BAYESIAN CONCEPT LEARNING

- Principles of probability for classification are an important area of machine learning algorithms. In practical life, decisions are affected by prior knowledge or belief about an event.
- Thus, an event that is otherwise very unlikely to occur may be considered seriously to occur in certain situations if that is known in the past, the event had certainly occurred when other events were observed.
- The same concept is applied in machine learning using Bayes' theorem.
- Bayesian concept provides the basis for machine learning concepts.
- The technique 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 more importantly, how probabilities should be revised when there is additional information available.

IMPORTANCE OF BAYESIAN METHODS

- Bayesian learning algorithms, like the naive Bayes classifier, are highly practical approaches to certain types of learning problems as they can calculate explicit probabilities for hypotheses.
- In many cases, they are equally competitive or even outperform the other learning algorithms, including decision tree and neural network algorithms.

  Bayesian classifiers use a simple idea that the training data are utilized to calculate an observed probability of each class based on feature values.
- When the same classifier is used later for unclassified data, it uses the observed probabilities to predict the most likely class for the new features.
- The application of the observations from the training data can also be thought of as applying prior knowledge or prior belief to the probability of an outcome, so that it has higher probability of meeting the actual or real-life outcome.
- This simple concept is used in Bayes' rule and applied for training a machine in machine learning terms.

Some of the real-life uses of Bayesian classifiers are as follows:

1. Text-based classification such as spam or junk mail filtering, author identification, or topic categorization
2. Medical diagnosis such as given the presence of a set of observed symptoms during a disease, identifying the probability of new patients having the disease
3. Network security such as detecting illegal intrusion or anomaly in computer network

- One of the strengths of Bayesian classifiers is that they utilize all available parameters to subtly change the predictions, while many other algorithms tend to ignore the features that have weak effects.
- Bayesian classifiers assume that 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.

Some of the features of Bayesian learning methods that have made them popular are as follows:

1. Prior knowledge of the candidate hypothesis is combined with the observed data for arriving at the final probability of a hypothesis. So, two important components are the

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

2. 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.
3. 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.
4. Through the easy approach of Bayesian methods, it is possible to classify new instances by combining the predictions of multiple hypotheses, weighted by their respective probabilities.
5. In some cases, when Bayesian methods cannot compute the outcome deterministically, they can be used to create a standard for the optimal decision against which the performance of other methods can be measured

- The success of the Bayesian method largely depends on the availability of initial knowledge about the probabilities of the hypothesis set.
- So, if these probabilities are not known in advance, then some background knowledge, previous data or assumptions about the data set, and the related probability distribution functions must be used to apply this method.
- Moreover, it normally involves high computational cost to arrive at the optimal Bayes hypothesis.

BAYES' THEOREM

- In Bayes' theorem, standard probability calculus is used to determine the uncertainty about the function f, and the classification can be validated by feeding positive examples.
- Lets define a concept set C and a corresponding function f(k). and also define f(k) = 1, when k is within the set C and f(k) = 0 otherwise.
- The aim is to learn the indicator function f that defines which elements are within the set C.
- So, by using the function f, it could be able to classify the element either inside or outside our concept set.
- 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.

Prior

- The prior knowledge or belief about the probabilities of various hypotheses in H(Best Hypothesis) is called Prior in context of Bayes' theorem.

- For example, to determine whether a particular type of tumour is malignant for a patient, the prior knowledge of such tumours becoming malignant can be used to validate current hypothesis and is a prior probability or simply called Prior
- Assume that P(h) is the initial probability of a hypothesis 'h' that the patient has a malignant tumour based only on the malignancy test, without considering the prior knowledge of the correctness of the test process or the so-called training data.
- Similarly, P(T) is the prior probability that the training data will be observed or, in this case, the probability of positive malignancy test results.
- Denote P(T|h) as the probability of observing data T in a space where 'h' holds true, which means the probability of the test results showing a positive value when the tumour is actually malignant.

Posterior

- The probability that a particular hypothesis holds for a data set based on the Prior is called the posterior probability or simply Posterior.
- For example, the probability of the hypothesis that the patient has a malignant tumour considering the Prior of correctness of the malignancy test is a posterior probability.
- In the notation P(h|T), which means whether the hypothesis holds true given the observed training data T. This is called the posterior probability or simply Posterior in machine learning language.
- So, the prior probability P(h), which represents the probability of the hypothesis independent of the training data (Prior), now gets refined with the introduction of influence of the training data as P(h|T).
- According to Bayes' theorem

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

combines the prior and posterior probabilities together.

- From the above equation, It can be deduced that P(h|T) increases as P(h) and P(T|h) increases and also as P(T) decreases.
- When there is more probability that T can occur independently of h then it is less probable that h can get support from T in its occurrence.
- Finding 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:

$$\begin{aligned} h_{\text{MAP}} &= \text{argmax}_{h \in H} P(h|T) \\ &= \text{argmax}_{h \in H} \frac{P(T|h)P(h)}{P(T)} \end{aligned}$$

- and as P(T) is a constant independent of h, in this case, can be written as

$$= \text{argmax}_{h \in H} P(T|h)P(h)$$

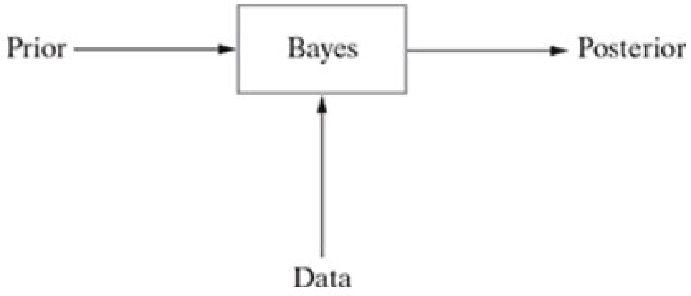
6.1

Likelihood

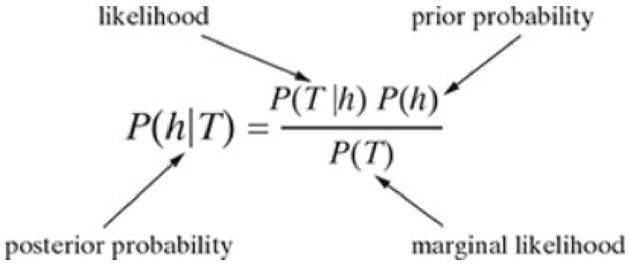
In certain machine learning problems, Equation 6.1 can be further simplified as if every hypothesis in H has equal probable priori as P(h) = P(h), and then, P(h|T) determined from the probability P(T|h) only.

Thus,P(T|h) is called the likelihood of data T given h, and any hypothesis that maximizes P(T|h) is called the **maximum likelihood (ML) hypothesis, h** .

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



Bayes Theorem



BAYES' THEOREM AND CONCEPT LEARNING

One simplistic view of concept learning can be that if feed the machine can be fed with the training data, then it can calculate the posterior probability of the hypotheses and outputs the most probable hypothesis. This is also called brute-force Bayesian learning algorithm, and it is also observed that consistency in providing the right probable hypothesis by this algorithm is very comparable to the other algorithms.

Brute-force Bayesian algorithm

- Using the MAP hypothesis output to design a simple learning algorithm called **brute-force map learning algorithm.**
- Assume that the learner considers a finite hypothesis space H in which the learner try to learn some target concept
  c:X → {0,1} where X is the instance space corresponding to H.
- The sequence of training examples is $\{(x_1 , t_1), (x_2 ,t_2),..., (x_m , t_m)\}$, where x is the instance of X and t_i is the target concept of x_i defined as $t_i= c(x_i)$.

- Without impacting the efficiency of the algorithm, assume that the sequence of instances of x {x_1 ,..., x_m } is held fixed, and then, the sequence of target values becomes T = {t_1 ,..., t_m }.
- For calculating the highest posterior probability, Bayes' theorem can be used

Concept of consistent learners

- The behavior of the general class of learner where , 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' theorem can characterize the behaviour of learning algorithms even when the algorithm does not explicitly manipulate the probability.
- As it can help to identify the optimal distributions of P(h) and P(T|h) under which the algorithm outputs the MAP hypothesis,
- the knowledge can be used to characterize the assumptions under which the algorithms behave optimally.
- There is a special case of Bayesian output which corresponds to the noise-free training data and deterministic predictions of hypotheses where P(T|h) takes on value of either 1 or 0, the theorem can be used with the same effectiveness for noisy training data and additional assumptions about the probability distribution governing the noise.

Bayes optimal classifier

- With the use of the MAP hypothesis , the most probable classification of the new instance given the training data can be found.
- To illustrate the concept, let us assume three hypotheses h_1 , h_2 , and h_3 in the hypothesis space H.
- Let the posterior probability of these hypotheses be 0.4, 0.3, and 0.3, respectively.
- There is a new instance x, which is classified as true by h_1 , but false by h_2 , and h_3.
- Then the most probable classification of the new instance (x) can be obtained by combining the predictions of all hypotheses weighed by their corresponding posterior probabilities.
- By denoting the possible classification of the new instance as c_i from the set C, the probability P(c_i |T) that the correct classification for the new instance is c_i is

$$P(c_i|T) = \sum_{h_i \in H} P(c_i|h_i)P(h_i|T)$$

The optimal classification is for which $P(c_i|T)$ is maximum is

$$\text{Bayes optimal classifier} = \underset{c_i \in C}{\overset{\text{argmax}}{}} \sum_{h_i \in H} P(c_i|h_i)P(h_i|T)$$

- The approach in the Bayes optimal classifier is to calculate the most probable classification of each new instance on the basis of the combined predictions of all alternative hypotheses, weighted by their posterior probabilities

Naïve Bayes classifier

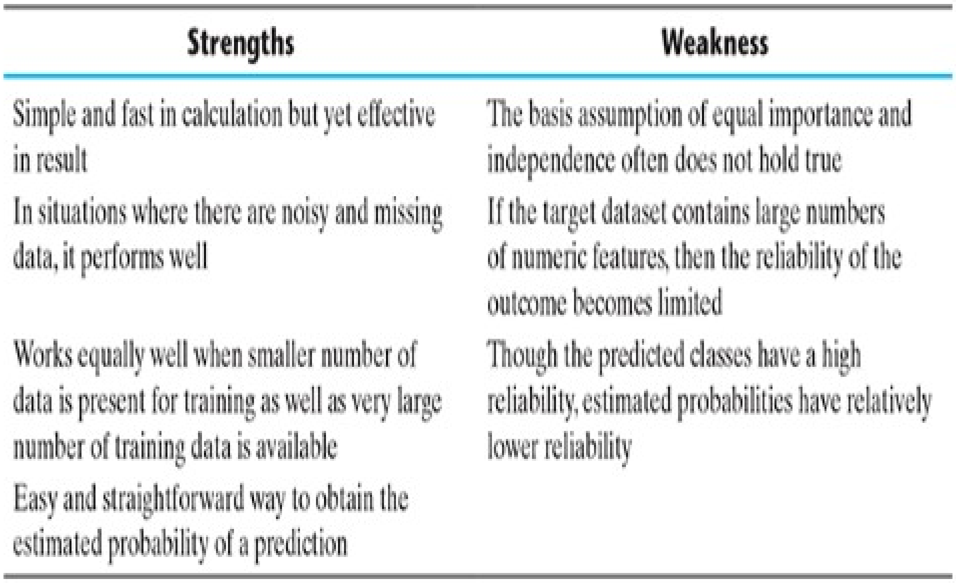
- Naïve Bayes is a simple technique for building classifiers: models that assign class labels to problem instances.
- The basic idea of Bayes rule is that the outcome of a hypothesis can be predicted on the basis of some evidence (E) that can be observed.
  From Bayes rule,
  1. A prior probability of hypothesis h or P(h): This is the probability of an event or before the evidence is observed.
  2. A posterior probability of h or P(h|D): This is the probability of an event after the evidence is observed within the population D.

$$\text{Posterior probability} = \frac{(\text{Prior probability} \times \text{Conditional Probability})}{\text{Evidence}}$$

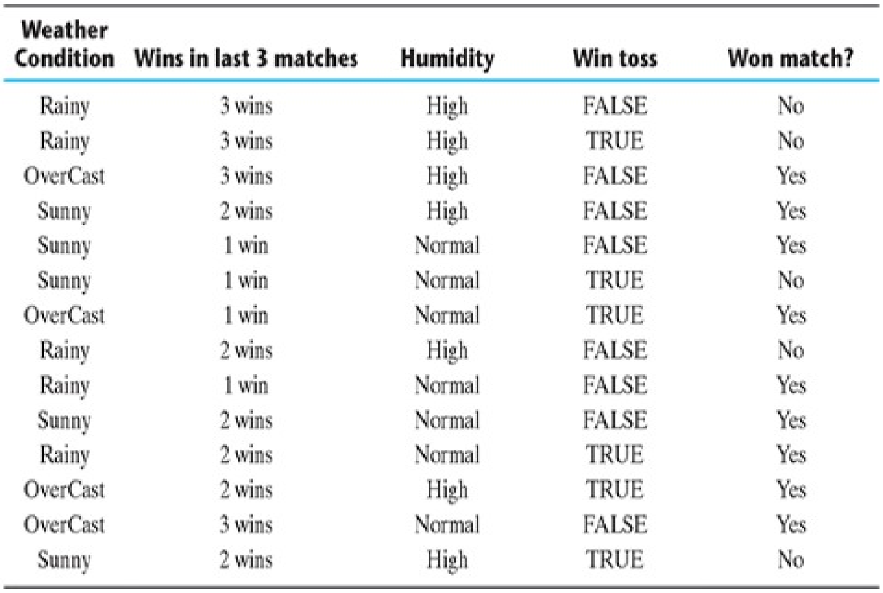
- For example, a person has height and weight of 182 cm and 68 kg, respectively, then the probability that this person belongs to the class ‘basketball player’ can be predicted using the Naïve Bayes classifier. This is known as probabilistic classifications.
- In machine learning, a probabilistic classifier is a classifier that can be foreseen, given a perception or information (input), a likelihood calculation over a set of classes, instead of just yielding (outputting) the most likely class that the perception (observation) should belong to.
- Parameter estimation for Naïve Bayes models uses the method of ML.
- Bayes’ theorem is used when new information can be used to revise previously determined probabilities
- Depending on the particular nature of the probability model, Naïve Bayes classifiers can be trained very professionally in a supervised learning setting.
- A Naïve Bayes classifier is a primary probabilistic classifier based on a view of applying Bayes’ theorem independence assumptions.
- The prior probabilities in Bayes’ theorem that are changed with the help of newly available information are classified as posterior probabilities.
- A key benefit of the naive Bayes classifier is that it requires only a little bit of training information (data) to gauge the parameters (mean and differences of the variables) essential for the classification (arrangement).
- In the Naïve Bayes classifier, independent variables are always assumed, and only the changes (variances) of the factors/variables for each class should be determined and not the whole covariance matrix.
- Because of the rather naïve assumption that all features of the dataset are equally important and independent, this is called Naïve Bayes classifier.
- Naïve Bayes classifiers are direct linear classifiers that are known for being the straightforward, yet extremely proficient result. The modified version of Naïve

Bayes classifier originates from the assumption that information collection (data set) is commonly autonomous (mutually independent).

Strengths and Weaknesses of Bayes Classifiers



Strengths	Weakness
Simple and fast in calculation but yet effective in result	The basis assumption of equal importance and independence often does not hold true
In situations where there are noisy and missing data, it performs well	If the target dataset contains large numbers of numeric features, then the reliability of the outcome becomes limited
Works equally well when smaller number of data is present for training as well as very large number of training data is available	Though the predicted classes have a high reliability, estimated probabilities have relatively lower reliability
Easy and straightforward way to obtain the estimated probability of a prediction	



Weather Condition	Wins in last 3 matches	Humidity	Win toss	Won match?
Rainy	3 wins	High	FALSE	No
Rainy	3 wins	High	TRUE	No
OverCast	3 wins	High	FALSE	Yes
Sunny	2 wins	High	FALSE	Yes
Sunny	1 win	Normal	FALSE	Yes
Sunny	1 win	Normal	TRUE	No
OverCast	1 win	Normal	TRUE	Yes
Rainy	2 wins	High	FALSE	No
Rainy	1 win	Normal	FALSE	Yes
Sunny	2 wins	Normal	FALSE	Yes
Rainy	2 wins	Normal	TRUE	Yes
OverCast	2 wins	High	TRUE	Yes
OverCast	3 wins	Normal	FALSE	Yes
Sunny	2 wins	High	TRUE	No

Training data for the Naïve Bayesian method

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 for ‘Won match = Yes’ and the probability for ‘Won match = No’ on the basis of all the attributes. Otherwise, simply multiply probabilities of all favourable conditions to derive ‘YES’ condition. Multiply probabilities of all non-favourable conditions to derive ‘No’ condition.

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

Applications of Naïve Bayes classifier

Text classification:

- Naïve Bayes classifier is among the most successful known algorithms for learning to classify text documents.
- It classifies the document where the probability of classifying the text is more.
- It uses the algorithm to check the permutation and combination of the probability of classifying a document under a particular 'Title'.
- It has various applications in document categorization, language detection, and sentiment detection, which are very useful for traditional retailers, e-retailors, and other businesses on judging the sentiments of their clients on the basis of keywords in feedback forms, social media comments, etc.

Spam filtering:

- Spam filtering is the best known use of Naïve Bayesian text classification.
- Presently, almost all the email providers have this as a built-in functionality,which makes use of a Naïve Bayes classifier to identify spam email on the basis of certain conditions and also the probability of classifying an email as 'Spam'.
- Naïve Bayesian spam sifting has turned into a mainstream mechanism to recognize illegitimate a spam email from an honest-to-goodness email (sometimes called 'ham'). Users can also install separate email filtering programmes. Server-side email filters such as DSPAM, Spam Assassin, Spam Bayes, and ASSP make use of Bayesian spam filtering techniques, and the functionality is sometimes embedded within the mail server software itself.

Hybrid Recommender System:

- It uses Naïve Bayes classifier and collaborative filtering. Recommender systems (used by e-retailors like eBay, Alibaba, Target, Flipkart, etc.) apply machine learning and data mining
  techniques for filtering unseen information and can predict whether a user would like a given resource.
- For example, when we log in to these retailer websites, on the basis of the usage of texts used by the login and the historical data of purchase, it automatically recommends the product for the particular login persona.
- One of the algorithms is combining a Naïve Bayes classification approach with collaborative filtering, and experimental results show that this algorithm provides better performance regarding accuracy and coverage than other algorithms.

Online Sentiment Analysis:

- The online applications use supervised machine learning (Naïve Bayes) and useful computing.
- In the case of sentiment analysis, let us assume there are three sentiments such as nice, nasty, or neutral, and Naïve Bayes classifier is used to distinguish between them. Simple emotion modeling combines a statistically based classifier with a dynamical model.
- The Naïve Bayes classifier employs 'single words' and 'word pairs' like features and determines the sentiments of the users.
- It allocates user utterances into nice, nasty, and neutral classes, labelled as +1, −1, and 0, respectively.
- This binary output drives a simple first- order dynamical system, whose emotional state represents the simulated emotional state of the experiment's personification.

- **BAYESIAN BELIEF NETWORK**

  A significant assumption in the Naïve Bayes classifier was that the attribute values a_1 , a_2 ,..., a_n are conditionally independent for a target value.The Naïve Bayes classifier generates optimal output when this condition is met.
- Though this assumption significantly reduces the complexity of computation, in many practical scenarios, this requirement of conditional independence becomes a difficult constraint for the application of this algorithm.
- The approach of Bayesian Belief network, which assumes that within the set of attributes, the probability distribution can have conditional probability relationship as well as conditional independence assumptions.
- This is different from the Naïve Bayes assumption of conditional independence of all the attributes as the belief network provides the flexibility of declaring a subset of the attributes as conditionally dependent while leaving rest of the attributes to hold the assumptions of conditional independence.

Independence and conditional independence

- The conditional probability of A with knowledge of K can be represented as P(A|K).
- The variables A and K are said to be independent if P(A|K) = P(A), which means that there is no influence of K on the uncertainty of A.
- Similarly, the joint probability can be written as P(A,K) =P(A)P(K).
- Extending this concept, the variables A and K are said to be conditionally independent given C if P(A|C) = P(A|K, C).
- This concept of conditional independence can also be extended to a set of attributes.
- Set of variables A_1 , A_2 ,..., A_n is conditionally independent of the set of variables B_1 , B_2 ,..., B_m given the set of variables C_1 ,C_2 ,..., C_1 if

$$P(A_1, A_2,..., A_n|B_1, B_2,..., Bm, C_1, C_2,..., C_1) = P(A_1, A_2,..., A_n|C_1, C_2,..., C_l)$$

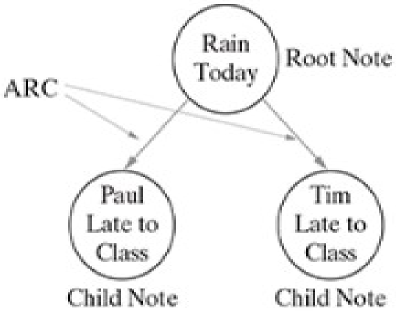
- There can be two main scenarios faced in the Bayesian Belief network learning problem.
- First, the network structure might be available in advance or can be inferred from the training data.

- Second, all the network variables either are directly observable in each training example or some of the variables may be unobservable.
- Learning the conditional probability tables is a straightforward problem when the network structure is given in advance and the variables are fully observable in the training examples.
- But in the case the network structure is available and only some of the variable values are observable in the training data, then the learning problem is more difficult.

The Bayesian Belief network can represent much more complex scenarios with dependence and independence concepts. There are three types of connections possible in a Bayesian Belief network.

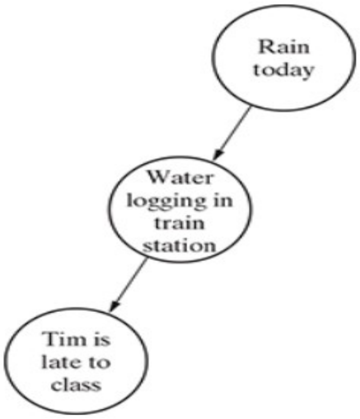
Diverging Connection:

In this type of connection, the evidence can be transmitted between two child nodes of the same parent provided that the parent is not instantiated.



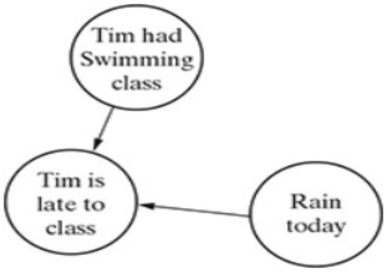
Serial Connection:

In this type of connection, any evidence entered at the beginning of the connection can be transmitted through the directed path provided that no intermediate node on the path is instantiated



Converging Connection:

In this type of connection, the evidence can only be transmitted between two parents when the child (converging) node has received some evidence and that evidence can be soft or hard


Tim had Swimming class
Tim is late to class
Rain today