) = 0 otherwise.
- to learn the function f that defines which elements are within the set C.
- So, by using the function f, we will be able to classify the element either inside or outside our concept set.
- In Bayes’ theorem, we will learn how to use standard probability calculus to determine the uncertainty about the function f, and we can validate the classification by feeding positive examples.

Bayes' probability rule

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

where A and B are conditionally related events and $p(A|B)$ denotes the probability of event A occurring when event B has already occurred.

As $P(h_2|\mathrm{PT})$ is higher than $P(h_1|\mathrm{PT})$, it is clear that the hypothesis h_2 has more probability of being true. So, *h*MAP = *h*2 = !MT.

This indicates that even if the posterior probability of malignancy is significantly higher than that of non-malignancy, the probability of this patient not having malignancy is still higher on the basis of the prior knowledge. Also, it should be noted that through Bayes' theorem, we identified the probability of one hypothesis being higher than the other hypothesis, and we did not completely accept or reject the hypothesis by this theorem. Furthermore, there is very high dependency on the availability of the prior data for successful application of Bayes' theorem.

BAYES’ THEOREM

$$h_{\mathrm{ML}} = \mathrm{argmax}_{h \in H} P(T|h)$$

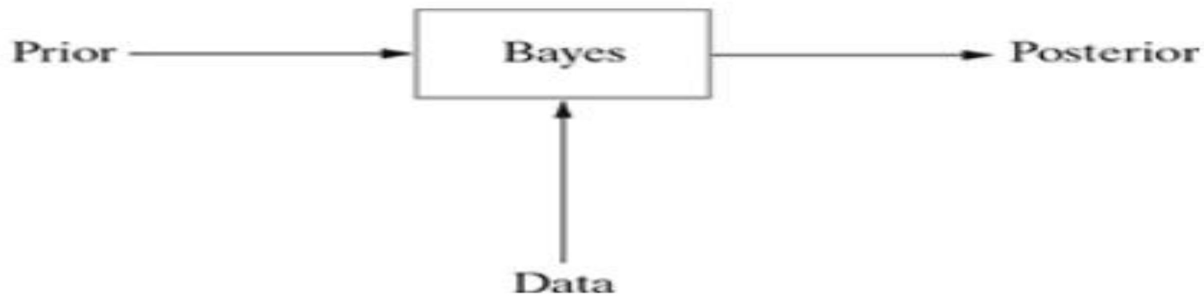


FIG. 6.1
Bayes’ theorem

likelihood

prior probability

$$P(h|T) = \frac{P(T|h)\,P(h)}{P(T)}$$

posterior probability

marginal likelihood

CONCEPT OF PRIOR,POSTERIOR

- A training data set has some observed data
- To determine the best hypothesis in space h by using the knowledge of T
- Prior probability p(h)
- The prior knowledge or belief about the probabilities of various hypotheses in H is ca
lled Prior probability.

Prosterior probability p(h/T)

The probability that a particular hypothesis holds for a data set based on the Prior is called the posterior probability

maximum a posteriori (MAP) hypothesis.

- to find out the maximum probable hypothesis h from a set of hypotheses H (h∈H) given the observed training data T. This maximally probable hypothesis is called the maximum a posteriori (MAP) hypothesis. By using Bayes' theorem, we can identify the MAP hypothesis from the posterior probability of each candidate hypothesis:

-

$$h_{\mathrm{MAP}} = \mathrm{argmax}_{h \in H} P(h|T)$$
$$= \mathrm{argmax}_{h \in H} \frac{P(T|h)P(h)}{P(T)}$$

and as $P(T)$ is a constant independent of h, in this case, we can write

$$= \mathrm{argmax}_{h \in H} P(T|h)P(h) \qquad (6.1)$$

According to Bayes' theorem

$$P(h|T) = \frac{P(T|h)P(h)}{P(T)}$$

combines the prior and posterior probabilities together.
P(h|T) increases as P(h) and P(T|h) increases and also as P(T) decreases.

$$\underbrace{P(h|T)}_{\text{posterior probability}} = \frac{\overbrace{P(T|h)}^{\text{likelihood}}\ \overbrace{P(h)}^{\text{prior probability}}}{\underbrace{P(T)}_{\text{marginal likelihood}}}$$

Likelihood

- if every hypothesis in H has equal probable priori as P(h) = P(h), and then, we can determine P(h|T) from the probability P(T|h) only. Thus, P(T|h) is called the likelihood of data T given h, and any hypothesis that maximizes P(T|h) is i j called the maximum likelihood (ML) hypothesis, h .

$$h_{\mathrm{ML}} = \mathrm{argmax}_{h \in H} P(T|h)$$

the example of malignancy identification in a particular patient's tumour

- We have two alternative hypotheses:
- (1) a particular tumour is of malignant type and
- (2) a particular tumour is non-malignant type.
- The priori available are—
- 1. only 0.5% of the population has this kind of tumour which is malignant,
- 2. the laboratory report has some amount of incorrectness as it could detect the malignancy was present only with 98% accuracy whereas could show the malignancy was not present correctly only in 97% of cases. This means the test predicted malignancy was present which actually was a false alarm in 2% of the cases, and also missed detecting the real malignant tumour in 3% of the cases.

EX

Solution: Let us denote Malignant Tumour = MT, Positive Lab Test = PT, Negative Lab Test = NT

h_1 = the particular tumour is of malignant type = MT in our example

h_2 = the particular tumour is not malignant type = !MT in our example

$$P(\text{MT}) = 0.005 \quad P(\text{!MT}) = 0.995$$
$$P(\text{PT}|\text{MT}) = 0.98 \quad P(\text{PT}|\text{!MT}) = 0.02$$
$$P(\text{NT}|\text{!MT}) = 0.97 \quad P(\text{NT}|\text{MT}) = 0.03$$

$$
\begin{aligned}
P(h_1|\mathrm{PT}) &= \frac{P(\mathrm{PT}|h_1).P(h_1)}{P(\mathrm{PT})} \\
&= P(\mathrm{PT}|\mathrm{MT})P(\mathrm{MT}) \\
&= 0.98 \times 0.005 \\
&= 0.0049 \\
&= 0.49\%
\end{aligned}
$$

$$
\begin{aligned}
P(h_2|\mathrm{PT}) &= \frac{P(\mathrm{PT}|h_2).P(h_2)}{P(\mathrm{PT})} \\
&= P(\mathrm{PT}|!\mathrm{MT})P(!\mathrm{MT}) \\
&= 0.02 \times 0.995 \\
&= 0.0199 \\
&= 1.99\%
\end{aligned}
$$

BAYES’ THEOREM AND CONCEPT LEARNING

Brute-force Bayesian algorithm

Concept of consistent learners

Bayes optimal classifier

Naïve Bayes classifier

Applications of Naïve Bayes classifier

Handling Continuous Numeric Features in Naïve Bayes Classifier

Brute-force Bayesian algorithm

The brute-force Bayesian algorithm applies Bayes' theorem to each possible hypothesis in the hypothesis space. This means for every possible hypothesis H_i, you calculate the posterior probability $P(H_i|E)$. The hypothesis with the highest posterior probability is then selected as the best explanation for the evidence.

$$P(h|T) = \frac{P(T|h)P(h)}{P(T)}$$

h_{MAP} with the highest posterior probability

$$h_{\mathrm{MAP}} = \mathrm{argmax}_{h \in H} P(h|T)$$

1. The training data or target sequence T is noise free, which means that it is a direct function of X only (i.e. $t_i = c(x_i)$)
2. The concept c lies within the hypothesis space H
3. Each hypothesis is equally probable and independent of each other

$$P(h) = \frac{1}{|H|} \text{ for all } h \text{ within } H$$

$P(T|h)$ is the probability of observing the target values t_i in the fixed set of instances $\{x_1, \ldots, x_m)$ in the space where h holds true and describes the concept c correctly. Using assumption 1 mentioned above, we can say that if T is consistent with h,

Using Bayes' theorem to identify the posterior probability

$$P(T|h) = \begin{cases} 1 & \text{if } t_i = h(x_i) \text{ for all } t_i \text{ within } T \\ 0 & \text{otherwise} \end{cases}$$

$$P(h|T) = \frac{P(T|h)P(h)}{P(T)}$$

$$P(h|T) = \frac{0 \times P(h)}{P(T)} = 0, \text{ when } h \text{ is inconsistent with } T,$$

and when h is consistent with T

$$P(h|T) = \frac{1 \times \frac{1}{|H|}}{P(T)} = \frac{1}{|H|P(T)}$$

if we define a subset of the hypothesis H which is consistent with T as H , then by using the total probability equation, we get

$$\begin{aligned} P(T) &= \sum_{h_i \in H_D} P(T|h_i)P(h_i) \\ &= \sum_{h_i \in H_D} 1.\frac{1}{|H|} + \sum_{h_i \notin H_D} 0.\frac{1}{|H|} \\ &= \sum_{h_i \in H_D} 1.\frac{1}{|H|} \\ &= \frac{|H_D|}{|H|} \end{aligned}$$

So, with our set of assumptions about $P(h)$ and $P(T|h)$, we get the posterior probability $P(h|T)$ as

$$P(h|T) = \begin{cases} \frac{1}{|H_D|} & \text{if } h \text{ is consistent with } T \\ 0 & \text{otherwise} \end{cases} \quad ($$

Concept of consistent learners

the group of learners who commit zero error over the training data and output the hypothesis are called consistent learners

If the training data is noise free and deterministic (i.e. $P(D|h) = 1$ if D and h are consistent and 0 otherwise) and if there is uniform prior probability distribution over H (so, $P(h_m) = P(h_n)$ for all m, n), then every consistent learner outputs the MAP hypothesis

Bayes optimal classifier

The **Bayes Optimal Classifier** is an idealized classifier that makes predictions by choosing the class with the highest posterior probability, given the data. It represents the best possible classification decision rule one could achieve under the Bayesian framework.

To determine the most probable classification for a new instance x, the **Bayes Optimal Classifier** considers the predictions of all hypotheses in the hypothesis space. The predictions are combined, each weighted by the posterior probability of the respective hypothesis. The goal is to maximize the probability of correctly classifying the new instance.

Steps to Determine the Most Probable Classification:

1. **Consider Possible Classifications**:
   - Let $C = \{c_1, c_2, \ldots, c_k\}$ be the set of all possible classifications for the new instance x. For example, in a binary classification problem, $C = \{\text{True}, \text{False}\}$.
2. **Calculate the Probability for Each Possible Classification** c:
   - For each possible classification c, the probability $P(c|T)$ that the new instance x is correctly classified as c given the training data T is computed by combining the predictions of all hypotheses weighted by their posterior probabilities:

$$P(c|T) = \sum_{h \in H} P(h|T) \cdot P(c|h)$$

Bayes optimal classifier EXAMPLE

The set of possible outcomes for the new instance x is within the set $C = \{\text{True}, \text{False}\}$ and

$$P(h_1 \mid T) = 0.4,\ P(\text{False} \mid h_1) = 0,\ P(\text{True} \mid h_1) = 1$$
$$P(h_2 \mid T) = 0.3,\ P(\text{False} \mid h_2) = 1,\ P(\text{True} \mid h_2) = 0$$
$$P(h_3 \mid T) = 0.3,\ P(\text{False} \mid h_3) = 1,\ P(\text{True} \mid h_3) = 0$$

Then,

$$\sum_{h_i \in H} P(\text{True} \mid h_i) P(h_i \mid T) = 0.4$$

$$\sum_{h_i \in H} P(\text{False} \mid h_i) P(h_i \mid T) = 0.6$$

Naïve Bayes classifier

- **Naïve Bayes algorithm** is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems.
- It is a probabilistic classifier, which means it predicts on the basis of the probability of an object.

Naïve: It is called Naïve because it assumes that the occurrence of a certain feature is independent of the occurrence of other features.

Bayes: It is called Bayes because it depends on the principle of Bayes' Theorem.

Bayes' Theorem:

- Bayes' theorem is also known as **Bayes' Rule** or **Bayes' law**, which is used to determine the probability of a hypothesis with prior knowledge. It depends on the conditional probability.
- The formula for Bayes' theorem is given as:

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

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

Where,

P(A|B) is Posterior probability: Probability of hypothesis A on the observed event B.

P(B|A) is Likelihood probability: Probability of the evidence given that the probability of a hypothesis is true.

P(A) is Prior Probability: Probability of hypothesis before observing the evidence.

P(B) is Marginal Probability: Probability of Evidence.

Naïve Bayes classifier steps

Step 1: First construct a frequency table. A frequency table is drawn for each attribute against the target outcome.

Step 2: Identify the cumulative probability.

Step 3: Calculate probability through normalization by applying the below formula

$$P(\text{Yes}) = \frac{P(\text{Yes})}{P(\text{Yes}) + P(\text{No})}$$

$$P(\text{No}) = \frac{P(\text{No})}{P(\text{Yes}) + P(\text{No})}$$

P(Yes) will give the overall probability of favorable condition in the given scenario.

P(No) will give the overall probability of non favorable condition in the given scenario.

Advantages of Naïve Bayes Classifier:

1. Naïve Bayes is one of the fast and easy ML algorithms to predict a class of datasets.
2. It can be used for Binary as well as Multi-class Classifications.
3. It performs well in Multi-class predictions as compared to the other Algorithms.
4. It is the most popular choice for **text classification problems**.

Disadvantages of Naïve Bayes Classifier:

Naive Bayes assumes that all features are independent or unrelated, so it cannot learn the relationship between features.

Applications of Naïve Bayes Classifier:

1. Text Classification
2. Spam Filtering
3. Hybrid Recommender system
4. Online Sentimental anlysis

EXAMPLE

Day	*Outlook*	*Temperature*	*Humidity*	*Wind*	*PlayTennis*
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	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

TEST DATA

$$\langle Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong \rangle$$

$$P(PlayTennis = yes) = 9/14 = .64$$

$$P(PlayTennis = no) = 5/14 = .36$$

Step-1

Outlook	Y	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5

Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

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

Windy		
Strong	3/9	3/5
Weak	6/9	2/5

Step-2

$$v_{NB} = \underset{v_j \in \{yes,no\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

$$= \underset{v_j \in \{yes,no\}}{\operatorname{argmax}} P(v_j) \quad P(Outlook = sunny|v_j) P(Temperature = cool|v_j)$$

$$\cdot P(Humidity = high|v_j) P(Wind = strong|v_j)$$

$$\langle Outlook = sunny, Temperature = cool, Humidity = high, Wind = strong \rangle$$

$$v_{NB} = \underset{v_j \in \{yes,no\}}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j)$$

$$= \underset{v_j \in \{yes,no\}}{\operatorname{argmax}} P(v_j) \quad P(Outlook = sunny|v_j) P(Temperature = cool|v_j)$$

$$\cdot P(Humidity = high|v_j) P(Wind = strong|v_j)$$

$$v_{NB}(yes) = P(yes)\ P(sunny|yes)\ P(cool|yes)\ P(high|yes)\ P(strong|yes) = .0053$$

$$v_{NB}(no) = P(no)\ P(sunny|no)\ P(cool|no)\ P(high|no)\ P(strong|no) = .0206$$

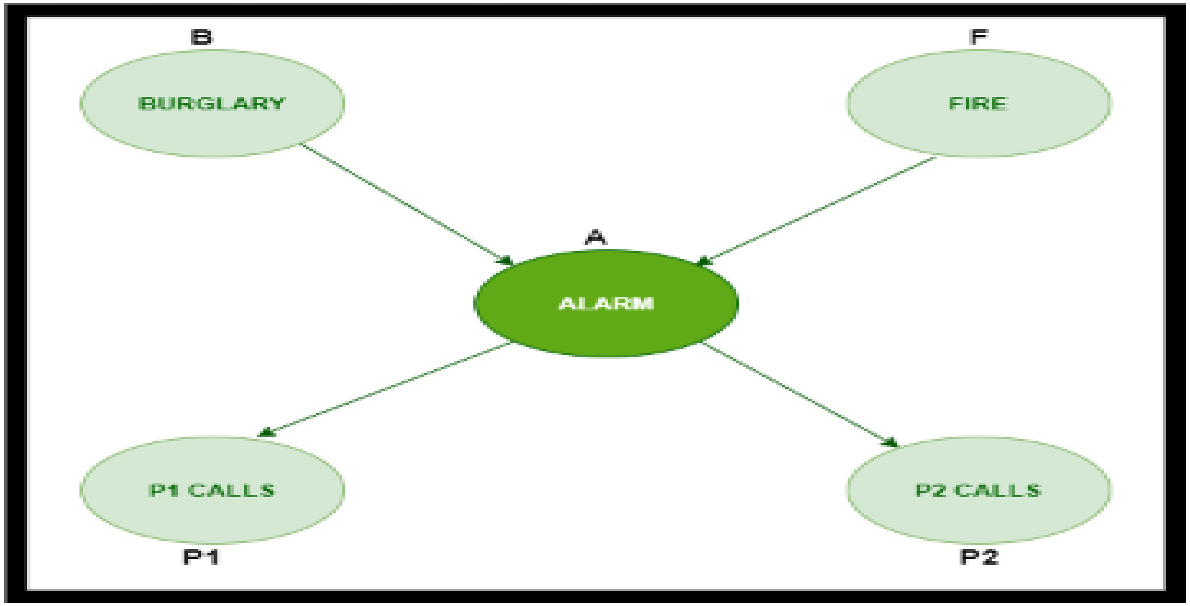
Step-3

$$v_{NB}(yes) = \frac{v_{NB}(yes)}{v_{NB}(yes)+v_{NB}(no)} = 0.205$$

$$v_{NB}(no) = \frac{v_{NB}(no)}{v_{NB}(yes)+v_{NB}(no)} = 0.795$$

by normalizing the above two probabilities, we can ensure that the sum of these two probabilities is **1**.

Bayesian Belief Network is a graphical representation of different probabilistic relationships among random variables in a particular set. It is a classifier with no dependency on attributes i.e it is condition independent. Due to its feature of joint probability, the probability in Bayesian Belief Network is derived, based on a condition — P(attribute/parent) i.e probability of an attribute, true over parent attribute.



- **Nodes** represent random variables, which can be discrete or continuous.
- **Edges** represent conditional dependencies between the variables. If there is an edge from node A to node B, it implies that B is conditionally dependent on A.

2. **Conditional Probability Distribution (CPD):**
   - Each node in the network has an associated CPD that quantifies the effect of its parent nodes. If a node has no parents, it has a prior probability distribution.
3. **Joint Probability Distribution:**
   - The BBN represents the joint probability distribution of the variables as a product of the CPDs of each node given its parents. If $X_1, X_2, \ldots, X_n$ are the variables in the network:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{Parents}(X_i))$$

4. **Inference:**
   - Given observed data (evidence), BBNs allow for the computation of posterior probabilities for other variables. This process is known as **Bayesian inference**.

Bayesian network is based on Joint probability distribution and conditional probability. So let's first understand the joint probability distribution:

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.

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

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

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

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

Explanation of Bayesian network:

Q) Find the probability that 'P1' is true (P1 has called 'gfg'), 'P2' is true (P2 has called 'gfg') when the alarm 'A' rang, but no burglary 'B' and fire 'F' has occurred.

=> **P (P1, P2, A, ~B, ~F)** [where- P1, P2 & A are 'true' events and '~B' & '~F' are 'false' events]

[**Note:** The values mentioned below are neither calculated nor computed. They have observed values]

Burglary 'B' –

- **P (B=T) = 0.001** ('B' is true i.e burglary has occurred)
- **P (B=F) = 0.999** ('B' is false i.e burglary has not occurred)

Fire 'F' –

- **P (F=T) = 0.002** ('F' is true i.e fire has occurred)
- **P (F=F) = 0.998** ('F' is false i.e fire has not occurred)

Alarm 'A' –

B	F	P (A=T)	P (A=F)
T	T	0.95	0.05
T	F	0.94	0.06
F	T	0.29	0.71
F	F	0.001	**0.999**

Person 'P1' –

A	P (P1=T)	P (P1=F)
T	**0.95**	0.05
F	0.05	0.95

Person 'P2' –

A	P (P2=T)	P (P2=F)
T	**0.80**	0.20
F	0.01	0.99

P (P1, P2, A, ~B, ~F)

= P (P1/A) * P (P2/A) * P (A/~B~F) * P (~B) * P (~F)

= 0.95 * 0.80 * 0.001 * 0.999 * 0.998

= 0.00075

Bayesian Belief Network Apllications

Real world applications are probabilistic in nature, and to represent the relationship between multiple events, we need a Bayesian network.
It can also be used in various tasks including

1.prediction,
2. anomaly detection,
3.diagnostics,
4.automated insight,
5.reasoning,
6.time series prediction, and
7.decision making under uncertainty.