UNIT – 4 :: Classification : Alternative Techniques

Classification: Alterative Techniques, Bayes' Theorem, Naïve Bayesian Classification, Bayesian Belief Networks.

Introduction: Alternative Techniques are using to build classification models. This includes rule-based and nearest neighbor classifier techniques.

5.1. Rule-Based Classifier: A rule-based classifier is a technique for classifying the records using a collection of "**if . . . . then . .**" rules.

→The rules for the model are represented in a disjunctive normal form $R = (r_1 \vee r_2 \vee \ldots r_k)$
→ In this R is known as the rule set and r_i is classification rule (or) disjuncts.

→Each classification rule can be expressed as $r_i = (\text{Condition}_i) \rightarrow y_i$
→ Where r_i is called "rule antecedent" or "precondition".
→ Where condition $_i = (A_1 \text{ op } v_1) \wedge (A_2 \text{ op } v_2) \wedge \ldots\ldots \wedge (A_k \text{ op } v_k)$
→ Where $A_1, A_2, \ldots$ are attributes.
→ Where $v_1, v_2, \ldots$ are data of attribute.
→ Where op is a relational operator such as $>, <, \geq, \leq, =, \neq$

This is illustrated using the following table and that can be classified by using the given rules.

Name	Body Temperature	Gives Birth	Aquatic Creature	Aerial Creature	Has legs	Class label
Python	cold- blooded	no	no	no	no	reptile
Salmon	cold- blooded	no	yes	no	no	fish
Whale	warm- blooded	yes	yes	no	no	mammal
Frog	cold- blooded	no	semi	no	yes	amphibian
Bat	warm- blooded	yes	no	no	yes	mammal
Pigeon	warm- blooded	no	no	yes	yes	bird
Cat	warm- blooded	yes	yes	no	yes	mammal
penguin	warm- blooded	no	semi	yes	no	bird
salamander	cold- blooded	no	semi	no	yes	amphibian

Application of Rule-Based Classifier:

- A rule r trigger an instance x if the attributes of the instance satisfy the condition of the rule

r1 : (Gives birth = no) ∧ (Aerial creature = yes) →birds
r2 : (Gives birth = no) ∧ (Aquatic creature = yes) →fishes
r3 : (Gives birth = yes) ∧ (Body temperature = warm blooded) →mammals
r4 : (Gives birth = no) ∧ (Aerial creature = no) →Reptiles
r5 : (Aquatic creature = semi) → Amphibian

5.1.1: Quality of classification rule: The quality of a classification rule can be evaluated using the measures such as coverage and accuracy.
→ Given a data set D and a Classific rule r : A → y
→ The coverage of the rule is defined as the fraction of records in D that trigger the rule r
→ The formal definition of these measures are

$$\text{Coverage}(r) = \frac{|A|}{|D|}$$

$$\text{Accuracy}(r) = \frac{|A \cap y|}{|A|}$$

Sreenivaas - 9948808818

→ Where | A | is the number of records that satisfy the rule antecedent.
→ Where | A ∩ y | is the number of records that satisfy both antecedent and consequent
→ Where | D | is the total number of records.

For example, the rule for coverage :
(Gives birth = yes) Λ (Body temperature = warm blooded) → mammals is covered 33% from the out of 9 records i.e. 3/9 = 33%.

For example, the rule for accuracy :
(Gives birth = yes) Λ (Body temperature = warm blooded) → mammals accuracy is 100%

5.1.2.How a rule-based classifier works: A rule-based classifier is classified by a test record based on the rule triggered by the record. This is illustrated using following table.

Name	Body Temperature	Gives Birth	Aquatic Creature	Aerial Creature	Has legs	Class label
Lemur	warm- blooded	yes	no	no	yes	
Turtle	cold- blooded	no	semi	no	yes	?

- From the table, Lemur is warm blooded and "gives birth=yes". So, it is triggered by rule r3 and result is classified as mammal. i.e.
  - (Gives birth = yes) Λ (Body temperature = warm blooded) → mammals
- From the table, Turtle is cold blooded and "gives birth=no". So, two rules are triggered ie. rule r4 & r5.
  r4 : (Gives birth = no) Λ (Aerial creature = no) →Reptiles
  r5 : (Aquatic creature = semi) → Amphibian

In this case, the result is classified as reptiles and amphibians. To reduce this difficulty, two important properties are generated. They are

1. Mutually Exclusive Rules 2. Exhaustive rules.

Mutually Exclusive Rules: The rules in a rule set R are mutually exclusive if no two rules in R are triggered by the same record.

a. Classifier contains mutually exclusive rules if the rules are independent of each other
b. Every record is covered by at most one rule.

Eg: r : (Gives birth = yes) →mammals

Exhaustive Rules: A rule set R has exhaustive if a rule is a combination of attribute values.

a. Classifier has exhaustive coverage if it accounts for every possible combination of attribute values
b. Each record is covered by at least one rule

Eg. r : (Gives birth = no) → non-mammals. This can be split into sub-conditions such as Reptiles, birds, amphibians etc. For example,

NAME	BT	GB	AQC	AC	has legs	class label
Turtle	cold- blooded	no	semi	no	yes	?

In this case, we need refer more than one rule to say the class label. i.e.
r : (Gives birth = no) Λ (Aerial creature = no) →Reptiles and
r : (Aquatic creature = semi) → Amphibian

To overcome this problem, two rules were developed. They are ordered rules and unordered rules.

Ordered Rule: In this approach the rules in a rule set is ordered in decreasing order of their priority. This can be defined in many ways. i.e. coverage and accuracy.

Unordered Rules: This approach allows a test record to trigger multiple classification rules and consider the consequent of each rule as a vote for a particular class.

5.1.3. Rule-ordering Schemes: Rule-ordering schemes can be implemented on a rule-by-rule basis (or) a class-by-class basis.

Rule-Based Ordering Scheme: This approach orders the individuals rules by some rule quality measure. This ordering scheme ensures that every test record is classified by the "best" rule covering.

For example, "Aquatic creature = semi → Amphibians".

Class-Based Ordering Scheme: In this approach, rules that belong to the same class appear together in the rule set R. The rules are then collectively sorted on the basis of their class information.

For example, the class label which is repeated more no. of times than other is first priority, the class label which is repeated less no. of times in the training set is last priority.

How to build a Rule-based Classifier: To build a rule-based classifier, there are two methods are using for extracting classification rules. 1. Direct method, 2. Indirect method.

5.1.4. Direct method: The Sequential covering algorithm is often used to extract rules directly from data. This algorithm extracts the rules one class at a time for data sets that contain more than two classes.

For the mammals and Non-mammals classification problem, the sequential covering algorithm generate the rules for classifying mammals first, followed by rules for classifying birds & amphibians second, and reptiles and fishes are at last.

The priority for deciding which class should be generated first depends on a number of factors i.e. records. This is given in the following algorithm.

Algorithm: Sequential covering algorithm:

Step 1: Let E be the training records and A be the set of attribute-value pairs, $\{(A_j, v_j)\}$

Step 2: Let Y_0 be an ordered set of classes $\{y_1, y_2, \ldots, y_k\}$.

Step 3: Let R = { } be the initial rule list

Step 4: **for** each class $y \in Y_0 - \{y_k\}$ **do**

Step 5: **while** stopping condition is not met **do**

Step 6: r ← Learn-one-rule (E, A, y)

Step 7: Remove training records from E that are covered by r

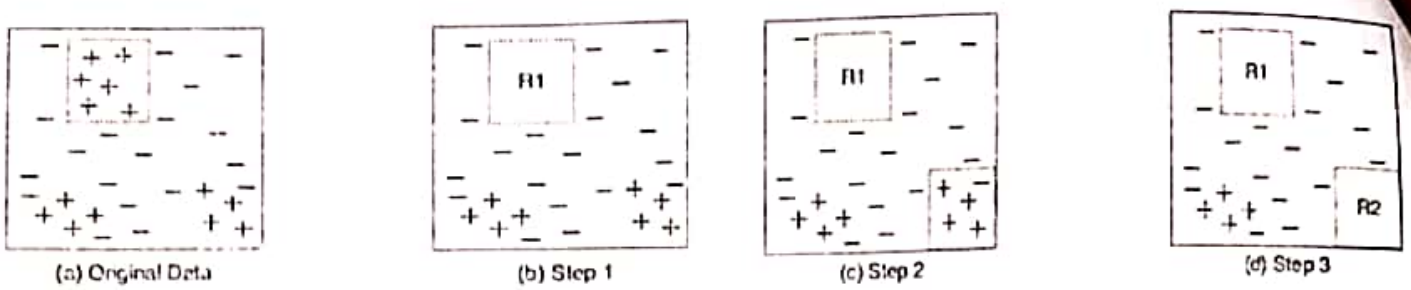
Step 8: Add r to thee bottom of the rule list: R → R V r.

Step 9: **end while**

Step 10: **end for**

Step 11: Insert the default rule, { } → y_k. to the bottom of the rule list R.

The above algorithm is demonstrated by using the sequential covering algorithm. The data set that contains a collection of positive and negative examples.



(a) Orginal Data (b) Step 1 (c) Step 2 (d) Step 3

→ The rule R1 coverage is shown in fig. (b) that covers the largest positive examples. i.e., 6

→ The rule R2, coverage is shown fig (d) is extracted (i.e. 4 positive and 1 negative symbol) after removing the R1 coverage.

→ Thus, from the graph R1 is classifier first and R2 is classifier in second.

Rule Evaluation: The rule evaluation is determined by using "sequential algorithm" and result is measured by using following 3 rules.

i) Statistical test using prune rule. ii) FOIL's info. Gain-algorithm. iii) RIPPER algorithm.

i) Statistical Test using prune rule: This rule is coverage by using the ratio statistical formula.

$$R = 2 \sum_{i=1}^{k} f_i \log (f_i / e_i)$$

→ Where k is the number of classes.

→ Where f_i is the frequency of class I values which is Covered by the rule R.

→ e_i is the expected frequency of a rule that makes random predictions.

→ e positive (e_+) = (Total no. of positive symbols in R1 x Total no. of positives in entire graph) / (Total no. of positive symbols + negative symbols in graph)

→ e negative (e_-) = (Total no. of negative symbols in R1 x Total no. of negative in entire graph) / (Total no. of positive symbols + negative symbols in graph)

ii) FOIL's info. Gain: This is evaluated by using the formula is

$$R = p_0 (\log_2 p_0/(p_0 + n_0)$$
$$R = p_1 (\log_2 p_1/(p_1 + n_1) - p_0 (\log_2 p_0/(p_0 + n_0) \quad \text{and so on.}$$

→ p_0 covers no. of positive symbols in R1 but no negative symbols in R1.

→ p_1 covers no. of positive symbols in R2 & only one negative symbol in R2.

iii) RIPPER algorithm: The RIPPER algorithm also works well with noisy data sets becz that prevent model overfitting. The formula is

$$(p - n) / (p + n)$$

→ Where 'p' refers no. positives and 'n' refers no. of negatives.

Suppose, If R1 covers 50 positives and 5 negatives, then (50 – 5) / (50 + 5) = 81%

For example, a data table with training set that contains 60 positive examples (+ve symbols) and 100 negative examples (-ve symbols). Total no. of symbols are 100 + 60 = 160.

Sreenivaas - 9948808818

→ Suppose, if we formed two rules using sequential coverage such as R1 & R2 as follows.

Rule R1 : Covers 50 +ve symbols and 5 –ve symbols i.e. totally R1 has 55.
Rule R2: Covers 2 +ve symbols and 0 –ve symbols i.e. totally R2 has 2.

i) Statistical Test using prune rule :

$e_+ = 55 \times 60/160 = 20.625$ $e_- = 55 \times 100/160 = 34.375$

$R1 = 2 \times 50 \times \log_2(50/20.625) + 5 \times \log_2(5/34.375)$

$e_+ = 2 \times 60/160 = 0.75$ $e_- = 2 \times 100/160 = 1.25$

$R2 = 2 \times 2 \times \log_2(2/0.75) + 0 \times \log_2(0/1.25) = 5.66$

ii) FOIL's information Gain:

$R1 = p_0 = 50$ +ve and no = 5 –ve Therefore

$= 50 \times (\log_2 5/(50+5) - \log_2 2/(2+0))$

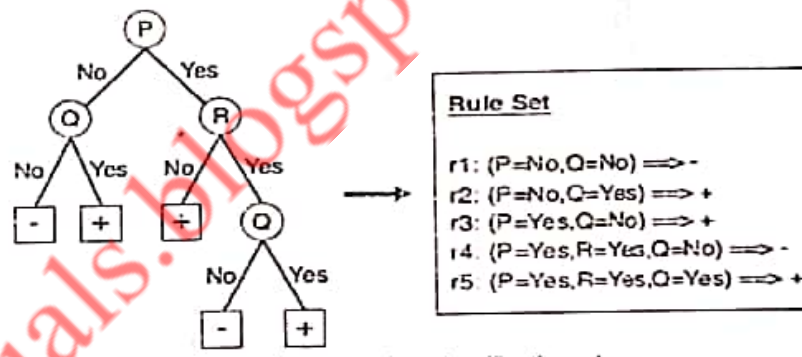
iii) RIPPER algorithm:

Formula $(p - n) / (p + n)$

Suppose, If R1 covers 50 positives and 5 negatives, then $(50 - 5) / (50 + 5) = 81\%$

5.1.5: Indirect method for rule extraction: This method generating the rule set from a decision tree. In this every path from the root node to the leaf node of a decision tree can be expressed as a classification rule.

The test conditions are encountered by using the rule set. Figure shows an example of a rule set generated from a decision tree. Notice that the rule set is exhaustive and contains mutually exclusive rules.



Converting a decision tree into classification rules.

Example 5.2. Consider the following three rules from above Figure:

r2 : (P = No) Λ (Q = Yes) → + (plus)
r3 : (P = Yes) Λ (R = No) → +
r5 : (P = Yes) Λ (R = Yes) Λ (Q = Yes) → + Observe that the rule set always predicts a positive class when the value of Q is Yes. Therefore, we may simplify the rules as follows:
r2': (Q = Yes) → + and r3: (P = Yes) Λ (R = No) → +.

Eg: The mammals and non-mammals classification problem is removed using a decision tree.

r1 : (Gives birth = no) Λ (Aerial creature = yes) →birds
r2 : (Gives birth = no) Λ (Aquatic creature = yes) →fishes
r3 : (Gives birth = yes) Λ (Body temperature = warm blooded) →mammals
r4 : (Gives birth = no) Λ (Aerial creature = no) →Reptiles
r5 : (Aquatic creature = semi) → Amphibian

5.1.6 : Characteristics of Rule-Based Classifiers : A rule based classifier has the following characteristics:

1. A decision tree can be represented by a set of mutually exclusive and exhaustive rules.
   i) Mutually exclusive rules

a. Classifier contains mutually exclusive rules if the rules are independent of each other
b. Every record is covered by at most one rule.

ii) Exhaustive rules

c. Classifier has exhaustive coverage if it accounts for every possible combination of attribute values
d. Each record is covered by at least one rule

2. Rule-based classifiers are generally used to produce descriptive models that are easier to interpret.
3. The class-based ordering approach adopted by many rule-based classifiers. Eg: RIPPER algorithm.

5.2: Bayesian Classifiers: What is Bayesian classifier?

In many applications the relationship b/w the attribute set and the class variable is non-deterministic. This situation creates noisy of data.

For example, to predict (extract) the heart disease persons based on diet. i.e. the people who eat healthy food and exercise regularly have less chance to get heart disease. But if they have smoking and alcohol habit then heart disease can increase.

To identify this type of disease people is difficult. This causes noisy of data. To overcome this, Bayesian theorem were used. The Bayesian theorem can predict the probabilities of memberships are called "Bayesian classifiers".

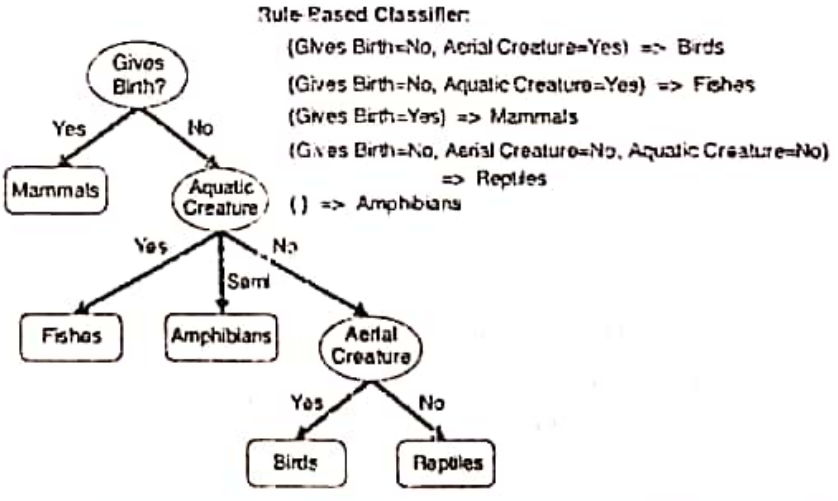
Advantages:

1. Bayesian classifiers are efficient like decision trees and neural network classifiers.
2. Bayesian are very accurate.
3. Bayesian classifiers can express high speed and class conditional independence. It means, the attribute values within a class are independent of each other.
4. It reduce the complexity.

5.2.1: Bayesian theorem:

This technique can predict the probabilities of memberships are called "Bayesian classifiers".

- 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.



Classification rules extracted from a decision tree for the vertebrate classification problem.

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

- The joint and conditional probabilities for X and Y are as:

  $P(X, Y) = P(Y|X) \times P(X) = P(X|Y) \times P(Y)$. This can be written as follows:

Bayes theorem:

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

- Where X is a data value (or) record and regarded (connected) as "evidence".
- Where Y is hypothesis. i.e. a data value (record) X belongs to a particular C.
- Where P(X) can be different hypothesis, expressed as

$P(X) = \sum_{i=1}^{k} (P(X_i | Y_i) \times P(Y_i)$ → This is called as the law of total probability.

Consider a football game b/w two rival teams : Team 0 and Team 1. Suppose Team 0 wins 65% of the time and Team 1 wins the remaining matches. Among the games won by Team 0, only 30% of them come from playing on Team 1's football filed. 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 b/w the two teams, which team will most likely emerge as the winner?

The Bayes theorem can be used to solve the prediction problem where noisy data occurred. For notational convenience, let X be the random variable that represents the team hosting the match and Y be the random variable that represents the winner of the match. Both X and Y can take on values from the set {0, 1}. We can summarize the information given in the problem as follows:

Probability Team 0 wins is $P(Y = 0) = 0.65$.
Probability Team 1 wins is $P(Y = 1) = 1 - P(Y = 0) = 0.35$.
Probability Team 1 hosted the match it won is $P(X = 1|Y = 1) = 0.75$.
Probability Team 1 hosted the match won by Team 0 is $P(X = 1|Y = 0) = 0.3$.

Our objective is to compute $P(Y = 1|X = 1)$, which is the conditional probability that Team 1 wins the next match it will be hosting, and compares it against $P(Y = 0|X = 1)$. Using the Bayes theorem, we obtain

$$P(Y = 1|X = 1) = \frac{P(X = 1|Y = 1) \times P(Y = 1)}{P(X = 1)} \rightarrow \frac{P(X = 1|Y = 1) \times P(Y = 1)}{P(X = 1, Y = 1) + P(X = 1, Y = 0)}$$

$$\rightarrow \frac{P(X = 1|Y = 1) \times P(Y = 1)}{P(X = 1|Y = 1)P(Y = 1) + P(X = 1|Y = 0)P(Y = 0)}$$

$$\rightarrow \frac{0.75 \times 0.35}{0.75 \times 0.35 + 0.3 \times 0.65} = 0.5738$$

- where the law of total probability was applied in the second line.
- Furthermore, $P(Y = 0|X = 1) = 1 - P(Y = 1|X = 1) = 0.4262$.
- Since $P(Y = 1|X = 1) > P(Y = 0|X = 1)$, Team 1 has a better chance than Team 0 of winning the next match.

What is prior and posterior Probability? Explain with an example.

- **Prior Probability P(Y):** It is the probability that the hypothesis Y exists without depending on the data record X.
- **Prior Probability P(X):** It is the probability that data record X exists without depending on the hypothesis Y.
- **Posterior Probability of Y on X is P(Y|X):** It is a probability that the hypothesis Y exists when a data record X is given. It means, data record X belongs to class C when the attributes of X is already given.

For example, suppose the world of fruits is described by their color and shape. Suppose that 'X' is red and round and Y is hypothesis that said X is an apple.

- Here P(Y) is the prior probability that expressed X is apple.
- Here P(Y|X) is posterior probability based on more information (ie. Background knowledge) such as X is red and round.
- ||ly P(X) is the prior probability of X. It is said that set of fruits are red and round.

- Suppose that hypothesis h1 (or) hypothesis h2 must occur, but not both. Also suppose that X_i is an observable event.
- Therefore Bayes theorem (or) rule is

$$P(Y_1 | X_i) = \frac{P(X_i | Y_1) P(Y_1)}{P(X_i | Y_1) P(Y_1) + P(X_i | Y_2) P(Y_2)}$$

Example 1: Suppose a person can decide to go to office in 3 modes of transportation. Such as car, bus, and train. Because of high traffic, if he/she decided to go by 'car' there is 50% chance he/she will late. If he goes by 'bus' 20% late and if he/she goes by 'train' 1% late. Find which is better.

Sol. Total no. of chances to reach office = 1/3.
i.e. pr{bus} = pr{car} = pr{train} = 1/3

- By car i.e. pr{late|car} = 0.5 (50%)
- By bus i.e. pr{late|car} = 0.2 (20%)
- By train i.e. pr{late|car} = 0.01 (1%)

$$\text{Bayes theorem for } pr\{car|late\} = \frac{Pr\{late|car\}\ pr\{car\}}{Pr\{late|car\}\ pr\{car\} + Pr\{late|bus\}\ pr\{bus\} + Pr\{late|train\}\ pr\{train\}}$$

$$= \frac{0.5 \times 1/3}{0.5 \times 1/3 + 0.2 \times 1/3 + 0.01 \times 1/3} = 0.7012$$

$$\text{Bayes theorem for } pr\{bus|late\} = \frac{Pr\{late|bus\}\ pr\{bus\}}{Pr\{late|bus\}\ pr\{bus\} + Pr\{late|car\}\ pr\{car\} + Pr\{late|train\}\ pr\{train\}}$$

$$\rightarrow \frac{0.2 \times 1/3}{0.2 \times 1/3 + 0.5 \times 1/3 + 0.01 \times 1/3} =$$

$$\text{Bayes theorem for } pr\{train|late\} = \frac{Pr\{late|train\}\ pr\{train\}}{Pr\{late|train\}\ pr\{train\} + Pr\{late|car\}\ pr\{car\} + Pr\{late|bus\}\ pr\{bus\}}$$

$$\rightarrow \frac{0.01 \times 1/3}{0.01 \times 1/3 + 0.5 \times 1/3 + 0.2 \times 1/3} = \frac{0.0033}{0.0033 + 0.1667 + 0.0667} = \frac{0.0033}{0.2367} = 0.0139$$

5.3. Naïve Bayes Classifier :

- Bayesian classifier is called "naïve Bayesian classifier". It is expressed high accuracy and speed when applied in large database.
- It is high performance than decision tree and neural network classifier.
- A naïve Bayes classifier estimate the class conditional probability by the values of attribute are conditional independent in the given class label y.
- The conditional independence assumption can be stated as $p(x|y) = y) \rightarrow \prod_{t=1}^{d} P(x_i|y = y)$
- where each attribute set $x = \{ x_1, x_2, x_3, \ldots . x_d)$, and → where d is an attribute.

5.3.1: Conditional Independence :

- Let X, Y and Z denote three sets of random variable.
- The variable in X are said to be conditionally independent of Y, given Z. Therefore the conditional holds as $P(X|Y, Z) = P(X|Z)$.

Therefore the conditional independence b/w X and Y is written as

$$P(X, Y|Z) = \frac{P(X, Y, Z)}{P(Z).} = \frac{P(X, Y. Z)}{P(Y, Z)} \times \frac{P(Y, Z)}{P(Z)} = P(X|Y, Z) \quad X \quad P(Y, Z) = P(X|Y) \text{ X } P(Y|Z)$$

5.3.2: How a Naïve Bayes Classifier works:

- Instead of computing the class-conditional probability for every combination of X, we estimate the conditional probability of each X_i, given , given Y.
- To classify a test record, the naïve Bayes classifier computes the posterior probability for each class Y:

$$P(Y|X) = P(Y) \times \prod_{t=1}^{d} P(X_i|y = y) / P(X)$$

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

$$P(Y) \times \prod_{t=1}^{d} P(X_i|Y)$$

- $P(X_i|Y)$ is $P(X_1|Y) \times P(X_2|Y) \times \ldots \ldots \times P(X_d|Y)$
  - where X_d represents values of attribute A_d.

For **estimating the conditional probability** $P(X_i|Y)$, several approaches were used. They are categorical and continuous attributes.

1. Estimating conditional probabilities for categorical attribute
2. Estimating conditional probabilities for continuous attribute

Example for the Naïve Bayes classifer: Consider the data set that shown in figure. This can compute the class conditional probability for each *categorical attribute* along with mean and variance for the *continuous attribute*.

Tid	Home Owner	martial status	Annual Income	Defaulted borrower
1	yes	single	125 k	no
2	no	married	100 k	no
3	no	single	70 k	no
4	yes	married	120 k	no
5	no	divorced	95 k	yes
6	no	married	60 k	no
7	yes	divorced	220 k	no
8	no	single	85 k	yes
9	no	married	75 k	no
10	no	single	90 k	yes

P(Home Owner =yes|no) = 3/7
P(Home Owner =no|no) = 4/7
P(Home Owner =yes|yes) = 0
P(Home Owner =yes|yes) = 1
P(Martial status = single|no) = 2/7
P(Martial status = single|yes) = 1/7
P(Martial status = married|no) = 4/7
P(Martial status = married|yes) = 2/3
P(Martial status = divorced|no) = 1/3
P(Martial status = divorced|yes) = 0

For Annual Income If class = No then Sample mean = 110 & Sample variance = 2975
If class = yes then Sample mean = 90 & Sample variance = 25

From the above to predict the class label of test record X=(Home owner = No, martial status = married, Annual income = 120 K). Its posterior probability is P(No|X) and P(Yes|X)
Total no. of records 'yes' and Total no. of records 'No' in the above data set is
P(yes) = 0.3 and P(No) = 0.7 out of 10 records.

P(X|No) = P(Home owner = No|No) x P(martial status = married|No) x P(Annual income = 120 K|No)
= 4/7 x 4/7 x 0.0072 = 0.0024

P(X|Yes) = P(Home owner = No|Yes) x P(martial status = married|Yes) x P(Annual income = 120 K|Yes)
= 1 x 0 x1.2 x 10^{-9} = 0

5.4 Bayes Error rate:

In statistical classification, the **Bayes error rate** is the lowest possible error rate for a given class of classifier.

The Bayes error rate finds important use in the study of patterns and machine learning techniques.

Example: 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 feet, we can express their class-conditional probabilities as follows:

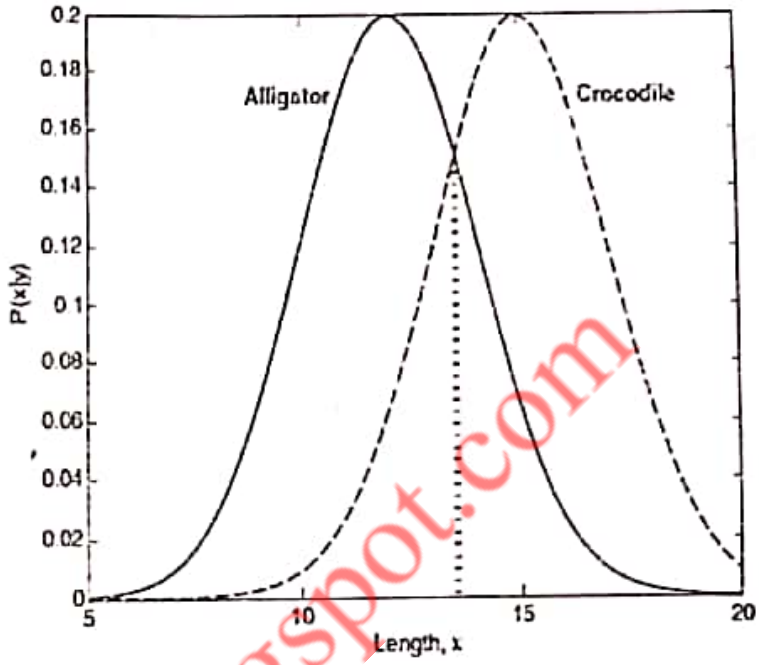


Figure. Comparing the likelihood functions of a crocodile and an alligator.

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

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

The above figure shows a comparison between the class-conditional probabilities for a crocodile and an alligator. Assuming that their prior probabilities are the same, the ideal decision boundary is located at some length ˆx such that P(X = ˆx|Crocodile) = P(X = ˆx|Alligator).
Using above Equations , we obtain

$$\left(\frac{\hat{x}-15}{2}\right)^2 = \left(\frac{\hat{x}-12}{2}\right)^2,$$

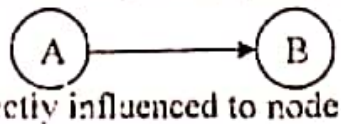
which can be solved to yield ˆx = 13.5.

5.5: Bayesian Belief Networks: The conditional independence assumption made by naïve bayes classifiers for classification problem in which the attributes are somewhat correlated. It also presents a more flexible approach for modeling the class-conditional probabilities P(X|Y). This approach allows us to specify which pair of attributes are conditionally independent.

5.5.1: Model Representation: A Bayesian belief network (BBN) provides a graphical representation of the probabilistic relational ships among a set of random variables. There are two key elements of a Bayesian network:

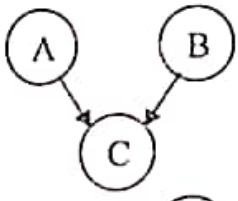
1. Directed acyclic graphic (dag). 2. Conditional probability table.

Directed acyclic graphic (dag): In direct acyclic graphic, an arc is directly established from one node another node. For example

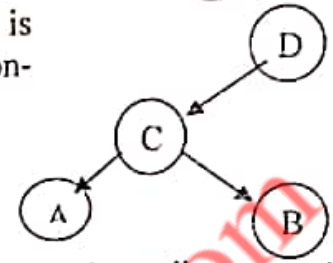


This examples shows a node A is directly influenced to node B.

Consider three random variables A, B, and 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. This is shown in fig.



Suppose, If there is a directed path in the network from D to C to A then A is **descendant** of D and D is an **ancestor** of B. Both B and D are non-descendants of A. Shown in fig.



Conditional probability table: A probability table associating each node to its immediate parent nodes. It means, each node in the graph represents a variable, and each arc asserts the dependence relationship b/w the pair of variables. If there is a directed arc from X to Y, then X is the **parent** of Y and Y is the **child** of X.

Conditional Independence: This is the property of Bayesian network. In this, a node in a Bayesian network is conditionally independent of its non-descendants, if its parents are known.

5.5.2: Model Building: Model building in Bayesian networks has steps. They are

1. Creating the structure of the network and
2. EstimatinD.g the probability values in the tables associated with each node.

The network topology can be obtained by encoding the subjective knowledge of domain experts. The Model Building algorithm presents a systematic procedure for the topology of a Bayesian network.

Algorithm: Algorithm for generating the topology of a Bayesian network.

Step 1: Let T = X1,X2, ... X_d) denote a total order of the variable.
Step 2: for j = 1 to d do
Step 3: Let $X_{T(j)}$ denote the j^{th} highest order variable in T.
Step 4: Let $\pi(X_{T(j)}) = X_{T(1)}, X_{T(2)}, \ldots, X_{T(j-1)})$ denote the set of variables preceding $X_{T(j)}$
Step 5: Remove the variables from $\pi(X_{T(j)})$ that do not affect X_j
Step 6: Create an arc b/w ($X_{T(j)}$) and the remaining variables in $\pi(X_{T(j)})$.

Example: Consider the variables shown in figure. After performing Step 1, assume that the variables are ordered in the following way: (E, D, HD, Hb, CP, BP). From step 2 to 7 starting with variable D, obtain the following conditional probabilities.

- P(D|E) is simplified to P(D).
- P(HD|E,D) cannot be simplified.
- P(Hb|HD,E,D) is simplified to P(Hb|D).
- P(CP|Hb,HD,E,D) is simplified to (CP|Hb,HD).
- P(BP|CP,Hb,HD,E,D) is simplified to (BP|HD).

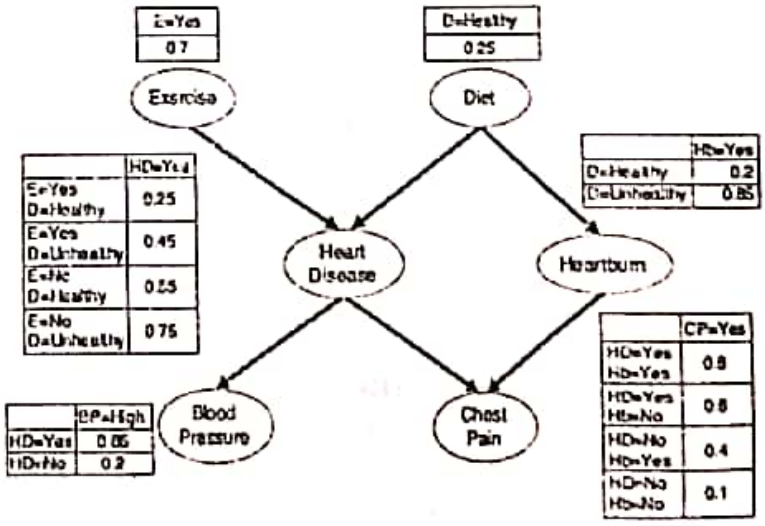


Figure 5.13. A Bayesian belief network for detecting heart disease and heartburn in patients.

Based on these conditional 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 the network structure shown in Figure.

Example of Inferencing Using BBN

Suppose we are interested in using the BBN shown in Figure 5.13 to diagnose whether a person has heart disease. The following cases illustrate how the diagnosis can be made under different scenarios.

Case 1: No Prior Information:

Without any prior information, we can determine whether the person is likely to have heart disease by computing the prior probabilities P(HD = Yes) and P(HD = No). To simplify the notation, let $\alpha \in$ {Yes, No} denote the binary values of Exercise and $\beta \in$ {Healthy, Unhealthy} denote the binary values of Diet.

$$P(HD = Yes) = \sum_{\alpha}\sum_{\beta} P(HD = Yes|E = \alpha, D = \beta)P(E = \alpha, D = \beta)$$

$$= \sum_{\alpha}\sum_{\beta} P(HD = Yes|E = \alpha, D = \beta)P(E = \alpha)P(D = \beta)$$

$$= 0.25 \times 0.7 \times 0.25 + 0.45 \times 0.7 \times 0.75 + 0.55 \times 0.3 \times 0.25 + 0.75 \times 0.3 \times 0.75$$

$$= 0.49.$$

Since P(HD = no) = 1 − P(HD = yes) = 0.51, the person has a slightly higher chance of not getting the disease.

Case 2: High Blood Pressure

If the person has high blood pressure, we can make a diagnosis about heart disease by comparing the posterior probabilities, P(HD = Yes|BP = High) against P(HD = No|BP = High). To do this, we must compute P(BP = High):

$$P(BP = High) = \sum_{\gamma} P(BP = High|HD = \gamma)P(HD = \gamma)$$

$$= 0.85 \times 0.49 + 0.2 \times 0.51 = 0.5185.$$

where $\gamma \in$ {Yes, No}. Therefore, the posterior probability the person has heart disease is

$$P(HD = Yes|BP = High) = P(BP = High|HD = Yes)P(HD = Yes) / P(BP = High)$$

$$= 0.85 \times 0.49 / 0.5185$$

$$= 0.8033.$$

Similarly, P(HD = No|BP = High) = 1 − 0.8033 = 0.1967. Therefore, when a person has high blood pressure, it increases the risk of heart disease.

Case 3: High Blood Pressure, Healthy Diet, and Regular Exercise

Suppose we are told that the person exercises regularly and eats a healthy diet. How does the new information affect our diagnosis? With the new information, the posterior probability that the person has heart disease is

$$P(HD = Yes|BP = High, D = Healthy, E = Yes)$$

$$= \left[\frac{P(BP = High|HD = Yes, D = Healthy, E = Yes)}{P(BP = High|D = Healthy, E = Yes)} \times P(HD = Yes|D = Healthy, E = Yes)\right]$$

$$= \frac{P(BP = High|HD = Yes)P(HD = Yes|D = Healthy, E = Yes)}{\sum_{\gamma} P(BP = High|HD = \gamma)P(HD = \gamma|D = Healthy, E = Yes)}$$

$$= 0.85 \times 0.25 \;/\; 0.85 \times 0.25 + 0.2 \times 0.75$$

$$= 0.5862,$$

while the probability that the person does not have heart disease is

$$P(HD = No|BP = High, D = Healthy, E = Yes) = 1 - 0.5862 = 0.4138.$$

The model therefore suggests that eating healthily and exercising regularly may reduce a person's risk of getting heart disease.

Characteristics of BBN

Following are some of the general characteristics of the BBN method:

1. BBN provides an approach for capturing the prior knowledge of a particular domain using a graphical model. The network can also be used to encode causal dependencies among variables.

2. Constructing the network can be time consuming and requires a large amount of effort. However, once the structure of the network has been determined, adding a new variable is quite straightforward.

3. Bayesian networks are well suited to dealing with incomplete data. Instances with missing attributes can be handled by summing or integrating the probabilities over all possible values of the attribute.

Because the data is combined probabilistically with prior knowledge, the method is quite robust to model overfitting.