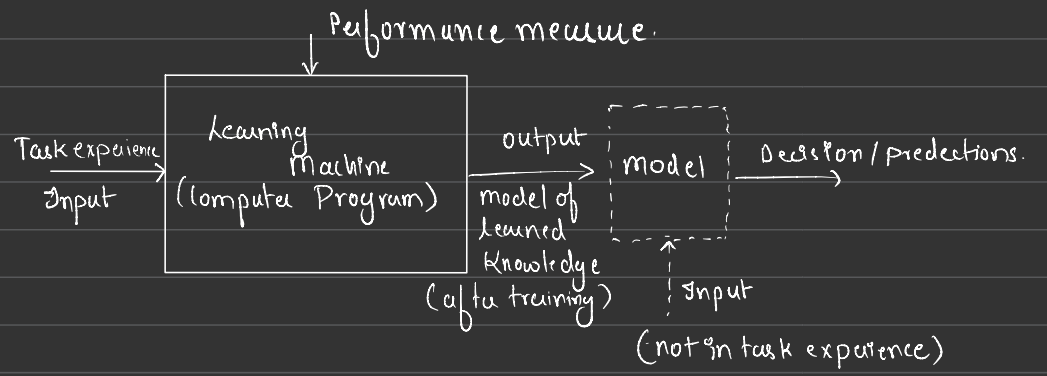
•> Metrics for Assessing Regression Accuracy.

⇒



•> Mean Square error

→ The mean is obtained from the training data as arithematic average. which measures the average square deviation of the predicted values from the true value.

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - h(w, x^{(i)}))^2$$

•> Root mean Square error

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y^{(i)} - h(w, x^{(i)}))^2}$$

•> Sum of Error Squares

⇒ Mathematical manipulation

$$= \sum_{i=1}^{N} (y^{(i)} - h(w, x^{(i)}))^2$$

•> Expected mean square error

$$EMSE = E\left[\sum_{i=1}^{N} (y^{(i)} - h(w, x^{(i)}))^2 \right]$$

E: statistical expectation operation

h: hypothesis funct.

•) Mean absolute Error

$$MAE \Rightarrow \frac{1}{N} \sum_{i=1}^{N} \left| y^{i} - h(w, u^{(i)}) \right|$$

•) Classification Accuracy

i) Miss classification Error

⇒ Formula = $\frac{\text{Number of data points for which } (y^{(i)} - \hat{y}^{(i)}) \neq 0}{\text{Number of samples } (N)}$

•) Confusion Matrix :

→

Actual values \ Predicted values	No	Yes
No	TN	FP
Yes	FN	TP

TN → True negative
FN → False negative
TP → True positive
FP → False negative.

⇒ The true +ve and true -ve are the accurate classifications.
⇒ A false +ve takes place when the result is inaccurately predicted as +ve but its -ve in reality.

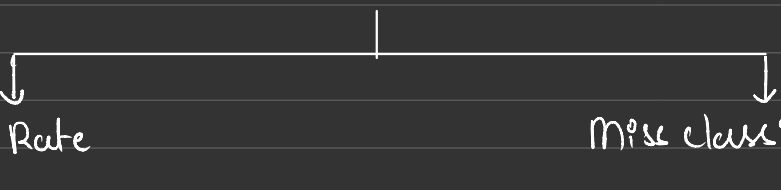
⇒ A false -ve is said to offer when the result is inaccurately predicted as -ve but in reality it is +ve

⇒ An ideal algorithm should create a diagonal matrix which FP = FN = 0

•) Metrices for calculation of Confusion matrix

1) Miss classification error or (Success Rate)

⇒ It has too parts



Success Rate

$$\frac{TP + TN}{TP+TN+FP+FN}$$

Miss classification Rate

$$\frac{FP+FN}{FP+FN+TP+TN}$$

True Positive Rate

$$\Rightarrow \frac{TP}{TP+FN}$$

True Negative Rate

$$\frac{TN}{FP+TN}$$

True Positive Rate tells how sensitive our decision method is in detecting the abnormal event

True Negative Rate tells how specific our decision method is in detecting the abnormal event

So a decision technique is said to be good if it is highly sensitive and highly specific.

Q)

AV \ PV	165	No	Yes	
	No	50 TN	10 FP	60
	Yes	5 FN	100 TP	105
		55	110	

Calculate the success rate, failure rate

Success = 0.9091
balue = 0.09

TP rate = 0.95
TN rate = 0.83

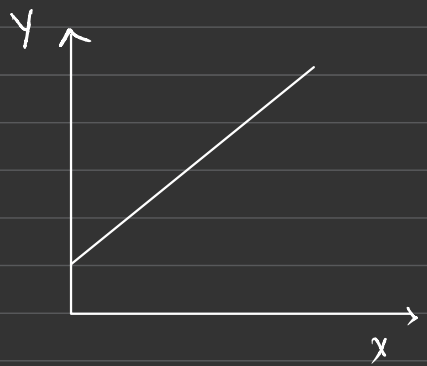
Unit - 7 : Generalized Linear Model (GLM)

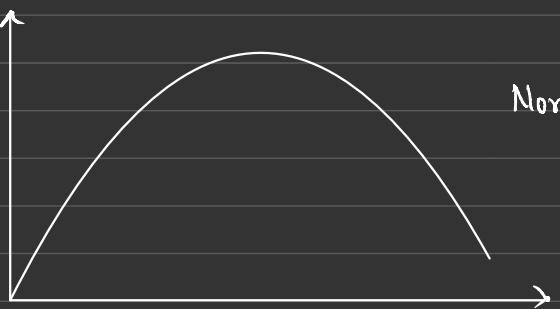
1) Why GLM
2) Assumptions Made
3) Component of GLM

•) Linear Models:

- Y (Dependent variable)
- X (Independent Variable)

→ When relation between Y & x are linear we will use linear model





Normal Distribution

error distribution

Definition: Generalised linear model allow to build a linear relationship Between the response and the predictors. Even though their underlying relationship is not linear. This is made possible by means of link function or activation function.

→ Here the errors in the response variable are assumed to follow an exponential family of distribution.

•) Assumptions of GLM

→ There are 2 - 3 assumptions that we will make.

1) Data should be independent and random
2) Y or response does not need to be normally distributed but distribution is from exponential family (Binomial, Beta, and gamma distributions)
3) The original response variable need not have a linear relationship with x. but the transformed response variable should be linearly dependent on x.

• Equation of logistic Regression:

$$\text{log odds} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots$$

Compurision between linear regression.

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots$$ (only LHS part is changed)

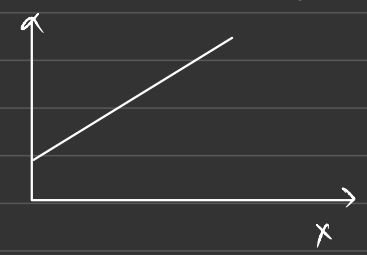
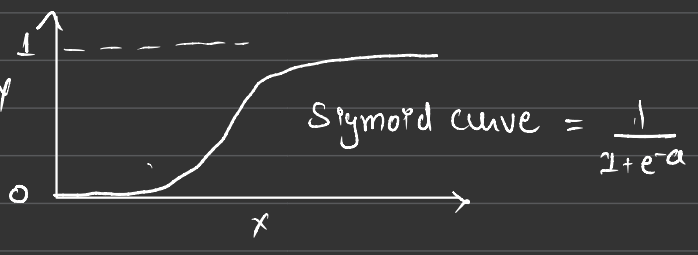
• Components of GLM

→ There are 3 components of GLM.

Systematic Component / linear predictor (X) Independent variable	Link function	Random Component / Probability Distribution Y (response variable)
1) It is the linear combination of predictors and the regression parameters ($\beta_1, \beta_2 \cdots$)	1) It specifies the link between random component & systematic component. It indicates how Y relates to the linear combination of X	1) It refers to the probability distribution from the exponential family of distribution of Y.

• Response variable | Distribution | Support | Link name | Link function

Only two possible | Bernoulli | Integers: {0, 1} | Logit | $XB = \ln(u/1-u)$

Linear Regression	Logistic Regression
→ Response : Continous → Y & X Linearly related	→ Response : binary / categorical → Y & X relation is not linear.
→ 	→ 
→ $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$ → least Square	log odds = $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots$ → maximum likelihood estimation.

→ The odds: ratio of the proportions for the two possible outcomes is a number between 0 and infinity

• What is logistic function

$$\log(\text{odds}) = \log \frac{P(\text{class } 1 \mid x)}{1 - P(\text{class } 1 \mid x)} = \omega_0 + \omega_1 x_1 + \omega_2 x_2 \cdots$$

$$P(\text{class } 1 \mid x) = \left(\frac{1}{1 + e^{-(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n)}} \right)$$

$$= \left(\frac{1}{1 + e^{-a}} \right) \rightarrow \text{Sigmoid function}$$

• Derivation / Parameters of logistic Regression.

⇒ We are given a set of observed data

→ Dataset → Input → output

$$D = \{ x^{(i)}, y^{(i)} \}; \; i = 1, \ldots, N$$

$$x = [x_1, x_2, \ldots, x_n]^T \in \mathbb{R}^n$$

$$y \in \{0, 1\}$$

∴ The proposed model

$$P(y=1 \mid x) = \frac{1}{1 + e^{-(\omega_0 + \omega_1 x_1 + \omega_2 x_2 + \cdots + \omega_n x_n)}}$$

$$= \frac{1}{1 + e^{-(\omega^T x + \omega_0)}} \quad \text{where } [\omega^T = \omega_1, \omega_2 \cdots]$$

$$P(y=0 \mid x) = 1 - P(y=1 \mid x)$$

⇒ Exponential family of distribution
- → Bernoulli
- → Sigmoid function $\frac{1}{1+e^{-a}}$

• Maximum Likelihood Criteria / Estimation

→ It is used for estimating the parameters of logistic regression model.

→ The parameter values are found such that they maximize the probability of observing the dataset of a given model.

⇒
$$\mathcal{L}(\{w, w_0\}, D) = \mathcal{L}(\bar{w}, D) = \prod_{i=1}^{N} P(y^{(i)} | x^{(i)})$$

⇒ In the maximum likelihood problem, our goal is to find w which will maximizes $\mathcal{L}$, i.e, we wish to find $\bar{w}^*$ where

$$\bar{w}^* = \arg\max_{\bar{w}} \mathcal{L}(\bar{w}, D)$$

⇒ Often we maximize $\log(\mathcal{L}(w, D))$ instead because it is analytically easier.

$$\log(\mathcal{L}(\bar{w}|D)) = \log\left(\prod_{i=1}^{N} P(y^{(i)} | x^{(i)})\right)$$
$$= \sum_{i=1}^{N} \log P(y^{(i)} | x^{(i)})$$

We model log-likelihood as

$$P(Y=y) = p^y (1-p)^{1-y}; \quad y \in [0, 1]$$

$$\log(\mathcal{L}(\bar{w}|D)) = \log\left(\prod_{i=1}^{N} P(y^{(i)} | x^{(i)})^{y^{(i)}} (1 - P(y^{(i)} | x^{(i)}))^{1-y^{(i)}}\right)$$
$$= \sum_{i=1}^{N} \log\left(g(x^{(i)}, \bar{w})^{y^{(i)}} (1 - g(x^{(i)}, \bar{w}))^{1-y^{(i)}}\right)$$

To maximize log-likelihood with respect to $\bar{w}$,

$$\log(L_i(\bar{w}|D)) = \log\left(g(x^{(i)}, \bar{w})^{y^{(i)}} (1 - g(x^{(i)}, \bar{w}))^{1-y^{(i)}}\right)$$
$$= y^{(i)} \log g(x^{(i)}, \bar{w}) + (1 - y^{(i)}) \log(1 - g(x^{(i)}, \bar{w}))$$

Taking derivative from above eqn

$$\nabla_{\bar{w}}(\log L_i(\cdot)) = \frac{y^{(i)}}{g(x^{(i)}, \bar{w})} \nabla_{\bar{w}}(g(x^{(i)}, \bar{w})) - \frac{1 - y^{(i)}}{1 - g(x^{(i)}, \bar{w})} \nabla_{\bar{w}}(g(x^{(i)}, \bar{w}))$$

Now,

$$g(x^{(i)}, \bar{\omega}) = \frac{1}{1+e^{-a}} \; ; \quad a = (\bar{\omega}^T x^{(i)})$$

Therefore

$$\nabla_{\bar{\omega}} (g(x^{(i)}, \bar{\omega})) = \frac{\partial}{\partial a}\left(\frac{1}{1+e^{-a}}\right) \frac{\partial a}{\partial \bar{\omega}}$$

$$= \left(\frac{1}{1+e^{-a}}\right)\left(1 - \frac{1}{1+e^{-a}}\right) \bar{x}^{(i)}$$

$$= g(x^{(i)}, \bar{\omega})(1 - g(x^{(i)}, \bar{\omega}))\, \bar{x}^{(i)}$$

This gives

$$\nabla_{\bar{\omega}}(\log(\mathcal{L}_i(\bar{\omega}|D))) = (y^{(i)} - g(x^{(i)}, \bar{\omega}))\, x^{(i)}$$

∴ The gradient ascent function will be as follows.

$$\Rightarrow \omega_{new} = \omega_{old} + \eta \frac{\partial(\text{likelihood function})}{\partial \omega}$$

$$= \omega_{old} + \eta \sum_{i=1}^{N} (y^{(i)} - g(\bar{x}^{(i)}, \bar{\omega}))\, x^{(i)}$$

•) Difference Between linear model & Generalized Linear Model

Generalized Linear Model	Linear model
1) Logistic Regression	1) Linear Regression
2) Catagorical output	2) Output is Continous.
3) Non Normal distribution	3) Normal Distribution.

• Logistic Regression.

→ It is a Supervised machine learning algorithm that can be used to model the probability of a certain class or event.

→ It is a non linear regression problem.

• Difference Between Probability & Likelihood.

→ Probability follows clear parameters and computations while a likelihood is based mealy on observed factors.

$$L(\mu, \sigma, data) = P(data; \mu, \sigma)$$

(μ: mean, σ: standard deviation.)

→ Probability	Likelihood
μ = Mean, σ = S.D } constant	μ, σ } may vary

→ $P(data; \mu, \sigma)$: It means "The probability density of observing the data with model parameter μ & σ". We can generalize this to any number of parameters and any distribution.

→ $L(\mu, \sigma; data)$: It means "The likelihood of the parameter μ and σ taking certain values given that we have observed a bunch of data.

Q) The dataset of pass or fail in an exam for 5 students is given in the table below

Hours of studies	Result (1 = Pass, 0 = Fail)
29	0
15	0
33	1
28	1
39	1

Use logistic regression as classifier where log of odds is given

$$\log(\text{odds}) = -64 + 2 \times \text{hours}$$

a) How to calculate the probability of pass for student who studied 33 hours.

b) Atleast how many hours the student should study that make suc will pass the course with the probability of more than 95%

⇒ Question on Calculating Probability with logistic Regression

Solution :-

a) We will use the sigmoid function to calculate the probability

Hours = 33

$= z = -64 + 2 \times 33 = 2$

$$p = \frac{1}{1 + e^{-z}} = \frac{1}{1 + e^{-2}}$$

$= 0.88 \Rightarrow 88\%$

b) $$0.95 = \frac{1}{1 + e^{-z}}$$

Solve for z $\Rightarrow$ $z = 2.944$

$\Rightarrow$ $2.944 = -64 + 2 \times \text{hours}$.

$\Rightarrow$ $2.94 + 64 = 2 \times \text{hours}$.

$=$ Hours $= 33.47$

Unit - 9

• Tree based learner (Classification Problem)
↳ Supervised learning technique.

→ Decision tree algorithm

→ Different ways to train a machine.

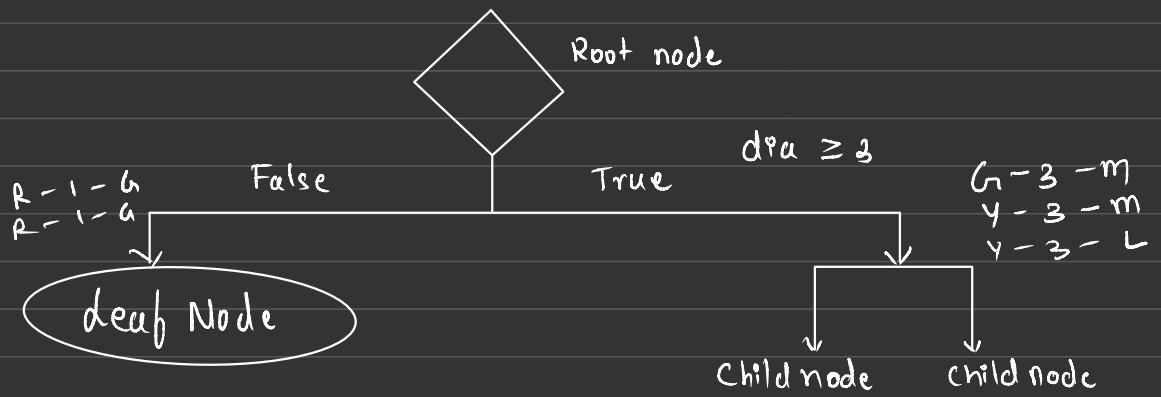
→ Decision tree → Common
→ random forest → many no of decision tree are combined randomly
→ naive bayes
→ KNN

⇒ Decision tree is a graphical representation of all possible solutions to a decision based on certain conditions.

⇒ Random forest build multiple decision tree and merge them together to get a more accurate and stable prediction.

Example :-

Colour	diameter	Label
Green	3	Mango
Yellow	3	Mango
Red	1	Grape
Red	1	Grape
Yellow	3	Lemon



•> Different Terminologies

1) Root node :- (Parent Node) It is a base node of a tree which represents the entire population or sample and further gets divided into two or more homogenous sets.

2) Leaf node :- It is one when we reach at the end of the tree it cannot further segregate into any other level.

3) Child node :- All root node is always the parent node while all other bottom node associated with that are called child node.

4) Splitting :- It is the dividing of root node into different subparts on the basis of some condition.

5) Branch / Subtree :- Branch gets formed when you split the tree

6) Pruning :- We are removing the subnodes of a decision tree.

7) Gini Index :- It is a measure of impurity or purity used in building a decision tree.

8) Information Gain :- It is the decrease in "entropy" after a dataset is split on the basis of an attribute for constructing a decision tree we will select the nodes which has highest Information gain.

9) Entropy :- $S = -P(yes)\log_2 P(yes) - P(No)\log_2 P(No)$

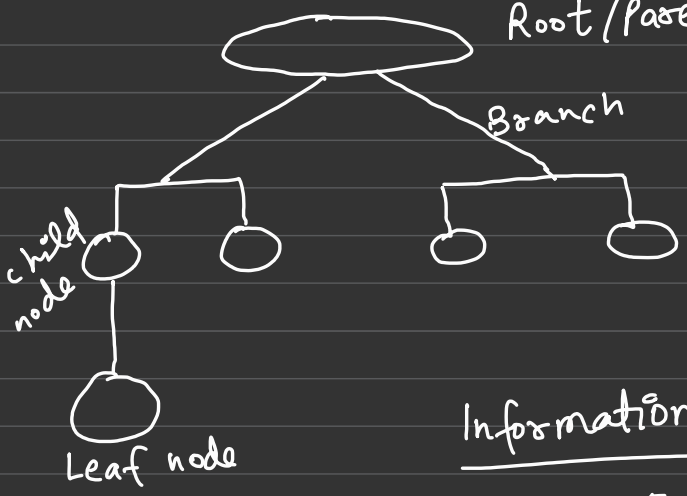
Unit - 9 : Decision Tree

⇓

Classification

⇓

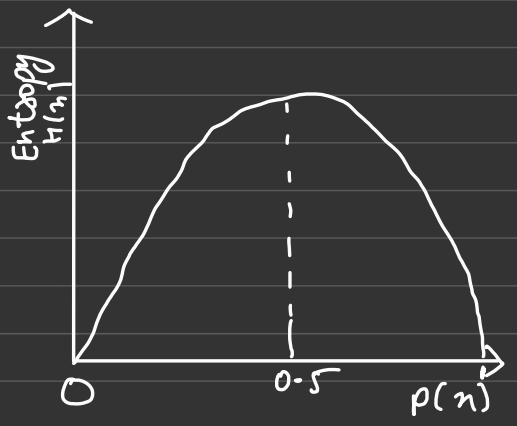
Supervised ML



Information Gain :

Entropy

As the probability is zero or one then in that case entropy is zero. When probability is 0.5 then the value of probability is maximum.



$$E(S) = -P_{yes} \log_2 P(yes) - P_{No} \log_2 P(No)$$

- if no. of YES = no. of No

  $\therefore P(S) = 0.5$

  $E(S) = 1$

- if it contains all YES or all NO

  $P(S) = 1$ / $P(S) = 0$

  $E(S) = 0$

S: Sunny O: Overcast R: Rain H: Hot C: Cold M: Mild

Problem — data missing

Instances	Outlook	Temperature	Humidity	Windy	Play Tennis
	S	H	H	W	No
	S	H	H	S	No
	O	H	H	W	Yes
	R	M	H	W	Yes
	R	C	N	W	Yes
	R	C	N	S	No
	O	C	N	S	Yes
	S	M	H	W	No
	S	C	N	W	Yes
	R	M	N	W	Yes
	S	M	N	S	Yes
	O	M	H	S	Yes
	O	H	N	W	Yes

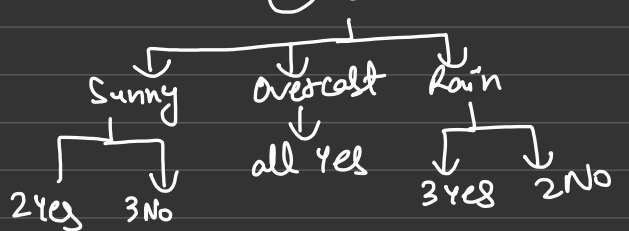
(i) Calculate the entropy for the dataset

$$\rightarrow E(S) = -\left(\frac{9}{14}\right) \times \log_2\left(\frac{9}{14}\right) - \left(\frac{5}{14}\right) \log_2\left(\frac{5}{14}\right)$$

$$= 0.94$$

(ii) Selection of Root Node from 4 features

① outlook ② Temp ③ humidity ④ Windy



Calculate Entropy for each call.

$$E(\text{outlook/Sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log\frac{3}{5} = 0.971$$

$$E(\text{outlook/Overcast}) = 0$$

$$E(\text{outlook/Rain}) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log\frac{2}{5} = 0.971$$

(iii) Calculate how much info you get from outlook

$$I(\text{outlook}) = [\text{weighted average} \times \text{Entropy (for all features)}]$$

$$= \left[\frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971\right]$$

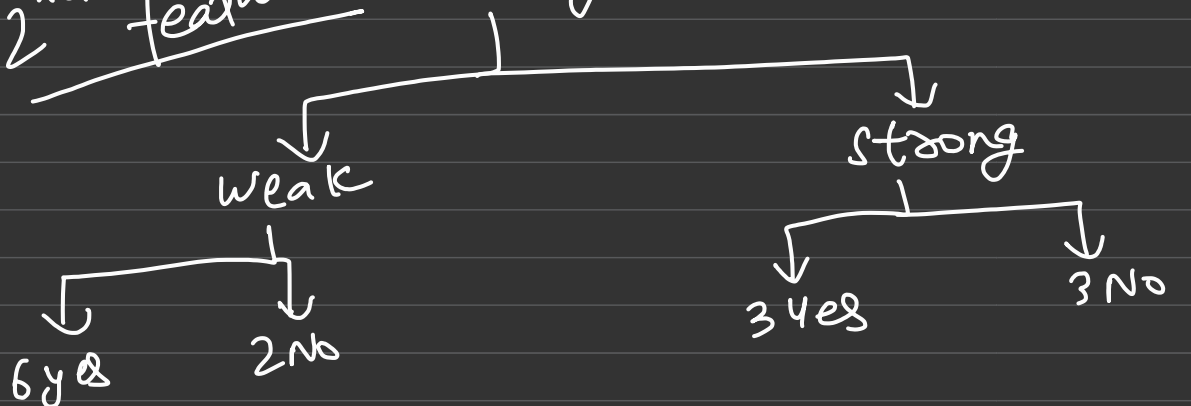
$$= 0.693$$

Information gain(outlook) = E(S) - I(outlook)

= 0.94 − 0.693

= 0.247

2nd feature : windy



$$E(\text{windy}/\text{weak}) = -\frac{6}{8}\log\frac{6}{8} - \frac{2}{8}\log\frac{2}{8} = 0.811$$

$$E(\text{windy}/\text{strong}) = -\frac{3}{6}\log\frac{6}{8} - \frac{3}{6}\log\frac{6}{8} = 1$$

$$I(\text{windy}) = \left[\frac{8}{14} \times 0.811 + \frac{6}{14} \times 1\right] = 0.892$$

Information gain(windy) = 0.94 − 0.892

= 0.048

Q)

DAY	Outlook	Temp	Humidity	Wind	Play Tennis
D_1	Sunny	Hot	High	W	N
D_2	Sunny	Hot	High	S	N
D_3	Overcast	Hot	High	W	Y
D_4	Rain	mild	High	W	Y
D_5	Rain	cool	Normal	W	Y
D_6	Rain	cool	Normal	S	N
D_7	Overcast	cool	Normal	S	Y
D_8	Sunny	mild	High	W	N
D_9	Sunny	cool	Normal	W	Y
D_{10}	Rain	mild	Normal	W	Y
D_{11}	Sunny	mild	Normal	S	Y
D_{12}	Overcast	mild	High	S	Y
D_{13}	Overcast	Hot	Normal	W	Y
D_{14}	Rain	mild	High	S	N

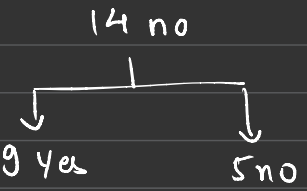
1st Method

$$I.G = E(S) - \text{Information}$$
$$= E(S) - [\text{weight} \times \text{Entropy (each feature) average}]$$

$$E(S) = -P_{yes}\log_2(P_{yes}) - P_{No}\log_2 P(No)$$

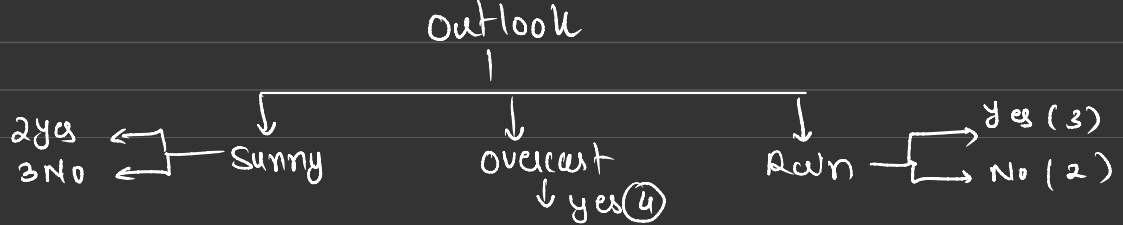
Step I: Calculate $E(S)$

$$E(S) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14} = 0.94$$



Step II Selection of Root node. Calculate I.G of each feature.

1st attribute



$$E(\text{Outlook} \mid \text{Sunny}) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$
$$= 0.971$$

$E(\text{Outlook} \mid \text{Overcast}) = 0$

$E(\text{Outlook} \mid \text{Rain}) = 0.971$

$$\text{Information (Outlook)} = \text{wt} \times \text{Entropy (each feature)}$$
$$= \frac{5}{14} \times 0.971 + \frac{5}{14} \times 0.971 = 0.693$$

$$\text{Information Gain} = E(S) - \text{Information}$$
$$= 0.94 - 0.693 = 0.247$$

The highest information gain of given feature will be the root node.

→ Then go to the child nodes and repeat the steps all over again until the decision tree is completed

•) Gini Index.

→ It is a metric used to quantify the amount of uncertainity or impurity present at a single node.

→ Opposite of information gain is gini index.

→ The feature having lowest gini index is considered as root node.

•) Frame Decision tree using Gini Index.

•> Unit - 4 (Numericals & short questions) {Recap}

→ What is Information gain?
→ What is Gini Index?
→ What is Pruning?

Long Question

→ Numericals (Form D.T.)
→ Find 1st Root node
→ Information Gain
→ highest → Root node

Decision tree is also know as ID_3 Algorithm.

⇒ ID_3 stands for Iterative Dichotomiser Algorithm
↓
It divides the nodes / Branches for better evaluation

→ ID_3 is named such because the algorithm iteratively or repeatedly dichotomizes or divides the feature into two or more groups.

→ First we are selecting the attribute / feature (Highest IG)
→ Then we stop when leaf node appears.

•) Unit - 10 PCA (Principal Component Analysis) / Dimensionality Reduction technique

⇒ "Curse of Dimentionality"

More the dimension difficult to evaluate the information.

⇒ What is PCA ?

→ PCA is a dimentionality reduction technique that helps to identify the corelation and patterns in a dataset so that it can be transformed into a dataset of significantly lower dimension without loss of any important information.

Example

A	B	C	D	Final

Original Dataset

→ Remove inconsistency
Redundant Data
High correlated feature

PCA

→

A	B	Final

Reduced dataset.

Step 1 :- Computation of mean value of variable

Step 2 :- Computation of covariance of matrix

Step 3 :- Computation of Eigen value, vector, Normalized eigen vector

Step 4 : Computation of Principal Component

Step 5 : Write the reduced dataset.

Q)

Features	Example 1	Ex 2	Ex 3	Ex 4
x	4	8	13	7
y	11	4	5	14

⇒ Given the following data use PCA to reduce the dimension from 2 to 1.

Ans → Step 1: Calculate the mean of x and y

$$\Rightarrow \bar{x} = \frac{4+8+13+7}{4} = 8$$

$$\Rightarrow \bar{y} = \frac{11+4+5+14}{4} = 8.5$$

Step 2: Covariance matrix calculation.

Ordered pairs are $(x,x)\ (x,y)\ (y,x)\ (y,y)$

$$Cov(x,x) = \frac{1}{N-1}\sum_{i=1}^{N}(x_i-\bar{x})^2 \qquad N = \text{No of sample.}$$

$$= \frac{1}{4-1}\left[(4-8)^2+(8-8)^2+(13-8)^2+(7-8)^2\right]$$

$$= 14$$

$$Cov(x,y) = \frac{1}{4-1}\left[(4-8)(11-8.5)+(8-8)(4-8.5)+(13-8)(5-8.5)+(7-8)(14-8.5)\right]$$

$$= -11$$

$$Cov(y,x) = Cov(x,y)$$

$$Cov(y,y) = \frac{1}{4-1}\left[(11-8.5)^2+(4-8.5)^2+(5-8.5)^2+(14-8.5)^2\right]$$

$$= \frac{1}{3}(6.25+20.25+12.25+30.25)$$

$$= 23$$

final Covariance matrix will be

$$S \Rightarrow \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix}_{2\times 2}$$

Step 3: Calculation of Eigen value.

$$\det |(S - \lambda I)| = 0$$

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\lambda I = \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$S - \lambda I = \begin{bmatrix} 14 & -11 \\ -11 & 23 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix}$$

$$= \begin{bmatrix} 14-\lambda & -11 \\ -11 & 23-\lambda \end{bmatrix} = 0$$

Simplify:

$$\Rightarrow (14-\lambda)(23-\lambda) - 121 = 0$$

$$\Rightarrow \lambda^2 - 37\lambda + 201 = 0$$

$$\Rightarrow \lambda_1, \lambda_2 = \frac{1}{2a}\sqrt{b^2 \pm 4ac} = \frac{1}{2}\sqrt{37^2 + 4\times 201}$$

$$\lambda_1 = 30.3849$$

$$= \frac{1}{2}\sqrt{37^2 - 4\times 201}$$

$$\lambda_2 = 6.6151$$

The value which is larger is considered as the first eigen value.

$\because \lambda_1 > \lambda_2$ we will have eigen value as 30.3849

Eigen vector for $\lambda_1 = 30.3849$.

Step 4

We will find out eigen vector of λ_1

$$(S - \lambda_1 I)\, \underline{U_1} = 0$$

$\underline{U_1}$ → eigen vector for λ_1

$$\Rightarrow \begin{bmatrix} 14-\lambda_1 & -11 \\ -11 & 23-\lambda_1 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} (14-\lambda_1)U_1 - 11U_2 \\ -11U_1 + (23-\lambda_1)U_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$(14-\lambda_1)U_1 - 11U_2 = 0 \quad - \text{eq } ①$$
$$-11U_1 + (23-\lambda_1)U_2 = 0 \quad - \text{eq } ②$$

from eqn ① $\quad \frac{U_1}{11} = \frac{U_2}{14-\lambda_1} = t$

When $t = 1$, $U_1 = 11$
$U_2 = 14 - \lambda_1$

Eigen Vector U_1 of $\lambda_1 = \begin{bmatrix} 11 \\ 14-\lambda_1 \end{bmatrix}$

Putting the value of $\lambda_1 = \begin{bmatrix} 11 \\ -16.3849 \end{bmatrix}$

c) Normalize eigen vector U_1 :

$$e_1 = \begin{bmatrix} 11 / \sqrt{(11)^2 + (-16.3849)^2} \\ -16.3849 / \sqrt{(11)^2 + (-16.3849)^2} \end{bmatrix} = \begin{bmatrix} 0.5574 \\ -0.8303 \end{bmatrix}$$

Similarly for $\lambda_2 = 6.6151$

$$e_2 = \begin{bmatrix} 0.8303 \\ 0.5574 \end{bmatrix}$$

Step 5:

first PC1	P_{11}	P_{12}	P_{13}	P_{14}

↳ Principal Component $x_1 - \bar{x}$

$$P_{11} = C_1^T \begin{bmatrix} 4-8 \\ 11-8.5 \\ y_1 - \bar{y} \end{bmatrix}$$

$$= \begin{bmatrix} 0.5574 & -0.8303 \end{bmatrix} \begin{bmatrix} -4 \\ 2.5 \end{bmatrix} = -4.3652$$

$P_{12} = 3.7361$

$P_{13} = 5.6928$

$P_{14} = -5.1238$

PC_1	-4.3062	3.7362	5.6928	-5.1238

" Solution ends Here "

• Naive Bayes Theorem / Classification

⇒ Test sample will be given & we have to find which class the test sample is using Naive Bayes.

Naive Bayes theorem = $P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$ → Liklihood

$P(A|B)$ ↓ Posterior probability

$P(B|A)$ → Prior

$P(B)$ → evidence

⇒ Posterior = $\frac{\text{Prior} \times \text{Likelihood}}{\text{evidence}}$.

Q) dataset

Input Attribute | Target attribute

Example No	Color	Type	Origin	Stolen
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				

Step 1 : Calculate priori probability

$\Rightarrow P(yes) = \frac{5}{10} = 0.5 \qquad p(No) = \frac{5}{10} = 0.5$

Step2: Calculate Conditional probability

$\Rightarrow$ Color Type Origin

Color	Yes	No
Red	3/5	2/5
Yellow	2/5	3/5

Type	Yes	No
Sport	4/5	2/5
SUV	1/5	3/5

Origin	Yes	No
Domestic	2/5	3/5
Imported	3/5	2/5

$\Rightarrow$ Step 3: Neo Instance calculation

$P(yes|\text{neo instance}) = P(yes) * P(\text{colour} = Red|yes) * P(\text{type} = SUV|yes) \times P(\text{Origin} = \text{Domatic}|yes)$

$= \frac{5}{10} * \frac{3}{5} * \frac{1}{5} * \frac{2}{5}$

$\Rightarrow \frac{1}{5} * \frac{1}{5} \times 3 \times \frac{1}{5} = \frac{3}{125} = 0.024$

$$P(\text{No} \mid \text{New instance}) = P(\text{No}) * P(\text{Color} = \text{Red} \mid \text{No}) * P(\text{Type} = \text{SUV} \mid \text{No}) * P(\text{Origin} = \text{Domestic} \mid \text{No})$$

$$= \frac{1}{10} * \frac{4}{5} * \frac{3}{5} * \frac{3}{5} = \frac{9}{125}$$

$$= 0.072$$

Compare these two posterior probability

$$P(\text{No} \mid \text{new instance}) > P(\text{yes} \mid \text{New instance})$$

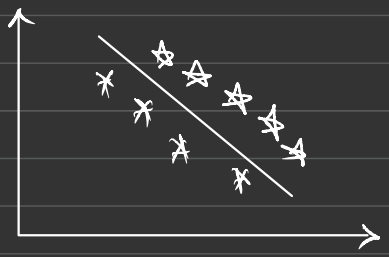
We can classify it as it belongs to No class.

Unit - II Support vector Machine (SVM)

SVM ⇒ Hyperplane is used (Objective is to find this hyperplane in n dimension)

→ What is SVM?

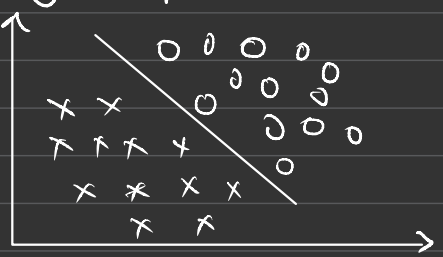
⇒ SVM is a discriminative classifier ie formaly designed by a seperative hyperplane such that the points of different catagories are seperated by a gap as wide as possible.



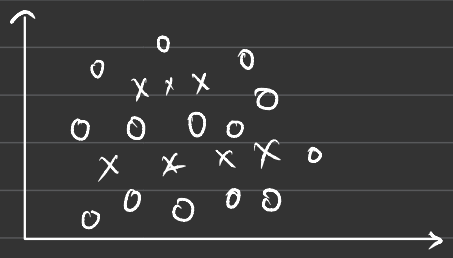
$$y = w^T x + w_0$$

Linearly Seperable Dataset

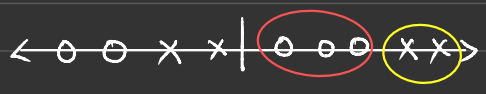
→



Non linear seperable Dataset



•> 1-D



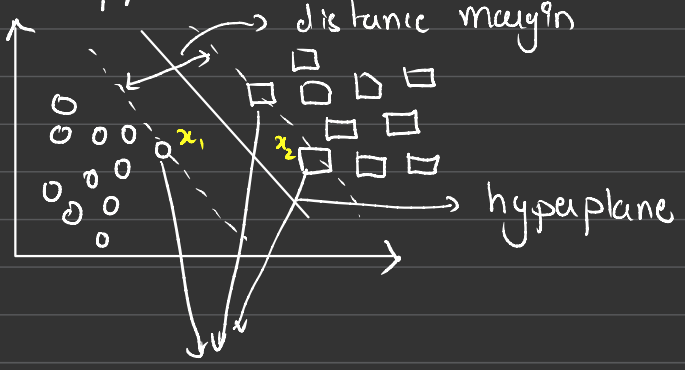
1-D plane to 2D plane

When there is no clear hyperplane it is necessary to move away from single dimensional view of the data to a 2-D view

This process is called as Kernelling where it transforms low dimensional data to high dimensional data to create a clear hyperplane.

•> Daivation of SVM

→ The Support Vectors are the extreme points in the dataset.



Hard margin

When we are strictly saying that classes should have no error & separated without any error.

$$r = \frac{g(x)}{\|w\|}$$ → equilidian norm/ distance of w.

⇒ We want to compute the distance between x_2 and x_1 so the equation will be.

$$w^T x_2 + w_0 = +1$$
$$w^T x_1 + w_0 = -1$$
$$w^T(x_2 - x_1) = +2$$

$$\frac{w^T}{\|w\|}(x_2 - x_1) = \frac{2}{\|w\|} \Rightarrow (x_2 - x_1) = \frac{2}{\|w\|}$$

⇒ Here we got our objective function, we need to update Objective function (w^{T*}, w_0^*) so that $\frac{2}{\|w\|}$ should be maximized.

$[w_1$ to $w_n]$

⇒ $\left[(w^{T*}, w_0^*) = \min \frac{\|w\|}{2}\right]$ eqn for hard margin

$$(w^{T*}, w_0^*) = \min \frac{\|w\|}{2} + C_i \sum_{i=1}^{n} \xi_i$$

- C_i: How many error model can consider
- $\sum_{i=1}^{n}$: eqn for soft margin
- ξ_i: value of error.