

(Classification Alternative Techniques)

(1) Bayesian classifier

(a) Bayes Theorem

(b) Using Bayes Theorem for classification

(c) Naïve Bayes Classifier

(d) Bayes Error Rate

(2) Bayesian Belief Networks

(a) Model Representation

(b) Model Building

BAYESIAN CLASSIFIERS

(1)

In Many Applications the relationship between the attribute set and the class variable is Non-deterministic.

In other words, the class label of a test record can not be predicted with certainty, even though its attribute set is identical to some of training Examples.

This Situation may arise because of noisy data or the presence of certain confounding factors that affect classification but are not included in the Analysis.

It presents an Approach for modeling probabilistic relationships between the attribute set and class Variable.

* Bayes Theorem :-

Let X and Y be a pair of random variables. Their Joint probability, $P(X=x, Y=y)$, refers to the probability that variable X will take on the value x and Variable Y will take on the value y .

A conditional probability is the probability that a random variable will take on a particular value given that the outcome for another random variable is known.

For Example, the conditional probability $P(Y=y|X=x)$ refers to probability that the variable Y will take on the value y , given that the variable X is observed to have value x .

The Joint and conditional probabilities for X and Y are related in the following way:

$$P(X, Y) = P(Y|X) \times P(X) = P(X|Y) \times P(Y)$$

Rearranging this Equation, leads to the following formula, known as Bayes theorem:

$$P(Y|X) = \frac{P(X|Y) P(Y)}{P(X)}$$

Problem:-

Consider a football game between two rival teams: Team '0' and Team '1'. Suppose Team '0' wins 65% of the time and team '1' wins remaining matches. Among the games won by Team '0', only 30% of them come from playing on Team 1's football field. On the other hand, 75% of the victories for Team 1 are obtained while playing at home. If Team 1 is to host the next match between two teams, which team will most likely emerge as the winner?

Solution:- * Let X be the random variable that represents the team hosting the match.

* Let Y be the random variable that represents the winner of the Match.

* Both X and Y can take on values from set $\{0, 1\}$.

We can Summarize the information given in problem as follows:

probability Team 'o' wins is $P(Y=0) = 0.65$

probability Team 'i' wins is $P(Y=1) = 1 - P(Y=0) = 0.35$

probability Team 'i' hosted the Match, it won is:

$$P(X=1|Y=1) = 0.75$$

probability Team 'i' hosted the Match, won by Team 'o' is:

$$P(X=1|Y=0) = 0.3$$

our objective is to compute $P(Y=1|X=1)$, which is conditional probability that Team 'i' wins the next match it will be hosting, and compares it against $P(Y=0|X=1)$.

Using the Bayes Theorem, we obtain:

$$\begin{aligned} P(Y=1|X=1) &= \frac{P(X=1|Y=1) \times P(Y=1)}{P(X=1)} \\ &= \frac{P(X=1|Y=1) \times P(Y=1)}{P(X=1, Y=1) + P(X=1, Y=0)} \\ &= \frac{P(X=1|Y=1) \times P(Y=1)}{P(X=1|Y=1) \times P(Y=1) + P(X=1|Y=0) \times P(Y=0)} \\ &= \frac{0.75 \times 0.35}{0.75 \times 0.35 + 0.3 \times 0.65} = 0.5738. \end{aligned}$$

$$P(Y=0|X=1) = 1 - P(Y=1|X=1) = 0.4262.$$

~~Since~~ Since $P(Y=1|X=1) > P(Y=0|X=1)$,

Team 'i' has a better chance than Team 'o' of winning Next Match.

* Naive Bayes Classifier

How

A Naive Bayes classifier estimates the conditional probability by assuming that the attributes are conditionally independent, given the class label 'y'.

The conditional independence assumption can be formally stated as follows:

$$P(X|Y=y) = \prod_{i=1}^d P(X_i|Y=y),$$
 where each attribute set $X = \{X_1, X_2, \dots, X_d\}$ consists of 'd' attributes.

Conditional Independence :-

Let X, Y, Z denote three sets of random variables. The variables in X are said to be conditionally independent of Y , given Z , if the following condition holds:

$$P(X|Y, Z) = P(X|Z).$$

An Example of conditional independence is the relationship between a person's arm length and his reading skills.

The conditional independence between X and Y can also be written into a form :

$$\begin{aligned} P(X, Y|Z) &= \frac{P(X, Y, Z)}{P(Z)} \\ &= \frac{P(X, Y, Z)}{P(Y, Z)} \times \frac{P(Y, Z)}{P(Z)} \\ &= P(X|Y, Z) \times P(Y|Z) \\ &= P(X|Z) \times P(Y|Z). \end{aligned}$$

How a Naive Bayes Classifier Works:-

With the conditional independence assumption, instead of computing the class-conditional probability for every combination of X , we only have to estimate the conditional probability of each X_i , given Y .

To classify a test record, the naive Bayes classifier computes the posterior probability for each class Y :

$$P(Y|X) = \frac{P(Y) \prod_{i=1}^d P(X_i|Y)}{P(X)}$$

Since $P(X)$ is fixed for every Y , it is sufficient to choose the class that maximizes the numerator term,

$$P(Y) \cdot \prod_{i=1}^d P(X_i|Y).$$

Estimating Conditional Probabilities for Categorical Attributes:-

For a categorical attribute X_i , the conditional probability $P(X_i = x_i | Y = y)$ is estimated according to the fraction of training instances in class 'y' that take on a particular attribute value x_i .

For Example, in the training set given in table, two out of the four people who repaid their loans also own a home.

As a result, the conditional probability for $P(\text{HomeOwner} = \text{yes} | \text{no})$ is $= 2/4$.

Estimating Conditional Probabilities for Continuous Attributes:

Similar
varia

There are two ways to estimate the class-conditional probabilities for continuous attributes in Naive Bayes classifiers:

1. We can discretize each continuous attribute and then replace the continuous attribute value with its corresponding discrete interval.

This approach transforms the continuous attributes into ordinal attributes.

The conditional probability $P(X_i | Y = y)$ is estimated by computing the fraction of training records belongs to class 'y' that falls within the corresponding interval for X_i .

2. A Gaussian distribution is usually chosen to represent the class-conditional probability for continuous attributes.

The distribution is characterized by two parameters, its mean, μ , and variance, σ^2 .

For each class y_j , the class-conditional probability for attribute X_i is

$$P(X_i = x_i | Y = y_j) = \frac{1}{\sqrt{2\pi}\sigma_{ij}} \exp \left(-\frac{(x_i - \mu_{ij})^2}{2\sigma_{ij}^2} \right)$$

The parameter μ_{ij} can be estimated based on sample mean of X_i ($\bar{x}$) for all training records that belong to the class y_j .

Similarly, σ_{15}^2 can be estimated from the Sample variance (s^2) of such training records.

Example of Naive Bayes Classifier :-

Consider the dataset, shown in the following table, we can compute the class conditional probability for each categorical attribute, along with sample mean and variance for continuous attributes. These probabilities are summarized below:

TID	Home owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125 K	No
2	No	Married	100K	No
3	No	Single	70 K	No
4	Yes	Married	120 K	No
5	No	Divorced	95 K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220 K	No
8	No	Single	85 K	Yes
9	No	Married	75 K	No
10	No	Single	90K	Yes

$$P(\text{Homeowner} = \text{yes} | \text{no}) = 3/7$$

$$P(\text{Homeowner} = \text{no} | \text{no}) = 4/7$$

$$P(\text{Homeowner} = \text{yes} | \text{yes}) = 0$$

$$P(\text{Homeowner} = \text{no} | \text{yes}) = 1$$

$$P(\text{Marital Status} = \text{Single} | \text{no}) = 4/7$$

$$P(\text{Marital Status} = \text{Divorced} | \text{no}) = 1/7$$

$$P(\text{Marital Status} = \text{Married} | \text{no}) = 4/7$$

$$P(\text{Marital Status} = \text{Single} | \text{yes}) = 2/3$$

$$P(\text{Marital Status} = \text{Divorced} | \text{yes}) = 1/3$$

$$P(\text{Marital Status} = \text{Married} | \text{yes}) = 0$$

For Annual Income:

If class = No: Sample mean = 110

Sample Variance = 2975

If class = Yes: Sample mean = 90

Sample Variance = 25

To predict the class label of a test record $X = (\text{Homeowner} = \text{no}, \text{Marital Status} = \text{married}, \text{Income} = 120K)$, we need to compute the posterior probabilities $P(\text{No} | X)$ and $P(\text{Yes} | X)$.

$$\begin{aligned} & (21-2)^2 + \\ & \Rightarrow (95-110)^2 + (85-110)^2 + (90-110)^2 \\ & \Rightarrow 115^2 + (25)^2 + (20)^2 \\ & \Rightarrow 1325 + 625 = 1950 \end{aligned}$$

The prior probabilities of each class can be estimated by calculating the fraction of training records that belong to each class. Since there are three records that belong to the class 'Yes' and seven records that belong to the class 'No', $P(\text{Yes}) = 0.3$; $P(\text{No}) = 0.7$

By using the above table information, the class-conditional probabilities can be computed as follows:

$$\begin{aligned} P(X|\text{No}) &= P(\text{HomeOwner}=\text{No}|\text{No}) \times P(\text{Status}=\text{Married}|\text{No}) \\ &\quad \times P(\text{Annual Income} = 120\text{K}|\text{No}) \\ &= \frac{4}{7} \times \frac{4}{7} \times 0.0072 = 0.0024. \end{aligned}$$

$$\begin{aligned} P(X|\text{Yes}) &= P(\text{HomeOwner}=\text{No}|\text{Yes}) \times P(\text{Status}=\text{Married}|\text{Yes}) \\ &\quad \times P(\text{Annual Income} = 120\text{K}|\text{Yes}) \\ &= 1 \times 0 \times 1.2 \times 10^{-9} = 0 \end{aligned}$$

* The posterior probability for class 'No' is $P(\text{No}|X) = \alpha \times 7/10 \times 0.0024 = 0.0016\alpha$, where $\alpha = 1/P(X)$ is a constant term.

* The posterior probability for class 'Yes' = 0.

Since $P(\text{No}|X) > P(\text{Yes}|X)$, the record is classified as 'No'.

Characteristics of Naive Bayes Classifiers :-

Naive Bayes classifiers, generally have the following characteristics:

- They are robust to isolated noise points, because such points are averaged out when estimating conditional probabilities from data.
- They are robust to irrelevant attributes. If X_i is an irrelevant attribute, then $P(X_i|Y)$ becomes almost uniformly distributed.
- Correlated attributes can degrade the performance of Naive Bayes classifiers.

* Bayes Error Rate

Suppose we know the true probability distribution that governs $P(X|Y)$. The Bayesian classification method allows us to determine the ideal decision boundary for the classification task, as given in the following Example:

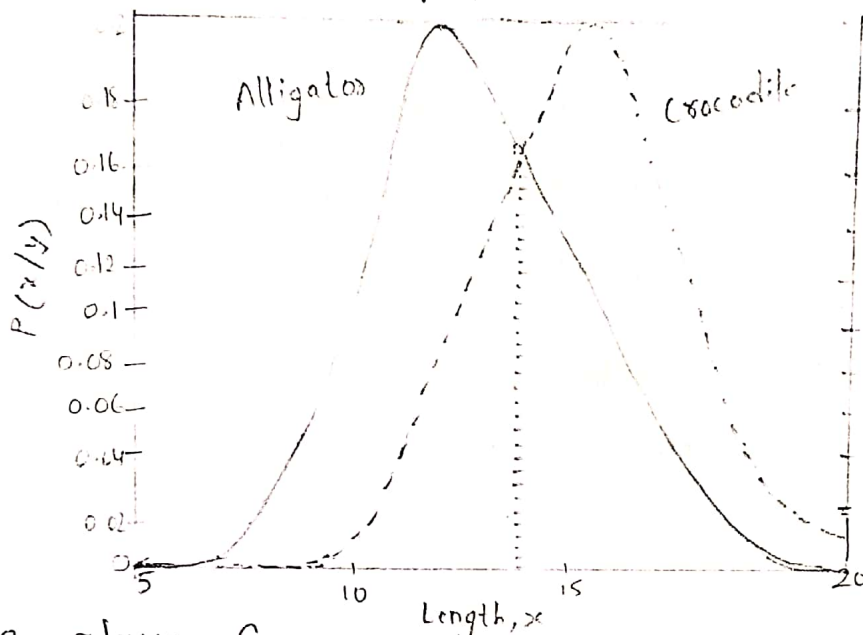
Ex:- consider the task of identifying Alligators and Crocodiles based on their respective lengths.

The Average length of an Adult crocodile is about 15 feet, while the average length of an Adult Alligator is about 12 feet.

Assuming that their length 'x' follows a Gaussian distribution with a standard deviation equal to 2, we can express their class-conditional probabilities as follows

$$P(X|\text{Crocodile}) = \frac{1}{\sqrt{2\pi} \cdot 2} \exp \left[-\frac{1}{2} \left(\frac{X-15}{2} \right)^2 \right] \rightarrow \textcircled{1}$$

$$P(X|\text{Alligator}) = \frac{1}{\sqrt{2\pi} \cdot 2} \exp \left[-\frac{1}{2} \left(\frac{X-12}{2} \right)^2 \right] \rightarrow \textcircled{2}$$



The above figure shows a comparison between the class-conditional probabilities for a Crocodile and an Alligator.

Assuming that their prior probabilities are same, the ideal decision boundary is located at some length $\hat{x}$ such that

$$P(X = \hat{x}|\text{Crocodile}) = P(X = \hat{x}|\text{Alligator})$$

Using Equations $\textcircled{1}$ & $\textcircled{2}$, we obtain: $\left(\frac{\hat{x}-15}{2} \right)^2 = \left(\frac{\hat{x}-12}{2} \right)^2$,

which can be solved to yield $\hat{x} = 13.5$.

* The decision boundary for this Example is located halfway between the two means.

⑦

The ideal decision boundary in this Example classifies all creatures whose lengths are less than $\hat{x}$ as Alligators and those whose lengths are greater than $\hat{x}$ as crocodiles.

The Error rate of the classifier is given by the sum of the area under posterior probability curve for crocodiles (from length 0 to $\hat{x}$) and the area under posterior probability curve for Alligators (from $\hat{x}$ to ∞):

$$\text{Error} = \int_0^{\hat{x}} P(\text{Crocodile} | x) dx + \int_{\hat{x}}^{\infty} P(\text{Alligator} | x) dx.$$

The total Error rate is known as Bayes Error Rate.

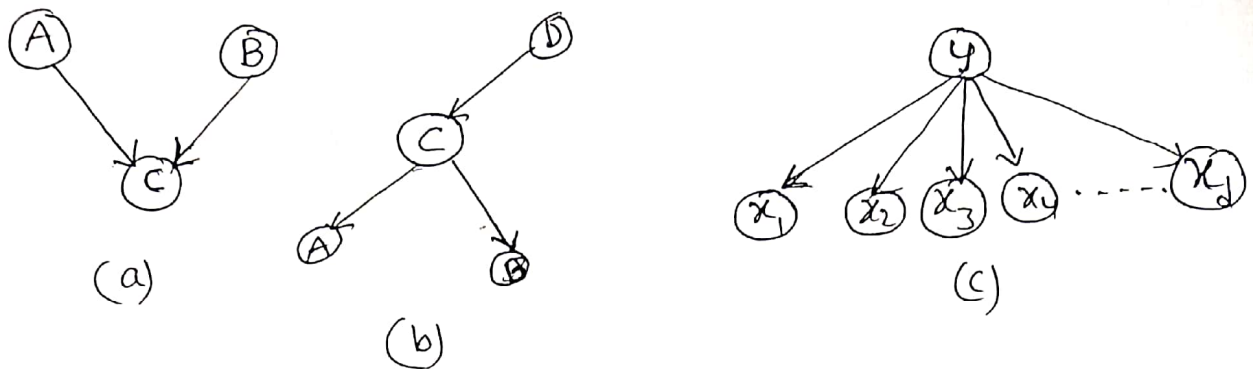
* Bayesian Belief Networks

① Model Representation:-

A Bayesian belief Network (BBN), or simply, Bayesian network, provides a graphical representation of the probabilistic relationships among a set of random variables. There are two key elements of a Bayesian network:

1. A Directed Acyclic Graph (dag) encoding the dependence relationships among a set of variables.

2. A probability table associating each node to its immediate parent nodes.



Consider 3 random variables, A, B, C in which A and B are independent variables and each has a direct influence on a third variable C . The relationships among the variables can be summarized into the directed acyclic graph as shown in fig: (a).

Each node in the graph represents a variable, and each arc asserts the dependence relationship between the pair of variables.

If there is a directed arc from X to Y , then X is parent of Y and Y is child of X .

If there is a directed path in the network from X to Z , then X is an Ancestor of Z , while Z is descendent of X .

From fig(b): A is descendent of D & D is an Ancestor of B . Both B and D are Non-descendants of A .

An important property of the Bayesian network can be stated as follows:

conditional Independence:- A Node in a Bayesian network is conditionally independent of its Non-descendants, if its parents are known.

From the above fig (b); A is conditionally independent of Both B and D given C because the nodes for B and D are non-descendants of node A.

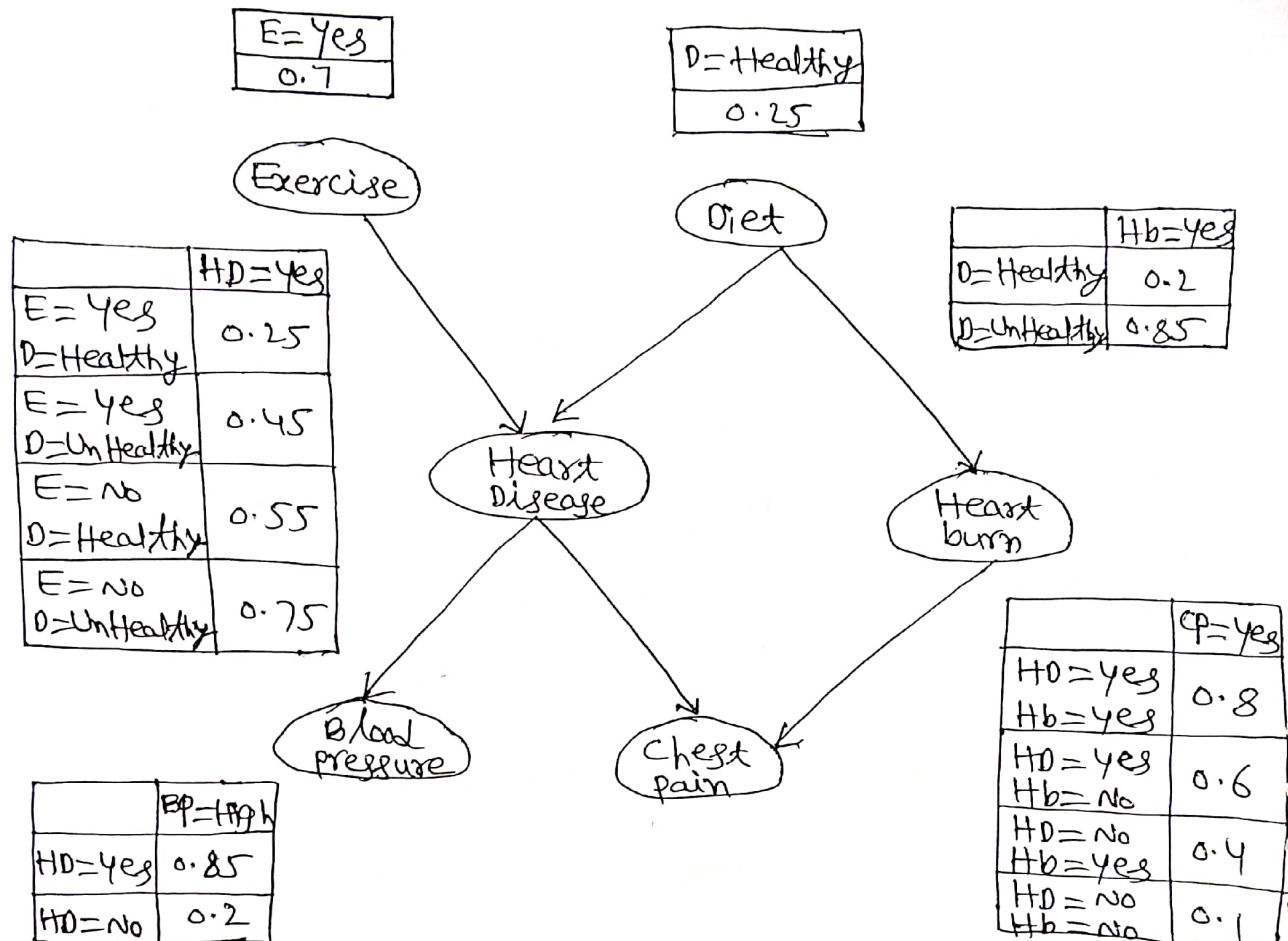
The conditional independence assumption made by a naive Bayes classifier can also be represented using a Bayesian network, as shown in fig (c).

where y is target class and $\{x_1, x_2, \dots, x_n\}$ is attribute set

* Besides the conditional independence condition, each node is also associated with a probability table

1. If a Node x does not have any parents, then the table contains only the prior probability $P(x)$.
2. If a Node x has only one parent y , then the table contains the conditional probability $P(x/y)$.
3. If a node x has multiple parents, $\{y_1, y_2, \dots, y_k\}$, then the table contains Conditional probability:
$$P(x/y_1, y_2, \dots, y_k).$$

The following figure shows An Example of Bayesian network for modeling patients with heart disease or heartburn problems.



Each variable in the diagram is assumed to be binary valued. The parent nodes for Heart disease (HD) correspond to risk factors that may affect the disease, such as Exercise (E) and Diet (D).

The child nodes for Heart Disease correspond to Symptoms of the disease, such as chest pain (CP) and high blood pressure (BP).

For Example, the diagram shows that Heartburn(Hb) may result from an unhealthy diet and may lead to chest pain.

The nodes for Heart Disease, Heartburn, and their corresponding symptoms contain conditional probability. Some of the probabilities have been omitted from diagram.

The omitted probabilities can be recovered by noting that $P(X=\bar{x}) = 1 - P(X=x)$ and

$P(X=\bar{x}|Y) = 1 - P(X=x|Y)$, where $\bar{x}$ denotes opposite outcome of x .

For Example, the conditional probability

$$\begin{aligned} &P(\text{HeartDisease}=\text{no} | \text{Exercise}=\text{no}, \text{Diet}=\text{Healthy}) \\ &= 1 - P(\text{HeartDisease}=\text{yes} | \text{Exercise}=\text{no}, \text{Diet}=\text{Healthy}) \\ &= 1 - 0.55 = 0.45, \end{aligned}$$

② Model Building :-

Model Building in Bayesian networks involves two steps: (1) creating structure of network, (2) estimating probability values in the tables associated with each node.

The network topology can be obtained by encoding the subjective knowledge of domain experts.

The following algorithm presents a systematic procedure for inducing the topology of a Bayesian network:

Algorithm:-

- 1: Let $T = (x_1, x_2, \dots, x_d)$ denote a total order of the variables.
- 2: for $j=1$ to d do
- 3: Let $x_{T(j)}$ denote the j^{th} highest order variable in T .
- 4: Let $\pi(x_{T(j)}) = \{x_{T(1)}, x_{T(2)}, \dots, x_{T(j-1)}\}$ denote set of variables preceding $x_{T(j)}$.
- 5: Remove the variables from $\pi(x_{T(j)})$ that do not affect x_j .
- 6: create an Arc between $x_{T(j)}$ and remaining variables in $\pi(x_{T(j)})$.
- 7: end for.

Example:- consider the variables shown in above figure:

- After performing Step i, let us assume the variables are ordered in the following way: (E, D, HD, Hb, CP, BP) .
- from Step 2 to 7, starts with variable D , we obtain the following conditional probabilities:
 - $P(D|E)$ is simplified to $P(D)$.
 - $P(HD|E, D)$ can not be simplified.
 - $P(Hb|HD, E, D)$ is simplified to $P(Hb|D)$.
 - $P(CP|Hb, HD, E, D)$ is simplified to $P(CP|Hb, HD)$.
 - $P(BP|CP, Hb, HD, E, D)$ is simplified to $P(BP|HD)$.
- Based on these probabilities, we can create Arcs between the nodes (E, HD) , (D, HD) , (D, Hb) , (HD, CP) , (Hb, CP) and (HD, BP) . These Arcs result in network structure as shown in above figure.