Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

SOHAN GUPTA

UNIT-2 DMCT 5AID-01

# Predictive Modelling

**classification:-** Finding a good model that is used to predict the class of object. whose class level is unknown.

Ex: milk packet. we don't know, milk packet which section belongs., if then we will put this packet in which section.,
—supervised learning algo.

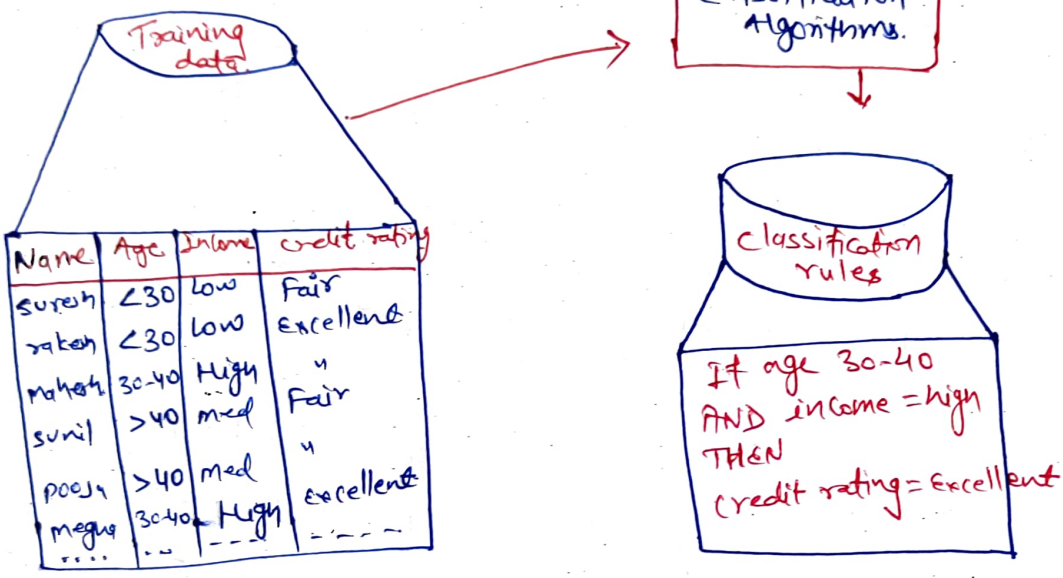
Ex: Grouping of the patient's based on their medical ~~records~~ Records.

Data classification is a two step's process—

In the first step-, a model is built describing a predetermined set of ~~data~~ classes of concepts. The model is constructed by analyzing database tuples described by attributes. Each tuples is assumed to belong to a predefined ~~data~~ class.

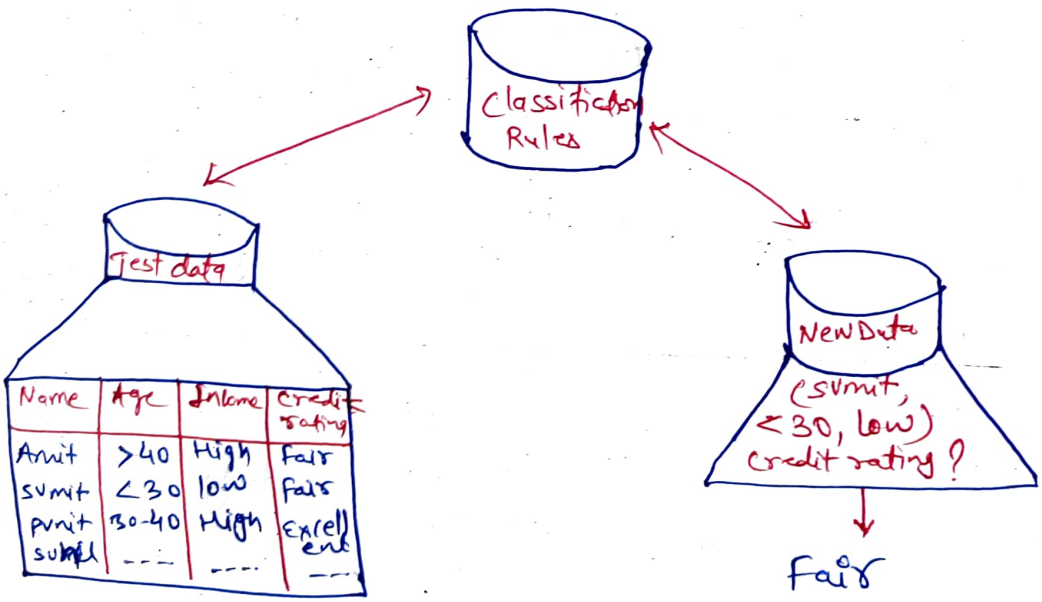
In the second ~~next~~ step,- the model is used for classification, First the predictive accuracy of the model is estimated, There are several method for estimating classifier accuracy.

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in




| Name | Age | Income | credit rating |
|---|---|---|---|
| sureh | <30 | low | Fair |
| rakesh | <30 | low | Excellent |
| mahesh | 30-40 | High | " |
| sunil | >40 | med | Fair |
| pooja | >40 | med | " |
| megha | 30-40 | High | Excellent |
| .... | .. | --- | ----- |

Fig a) learning: Traning data are analyzed by a classification algo.




| Name | Age | Income | credit rating |
|---|---|---|---|
| Amit | >40 | High | Fair |
| sumit | <30 | low | fair |
| punit | 30-40 | High | Excellent |
| sunil | --- | ---- | --- |

Fig b) Classification :- Test data is used to estimate the accuracy of the classification rules.

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
. E-mail: info@skit.ac.in Web: www.skit.ac.in

**Prediction :—** predicting a missing/unknown value based on past/current data.

- o/p - is a continuous data.

Ex: Predicting the correct treatment for a person based on their medical condition.

- unsupervised learning

Prediction can be viewed as the construction and use of a model to access the class of an unlabeled object or to access the value or value ranges of an attributes that a given object is likely to have. In this view-

classification and regression are two majors type of prediction problem where classification is used to predict discreate or nominal value, while regression is used to predict continuous or ordered values.

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

## A General Approach to classification –

classification is a two step process involve.

a) learning step – It is a step where the classification model is to be constructed. In this phase, training data are analyzed by a classification algorithms.

b) classification step – It's a step where the model is employed to predict class labels for given data. In this phase, test data are wont to estimate the accuracy of classification rules.

⇒ Issue Regarding classification and prediction –

1. Preparing the data for classification and prediction.
   - Data cleaning
   - In Consistency
   - Data transformation & reduction

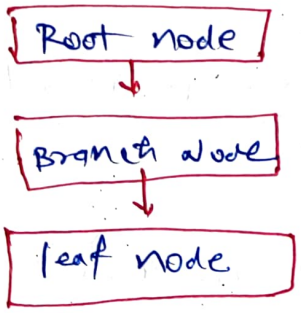
2. Comparision of classification and prediction method –
   - accuracy
   - speed
   - Robustness
   - scalability
   - Interpretability

– model specific measure.

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

**Decision Tree** Induction :- It is apply to classification and prediction

- A decision tree is a flow chart like a tree structure.
- It is used to visually define the rules for simple interpretation and understanding.
- It is a represent the rules ( Generate rule).
- There are 3 nodes in the Decision tree




(main question)

Intermediate node

Answer.

Ex' ID3

There two important ~~two~~ factors for draw a decision tree

**Information Gain** → measure of how much information the answer to a specific question provides.

**Entropy** — measure of how much uncertainity there is in the information.

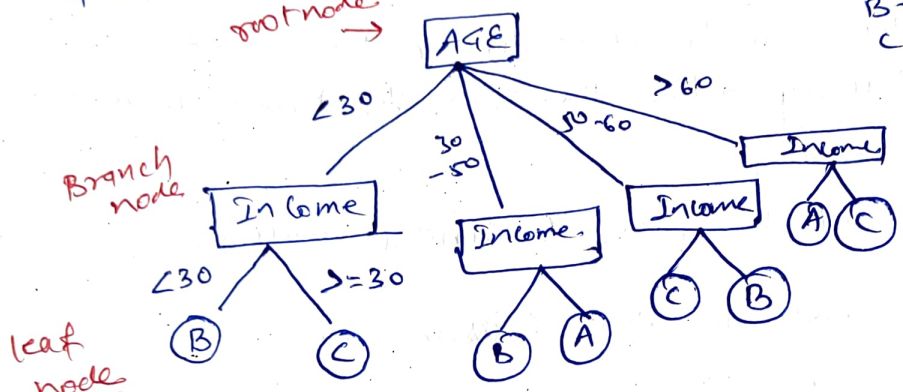
information Gain ↑ & Entropy ↓

when information gain increase then entropy will decrease.

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

A decision tree is shown below.

A - Average
B - Bad
C - Good
D - Excellent




Rule:- → if Age > 30 and Income ≥ 30k It is lie in Group B.

→ if Age ≤ 30, and ≥ 30k then credit score Good.

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

⇒ Naive Bayes classification —

- Naive Bayes classification is a supervised learning algo.
- It is based on Bays theorm which is based on probability for classification. They can predict the probabilities of class item
- It is used for solving a classification problem.
- It is mainly used in text classification that includes a high dimensional training data set.

Bay's theorm:— It is determine the probability of a hypothesis with prior knowledge. It depend on conditional probability. The Bays theorm is given by-

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

where $P(A|B)$ is a 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.

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

Ex: ~~the fruit is~~

Ex: Suppose we have a dataset of weather conditions and corresponding target variable "play". So using this dataset we need to decide that weather we should play or not on a particular day according to the weather conditions. So to solve this problem, we need to follow the below steps—

1. Convert the given dataset into frequency tables.
2. Generate ~~likely~~ likelihood table by finding the probabilities of given feature.
3. Now, use Bays theorm to calculate the posterior probability.

Problem— If the weather is sunny then the player should play or not ?

Solution— To solve this first consider the below dataset

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

| | Outlook | play |
|---|---|---|
| 0 | Rainy | Yes |
| 1 | sunny | Yes |
| 2 | overcast | Yes |
| 3 | overcast | Yes |
| 4 | sunny | No |
| 5 | rainy | Yes |
| 6 | sunny | Yes |
| 7 | overcast | Yes |
| 8 | Rainy | No |
| 9 | sunny | No |
| 10 | sunny | Yes |
| 11 | Rainy | No |
| 12 | overcast | Yes |
| 13 | overcast | Yes |

⇒ Frequency Table for the weather Condition –

| Weather | Yes | No |
|---|---|---|
| overcast | 5 | 0 |
| Rainy | 2 | 2 |
| sunny | 3 | 2 |
| Total | 10 | 4 |

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

→ likelihood table weather condition —

| weather | NO | yes | |
|---|---|---|---|
| overcast | 0 | 5 | 5/14 = 0.35 |
| Rainy | 2 | 2 | 4/14 = 0.29 |
| sunny | 2 | 3 | 5/14 = 0.35 |
| All | 4/14 = 0.29 | 10/14 = 0.71 | |

Apply the Bay's theorm

$$P(yes|sunny) = \frac{P(sunny|yes) * P(yes)}{P(sunny)}$$

$P(sunny|yes) = 3/10 = 0.3$, $P(NO) = 0.29$

$P(sunny) = 0.35$

$P(yes) = 0.71$

$P(sunny|No) = 2/4 = 0.5$

so $P(yes|sunny) = \frac{0.3 * 0.71}{0.35} = 0.60$

$$P(NO|sunny) = \frac{P(sunny|NO) * P(NO)}{P(sunny)} = \frac{0.5 * 0.29}{0.35} = .41$$

so we can say from above Calculation $P(yes|sunny) > P(NO|sunny)$

Hence on a sunny day, players can play the game.

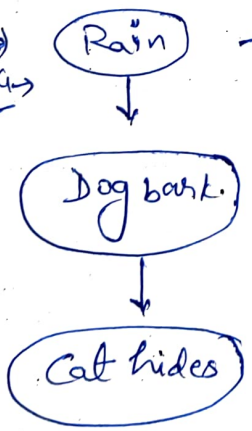
(2-10), 2-11

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

## Bayesian Belief Networks :—

- Bayesian Belief Network is a probabilistic graphical model (PGM) that represents conditional dependencies b/w random variables through DAG.
- It is also suitable for representing probabilistic relation b/w multiple events (more than 2 events)
- There are two concept involved in this method
  a) Direct Acyclic graph ( DAG)
  b) Conditional Probability Table ( CPT)

Ex.
A) DAG →




→ Node → Random variable → Hypothesis

- when raining then Dog bark, if Dog bark then Cat hides.
- it may or may not rain, the dog may be bark and cat also hides.
- all are explained in direction of form of arrow so it is acylic graph.

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

b) CPT –

| | R | ~R |
|---|---|---|
| B | 9/48 | 18/48 |
| ~B | 3/48 | 18/48 |

$(B=T \leftarrow R=T) = \frac{9}{48} = 0.19$

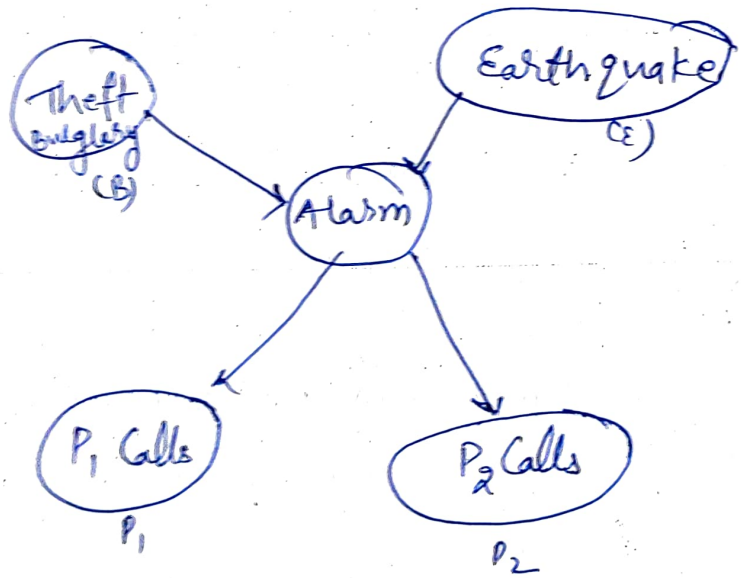
$(B=T \leftarrow R=F) = 18/48 = 0.375$

$(B=F \leftarrow R=T) = 3/48 = 0.06$

$(B=F \leftarrow R=F) = 15/48 = .375$

R means rain
~R means not rain
B means dog barks
~B means not dog barks

Ex.



Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

Given probabilities are.

$P(B=T) = 0.001$
$P(B=F) = 0.999$

$P(E=T) = 0.002$
$P(E=F) = 0.998$

Probability of Alarm →

| Theft (Bulglary) | Earthquake | 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 |

Probability of $P_1$ (in respect of parent)

| Alarm (A) | $P(P_1=T)$ | $P(P_1=F)$ |
|---|---|---|
| T | 0.90 | 0.10 |
| F | 0.05 | 0.95 |

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

Probability of $P_2$

| A | $P(P_2 = T)$ | $P(P_2 = F)$ |
|---|---|---|
| T | 0.70 | 0.30 |
| F | 0.01 | 0.99 |

Ex: Find the probability of $P_1$ is T, $P_2$ is T, A is T, B is F and E is F.

ie $P(P_1, P_2, A, \neg B, \neg E)$

$= P(P_1/A) \cdot P(P_2/A) \cdot P(A/\neg B, \neg E) \cdot P(\neg B) P(\neg E)$

$= 0.90 * 0.70 * 0.001 * 0.999 * 0.998$

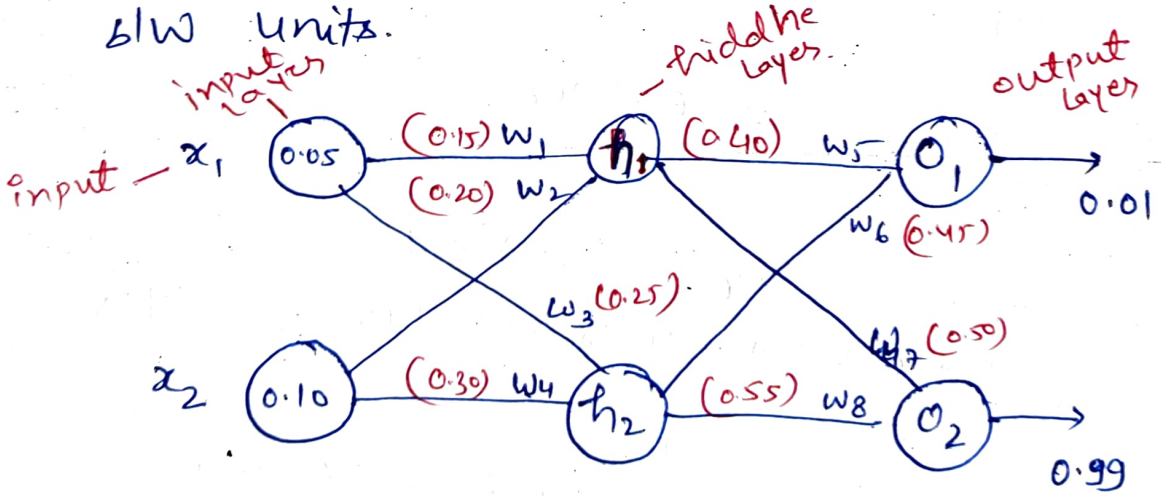
$= 0.00062$

## Classification by Backpropagation :— (Back propagation of errors)

Backpropagation is a neural network algorithms.

* Neural Network ~~algo~~ is a set of connected input/output unit, where each connection has a weight associated with it.

During the learning phase, the network learns by adjusting the weight so as to be able to predict the correct class label of the input samples. Neural network learning is also referred to as "connectionist learning" due to connection b/w units.




- Forward Backpropagation $b_1 = 0.35$ $b_2 = 0.60$
- Back Back propagation

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

$h_{1(in)} = w_1 * x_1 + w_2 * x_2 + b_1$

$= 0.15 * 0.05 + 0.2 * 0.1 + 0.35$

$= 0.377$

$h_{1(out)} = \frac{1}{1 + e^{-h(in)}}$ by using sigmoid function

$= \frac{1}{1 + e^{-(0.377)}} = 0.5932$

$h_{2(in)} = x_1 * w_3 + x_2 * w_4 + b_1$

$= 0.05 * 025 + 0.10 * 0.30 + 0.35$

$= 0.3915$

$h_{2(out)} = \frac{1}{1 + e^{-0.3915}} = 0.5968$

$O_{1(in)} = w_5 * h_{1(out)} + w_6 * h_{2(out)} + b_2$

$= 0.4 * 0.593 + 0.45 + 0.596 + 0.6 = 1.105$

$O_{1(out)} = \frac{1}{1 + e^{-O_{1(in)}}} = \frac{1}{1 + e^{-1.105}} = 0.7513$

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

$O_2(in) = w_7 * h_1(out) + h_2(out) * w_8 + b_2$

$= 0.5 * 0.5932 + 0.5968 * 0.55 + .60$

$= 1.22484$

$O_2(out) = \frac{1}{1+e^{-O_1(in)}} = \frac{1}{1+e^{-1.22484}} \quad \frac{1}{1+.2938} = 0.7729$

these are no match output value as the target value.

So there are errors to be occur.

errors factor $E_{(total)} = \sum \frac{1}{2} (target - o/p)^2$

$E_{O_1} = 0.274$

$E_{O_2} = 0.0235$

$E_{total} = E_{O_1} + E_{O_2}$

$= 0.274 + 0.0235 = 0.2983$

So there are errors factors to be occure (more errors) now we are doing back propagation to back on network.

Partial derivative

$W_5 = W_5 - \eta \frac{\partial E_{total}}{\partial W_5}$

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

Now alter the weight of the unit as the required ~~Sigmoid function~~ output value. This ~~is~~ is trial & error method, it may be done more times.

⇒ Calculating backpropagation errors—

output layer → hidden layer

$w_5$, $w_6$, $w_7$, $w_8$


First let us adjust $w_5$

$$w_5^* = w_5 - \eta \frac{d E_{total}}{d w_5}$$

($w_5^*$ = new value, $w_5$ = old value)

$\eta = 0.6$ (learning rate)

$$\frac{d E_{total}}{d w_5} = \frac{d E_{total}}{d out_{O_1}} * \frac{d out_{O_1}}{d net_{O_1}} * \frac{d net_{O_1}}{d w_5}$$

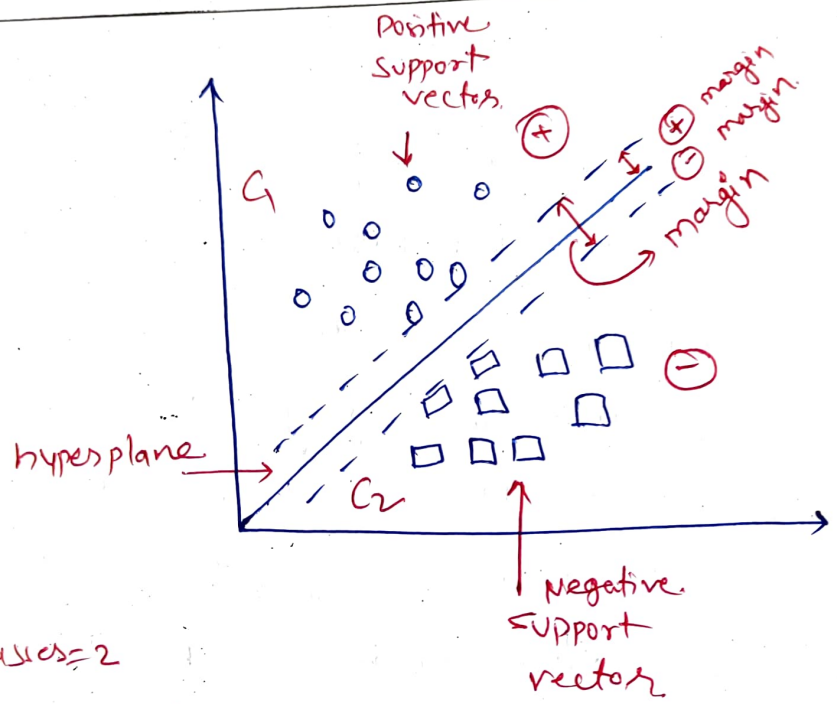
Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

## Support Vector Machine.

- Lazy learners
- Eager learners

- It is supervised learning algorithms which is used for classification as well as regression problems.
- The goal of svm algorithms is to create the best line or decision boundary that can segregate n-dimensional space into classes. so that we can easily put the new data point in the correct category in the future.
- This best decision line/boundary is called a hypersplane.
- svm chooses the extreme point/vector that help in creating the hypersplane. these extreme cases are called as support vectors and hence algorithms is termed as support vectors machine.
- these are two different classes, that are classified by using decision boundary or hypersplane.
-

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in




classes = 2

$C_1$ = circle

$C_2$ = quadrilateral.

Ex Face detection, image classification, text categorization etc.

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

## Type of support vector machine (SVM) →

SVM can be of two types–

a) linear SVM— It is used for linearly separable data which means if a data set can be classified into two classes by using a single straight line, then such data is termed as linearly seperable data, and classifier is used called as Linear SVM Classifier.

b) Non linear SVM:— It is used for non linearly seperated data, which means if a dataset can not be classified by using a straight line, then such data is termed as non linear data and classifier used is called as non linear SVM.

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

## Application of SVM:-

- Face observation
- Text and hypertext arrangement
- Grouping of portrayals. (Compare the piece & take decision)
- Bioinformatics (medical science, DNA, research,
- Handwriting remembrance.
- protein fold and remote homology, spooting
- Generalized predictive Control (GPC)

Swami Keshvanand Institute of Technology, Management & Gramothan,
Ramnagaria, Jagatpura, Jaipur-302017, INDIA
Approved by AICTE, Ministry of HRD, Government of India
Recognized by UGC under Section 2(f) of the UGC Act, 1956
Tel : +91-0141-3500300 Fax: +91-0141-2759555
E-mail: info@skit.ac.in Web: www.skit.ac.in

## Question Bank

Q.1 Write and explain the concept of classification and prediction along with various issues incorporated with them [RTU 2016]

Q.2 What is the advantage and disadvantage of decision tree approach over other approaches of data mining. [RTU 15]

Q.3 Explain the decision tree construction [RTU 2015]

Q.4 Explain the Naive Bayes classification.

Q.5 Briefly outline the major idea of Bayesian belief network classification.

Q.6 Write short notes on Back Propagation algorithms with its example.

Q.7 Explain support vector machine with its types. also explain its application.