Bayesian Belief Network (BBN)

Definition of BBN

It is a acyclic (with no cycles) directed graph where the nodes of the graph represent evidence or hypotheses and arc connecting two nodes represents dependence between them. If there is an arc from node X to another node Y (i.e., $X \rightarrow Y$), then X is called a *parent* of Y, and Y is a *child* of X. The set of parent nodes of a node X_i is represented by parent_nodes(X_i).

Need of BBN

Joint probability distribution for *n* variables requires 2^n entries with all possible combinations. The time and storage requirements for such computations become impractical as n grows. Therefore, inferring with such large numbers of probabilities does not seem to model human process of reasoning. Human tends to single out few propositions which are known to be causally linked when reasoning with uncertain beliefs. This leads to the concept of forming belief network called a *Bayesian belief network.*

Joint probability distribution of two variables A and B are given in the following table.

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

BBN is a probabilistic graphical model that encodes probabilistic relationships among set of variables with their probabilistic dependencies. This belief network is an efficient structure for storing joint probability distribution.

Joint Probability of n variables

Joint probability for 'n' variables (dependent or independent) is computed as follows.

For the sake of simplicity we write $P(X_1 , \ldots , X_n)$ instead of $P(X_1$ and $\ldots$ and $X_n)$.

$$P(X_1 , \ldots ,X_n) = P(X_n \mid X_1 ,\ldots, X_{n-1}) * P(X_1 , \ldots , X_{n-1})$$

Or

$$P(X_1 , \ldots , X_n) = P(X_n \mid X_1 , \ldots , X_{n-1}) * P(X_{n-1} \mid X_1 , \ldots , X_{n-2}) * \ldots * P(X_2 \mid X_1) * P(X_1)$$

Joint Probability of 'n' Variables using B-Network

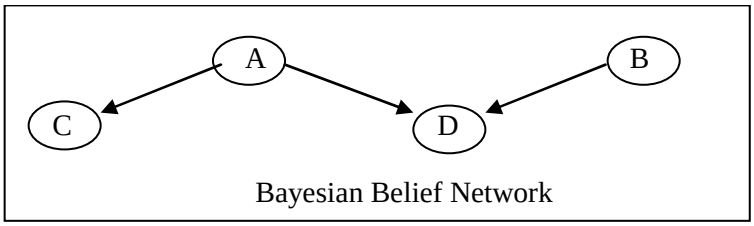
In Bayesian network, the joint probability distribution can be written as the product of the local distributions of each node and its parents such as:

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

This expression is reduction of joint probability formula of 'n' variables as some of the terms corresponding to independent variables will not be required. If node Xi has no parents, its probability distribution is said to be unconditional and it is written as P(Xi) instead of P(Xi | parent_nodes(Xi)). Nodes having parents are called conditional. 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 node and nodes with no parents are called independent nodes.

Example

The following graph is a Bayesian belief network. Here there are four nodes with {A, B} representing evidences and {C, D} representing hypotheses that is A and B are unconditional nodes and C and D are conditional nodes.



Bayesian Belief Network

To describe above Bayesian network, we should specify the following probabilities.

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

The conditional probability tables are as follows:

Conditional Probability Tables

P(A)	P(B)
0.3	0.6

A	P(C)
T	0.4
F	0.2

A	B	P(D)
T	T	0.7
T	F	0.4
F	T	0.2
F	F	0.01

Using Bayesian belief network, only 8 probability values in contrast to 16 values are required in general for 4 variables {A, B, C, D} in joint distribution probability.

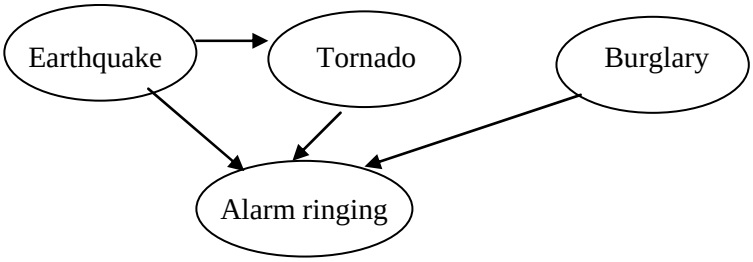
Joint probability using Bayesian Belief Network is computed as follows:

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

Example of Simple B-Network

Assume that there are three events namely earthquake, burglary or tornado which could cause ringing of alarm in a house. This situation can be modeled with Bayesian network as follows.

Here the names of the variables have been abbreviated to *A = Alarm*, *E = Earthquake*, and *B = Burglary* and *T = Tornado*. All four variables have two possible values T (for true) and F (for false).



Prior probability of 'earthquake' is 0.4 and if it is earthquake then probability of 'tornado' is 0.8. and if not then the probability of 'earthquake' is 0.5. Table contains the probability values representing complete Bayesian belief network.

- **Conditional Probability Tables**

P(E)	P(B)
0.4	0.7

E	P(Tor)
T	0.8
F	0.5

E	B	Tor	P(A)
T	T	T	1.0
T	T	F	0.9
T	F	T	0.95
T	F	F	0.85
F	T	T	0.89
F	T	F	0.7
F	F	T	0.87
F	F	F	0.3

The joint probability is computed as follows:

P(E, B, T, A) = P(A| E, B, T) * P(T|E) * P(E) * P(B)

= 1.0 * 0.8 * 0.4 * 0.7 = 0.214

From this model, using the conditional probability formula further we can compute the following:

- the probability that it is earthquake, given the alarm is ringing - P(E|A)
- the probability of burglary, given the alarm is ringing - P(B|A)
- the probability of ringing alarm if both earthquake and burglary happens - P(A|E, B)

Advantages of Bayesian Belief Network

- It can easily handle situations where some data entries are missing as this model encodes dependencies among all variables.
- It is intuitively easier for a human to understand direct dependencies than complete joint distribution.
- It can be used to learn different kinds of inferences (diagnostic, casual, inter casual, mixed).
- It is an ideal representation for combining prior knowledge (which often comes in causal form) and data because the model has both causal and probabilistic semantics.

Disadvantages of Bayesian Belief Network

- 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 only.
- The quality and extent of the prior beliefs used in Bayesian inference processing are major shortcomings.
- 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 distribution for modeling the data is very important.

Dempster–Shafer Theory

Dempster–Shafer theory is a mathematical theory of evidence. It allows one to combine evidence from different sources and arrive at a degree of belief.

Belief function 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. These degrees of belief may or may not have the mathematical properties of probabilities. The difference between them will depend on how closely the two events are related. It also uses numbers in the range [0, 1] to indicate amount of belief in a hypothesis for a given piece of evidence.

Degree of belief in a statement depends upon the number of answers to the related questions containing the statement and the probability of each answer. In this formalism, a degree of belief (also referred to as a mass) is represented as a belief function rather than a Bayesian probability distribution

Example

Mary and John are friends.

- Suppose Mary tells John that his car is stolen. Then John's belief on the truth of this statement will depend on the reliability of Mary. But it does not mean that the statement is false if Mary is not reliable.
- Assume that probability of John's opinion about the reliability of Mary is given as 0.85. Then the probability of Mary to be unreliable for John is 0.15.
- So her statement justifies a 0.85 degree of belief that a John's car is stolen and John has no reason to believe that his car is not stolen so it is zero degree of belief that John's car is not stolen.

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 represented by ϕ. The theory of evidence assigns a belief mass to each subset of the power set.

A function m: P(U) $\rightarrow$ [0,1] is called a *basic belief assignment* (BBA) function. It satisfies the following axioms:

- $m(\phi) = 0$;
- $\sum m(A) = 1, \ \forall A \in P(U)$. The value of m(A) is called *mass assigned to A* on the unit interval.

It makes no additional claims about any subsets of A, each of which has, by definition, its own mass.

Dempster's Rule of Combination

The original combination rule, known as Dempster's rule of combination, is a generalization of Bayes' rule.

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 $m1(A) \neq 0$, and $m2(B) \neq 0$.
- The Dempster's rule for combining two belief functions to generate an *m3* function may be defined as:

$$m3(\phi) = 0$$

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

- This belief function gives new value when applied on the set C = A ∩ B.
- The combination of two belief functions is called the *joint mass*. Here m3 can also be written as (m1 o m2).
- The expression [$\sum_{A \cap B = \phi}$ (m1(A) * m2(B))] is called normalization factor. It is a measure of the amount of conflict between the two mass sets.
- The normalization factor has the effect of completely ignoring conflict and attributing any mass associated with conflict to the null set.

Example: Diagnostic System

Suppose we have mutually exclusive hypotheses represented by a set U = {flu, measles, cold, cough}. The goal is to assign or attach some measure of belief to the elements of U based on evidences.

It is not necessary that particular evidence is supporting some individual element of U but rather it may support subset of U. For example, an evidence of 'fever' might support {flu, measles}. So a belief function 'm' is defined for all subsets of U. The degree of belief to a set will keep on changing if we get more evidences supporting it or not.

- Initially assume that we have no information about how to choose hypothesis from the given set U.
- So assign m for U as 1.0 i.e., m(U) = 1.0. This means we are sure that answer is somewhere in the whole set U.
- 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.
- Let us define two belief functions m1 and m2 based on evidence of fever and on evidence of headache respectively as follows:

m1({flu, measles})	=	0.8
m1(U)	=	0.2
m2({flu, cold})	=	0.6
m2(U)	=	0.4

- We can compute their combination m3 using these values.

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

- Now previous belief functions are modified to m3 with the following belief values and are different from earlier beliefs.

m3({flu}) = 0.48

m3({flu, cold}) = 0.12

m3({flu, measles}) = 0.32

m3(U) = 0.08

- Further, if we have another evidence function m4 of sneezing with the belief values as:

m4({cold, cough}) = 0.7

m4(U) = 0.3

- Then the combination of m3 and m4 gives another belief function as follows:

Combination of m3 and m4	**m4({cold, cough}) = 0.7**	**m4(U) = 0.3**
m3({flu}) = 0.48	m5(ϕ) = 0.336	m5({flu}) = 0.114
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

- If we get 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 ($1 - \sum A \cap B = \phi \ (m1(A) * m2(B))$) in the denominator of belief values for non empty sets.

- From the table we get multiple belief values for empty set (ϕ) and its total belief value is 0.56.

- So according to formula, we have to scale down the remaining values of non empty sets by dividing by a factor (1 - 0.56 =0.44).

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

- While computing new belief we may get same subset generated from different intersection process. The ‘m’ value for such set is computed by summing all such values.