8 Chapters.

		Marks.
1.	Introduction to Machine Learning.	15-20 M
2.	Learning with Regression.	5-10 M
3.	Learning with Trees.	10-15M
4.	Support Vector Machine (SVM)	15 M
5.	Learning with classification.	20-30M
6.	Dimensionality Reduction	10M
7.	Learning with Clustering	15-25M
8.	Reinforcement Learning	15-20M

Scanned by CamScanner

CHAP 1 : Introduction to Machine Learning

Contents :-

- What is Machine Learning
- Key Terminology
- Types of Machine Learning
- Issues in Machine Learning
- Application of Machine Learning
- How to choose Righ algorithm.
- Steps in developing a Machine Learning Applicat

1) What is Machine Learning?

Def^n :- Machine Learning is a type of Artifical Intelligence that provides computer with the ability to Learn without being explicitly programmed.

If a computer program can improve how it perform a certain task based on a past experience than you say it has learned.

ML explores the study & construction of algorithms that can teach themselves to grow & change when exposed to new data.

Focus :- ML focuses on the development of computer programs that can teach themselves to grow & change when exposed to new data.

Scanned by CamScanner

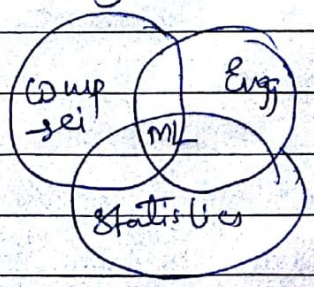
Machine learning is actively being used today perhaps in many more places than you'd expect its your freinds birthday & you want to send her a card via mail.
You Search for cards & search engine show you the 10 most relevant links you click on the second link, the search engine learns from this.

→ Examples of ML : face Recognition, handwriting Recognition, spam filtering in email & product recommendation on online sites.

→ In all but most trivial cases insight or knowledge you are trying to get out of the raw data wont be obvious from just looking at the data.

for eg: In detecting spam email, looking for the occurence of a single word may not be very helpful. But looking at the occurence of certain words used together combined with the length of email & other factors, you could get a much clearer picture of whether the email is spam or not.

→ Machine learning is turning data into Information.



ML lies at the intersection of Computer Science, Engineering & Statistics

Scanned by CamScanner

→ ML can be applied to many fields from politics to geoscience. It is a tool that can be applied to many problems. Any field that needs to interpret & act on data can benefit from machine learning techniques

→ Companies are using ML to improve business decisions, increase productivity, detect disease weather & do many more things.

→ ML uses statistics. There are many problems where the solution isn't deterministic. That is, we don't know enough about the problem or dont have enough computing power to properly model the problem. For those problems we need statistics. Eg: The motivation of human is a problem that is currently too dificult model.

How statistics work? ⟹ Take data + understand it + process it + extract value from it + Visualize it + Communicate it

Scanned by CamScanner

2) Key Terminology :-

Eg :- Bird Classification System.

Bird species classification based on 4 features

features

Sr No	wt (gm)	Wing span (mm)	Webbed feet?	Back Color	Species
1	1000.1	125.0	No	Brown	Buteo
2	3000.7	200.0	No	Gray	Sagittarius
3	3300.0	220.3	No	Gray	Sagittarius
4	4100.0	136.0	Yes	Black	Gavia

instance

target attributes

1) Expert system :-
Expert System is a system which is developed using some training set, testing set, Knowledge Representation, Features, Algorithm and Classification terminology.

2) Training set :-
A training set is the set of training example which is used to train our machine learning algorithms

3) Testing Set :-
To test ML algorithms what's usually done is to have a training set of data & a separate dataset called a test set.

Each training eg. has some features & one target variable. The target variable is what we will be trying to predict with our ML algorithms

Scanned by CamScanner

In a training set target variable is known The ML by finding some relationship between the features & target variable. In classification problem the target variable are called classes & there assumed to be a finite number of classes.

4) Knowledge Representation :
Knowledge Representation may be in the form of a set of rules. It may be a probability distribution or an example from the training set.

5) Features :
Important properties or attributes.

6) Classification : We classify the data based on features.

7) Process :-
say we have decided on a ML algorithm to use for classification what we need to do next is to train the algorithm or allow it to learn. To train the algorithm we feed it with quality data known as training set.
The target variable or class that the training example belongs to is then compared to the predicted value & we can get a sense for how accurate the algorithm is.

Scanned by CamScanner

Testing :-

wt	wingspan	webbed feet	backcolor	Species
1000.1	125.0	No	Brown	Buteo

features — target variable

features & target variables identified.

3) Types of Machine Learning :

There are two main types of Machine learning algorithms:

i) Supervised Learning
ii) Unsupervised Learning

i) Supervised learning :-

→ In Supervised learning we train the machine using data which is well labeled that means data is already tagged with correct outcome
→ we tell the algorithm what to predict.
→ Classification & Regression are examples of supervised learning.

In classification, our job is to predict what class an instance of data should fall into. Another task in ML is regression. Regression is the prediction of a numeric value. Regression problem is to fit the best-fit line through some data points to generalize the data points.

Scanned by CamScanner

ii) Unsupervised learning :-

→ In unsupervised learning we train the machine with dataset that does not have any labels.

→ There is no label or target value given for the data.

→ Example for unsupervised learning is clustering & density estimation

→ A task where we group similar items together is known as clustering. In unsupervised learning, we may also want to find statistical values that describe the data. This is known as density estimation.

Supervised learning tasks

Classification	Regression
K-nearest neighbour	Linear
Naive Bayes	Locally weighted Regression
SVM	Ridge
Decision Trees	Lasso

Unsupervised learning tasks

Clustering	Density Estimation
K-means	Expectation Maximization
DBSCAN	Parzen window.

Scanned by CamScanner

4) Issues in ML →

i) What algorithms exists for learning general target functions from specific training examples? In what setting will particular algorithms converge to the desired function given sufficient training data?
Which algorithms perform best for which type of problems & representation?

Deciding on algorithm

ii) How much training data is sufficient? What general bounds can be found to relate the confidence in learned hypothesis to the amount of training experience & the character of the learners hypothesis-space?

Training data

iii) When and how can prior knowledge held by the learner guid the process of generalizing from examples?
Can prior knowledge be helpful even when it is only approximately correct?

Prior knowledge

iv) What is the best strategy for choosing a useful next training experience, and how does the choice of this strategy after the complexity of the learning problem?

Next training experience

v) What is the best way to reduce the learning tasks to one or more function approximation problem?
Put another way, what specific function should the system attemp to learn?
Can this process itself be automated?

Learning task to function

vi) How can the learner automatically alter its representation to improve its ability to represent & learn the target function?

can m/c learn automatically?

5) How to choose the right algorithm:-

There are two important aspects that need to considered while choosing the right Machine learning algorithm:

i) first you need to consider your goal — what are you trying to get out of this?

ii) what data do you have or you can collect?

1) Goal:

→ If system is trying to predict or forecast a target value, then you need to look into Supervised learning, if not then Unsupervised learning.

→ If you have choosed Supervised learning what's your target value?

target value → discrete (yes/No, 1/2/3, A/B/C) —use→ Classification

target value → No. of values (0-100, -99 to 99) —use→ Regression

Scanned by CamScanner

If you are considering Unsupervised learning.

If you want to fit your data into some discrete groups → clustering

If you want to find numerical estimate of how strong the fit is into each group → density estimation algorithm.

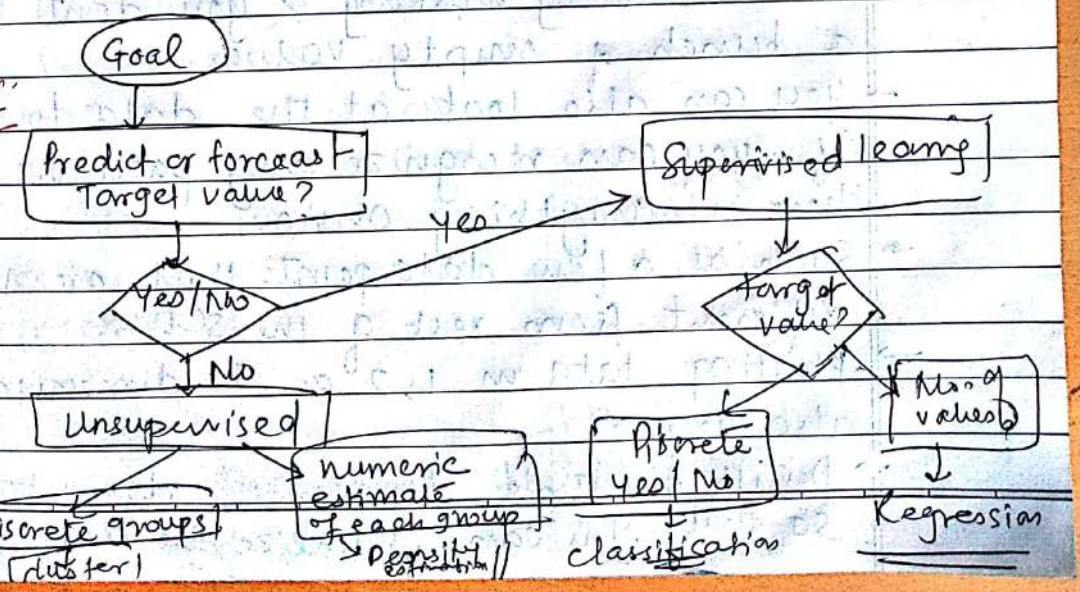
2) Based on Data

Are the features nominal or Continuos?
Are there any missing values in features?
If missing, why are there missing value?
Are there outliers in data?
All of these features about your data can help you narrow the algorithm selection process.

	Supervised	Unsupervised
Discrete	classification	clustering
continuos	Regression	density Estimation

Goal

To Remember:



6) Steps in developing a ML application.

i) Collect data :-
→ you can collect data samples by Scraping a website & extracting data.
→ from RSS feed or an API.
→ from devices for eg: blood ~~glu~~ sugar level or wind speed measurment
→ publicly available data.

ii) Prepare the input data:-
→ Once you have data, you need to make sure its in a usable format.
→ Some algo need features in a special format
→ Some algo can deal with target variables & features as string, some need them to be integer.

iii) Analyze the Input data:
→ looking at the data you have parsed in a text editor to make sure steps 1 & 2 are actualy working & you don't have a bunch of empty values.
→ you can also look at the data to see if you can recognize any patterns or if there is anything obvious.
→ Such as a few data points that are vastly different from rest of the set.
→ plotting data in 1, 2 or 3 dimensions can also help.
→ Distill multiple dimensions down to 2/3 so that you can visualize the data.

Scanned by CamScanner

iv) Cleaning the data:
- The value of this step is that it makes you understand you don't have garbage coming in.

v) Train the algorithm:
This step is where m/c learning actually takes places. This step & next step is "core" of machine learning.
You feed the algorithm with good & clean data from the first two steps & extract knowledge or information. This knowledge you often store in a format that's readily usable by machine for next two steps.
In case of unsupervised learning, there's no training step because you don't have a target value. Everything is used in the next step.

vi) Test the algorithm:
→ This is where the information learned in the previous step is put to use when you are evaluating an algorithm. You will test it to see how well it does, in case of supervised you have some known values you can use to evaluate the algorithm.
In case of unsupervised learning, you may have to use some other metrics to evaluate the success.
→ In either case, if you are not satisfied you can go back to step 4, change some things & try testing again.
→ Often collection or preparation of data may have been the problem, so you have to go back to step 1.

Scanned by CamScanner

vii) Use it :-

→ Here you make a real program to do some task & once again you see if all the previous steps worked as you expected.

→ you migh encounter some new data & have to revisit steps 1-5.

7) Applications of Machine learning :

① Learning Associations :-

→ Association between products or services is discovered.

→ In the case of retail for eg. a supermarket chain one application of ML is Market basket analysis, which finding associations between products bought by customers.

→ If people buy X typically also buy Y & if there is a customer who buys X & does not buy Y, he or she is a potential Y customer. Once we find such customers we can target them for cross-selling.

$$X \longrightarrow Y$$

→ Association rule are determined using conditional probability.

$P(Y|X)$ → where Y is product to be conditioned on X.

eg. $P(\text{milk}|\text{bread}) = 0.7$

Indicates 70% of customers who buy bread also buy milk.

Scanned by CamScanner

ii) Classification :-

→ Determines data fall into which class.

→ Eg: Credit Scoring system.

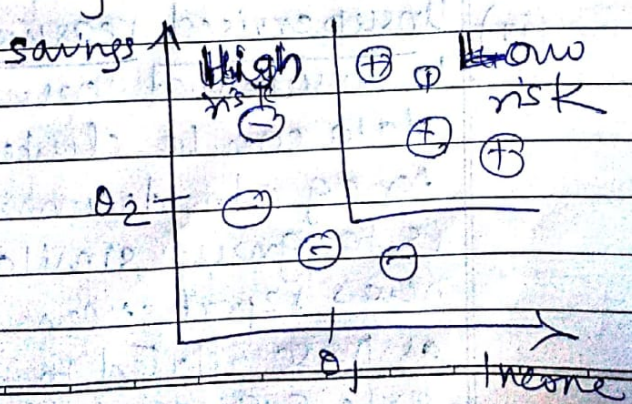
→ Credit is an amount of money loaned by a financial institute. It is important for the bank to be able to predict in advance the risk associated with a loan.
which is probability that the customer will default & not pay the whole amount back.

→ In Credit Scoring, the bank calculates the Risk given the amount of credit & the information about the customer (income, saving, collaterals, profession, age, past financial history).

→ From this data the aim is to infer a general rule coding the association between a customers attributes & his risk.

→ ML system fits a model to the past data to be able to calculate the risk for a new application & then decides to accept or refuse it accordingly.

If income > θ_1 AND saving > θ_2
then low-risk else High-risk.



other classification eg:
- optimal character recognition
- face recognition
- medical diagnosis
- speech recognition
- Biometric

iii) Regression :-

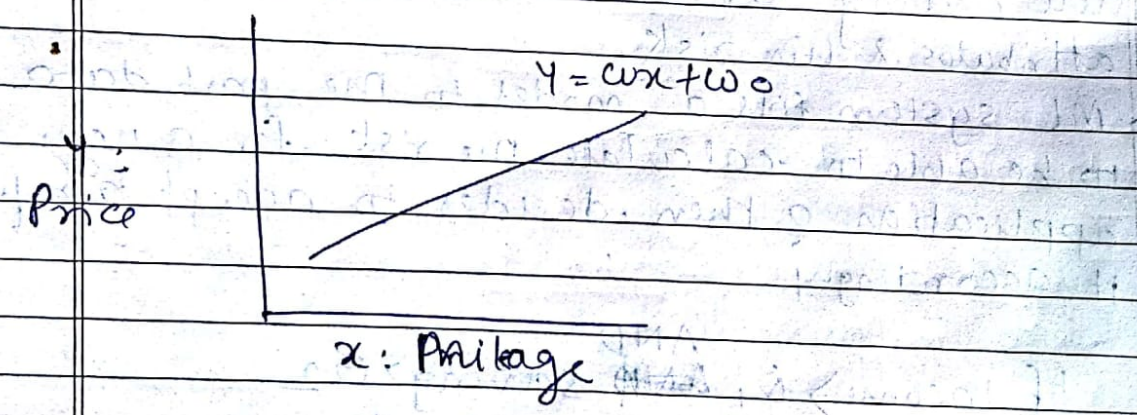
- prediction of numeric value.
- Eg: we want to predict the price of used car.

Inputs are features of car - Brand, Year, Engine capacity, Milage & Other Information that we believe affect a car's worth.

- The o/p is the price of a car.
- let x denotes car features / attributes & y denotes price.

Again surveying past transaction we can collect a training data & the ML algo fits a function to this data to learn y as a function of x

$$\boxed{y = \omega x + \omega_0}$$



iv) Unsupervised Learning :-

In unsupervised learning similar types of data can be clustered.

for eg: In document clustering the aim is to group similar documents for eg: news reports can be subdevided as those related to politics, sports, fashion, arts & so on.

Scanned by CamScanner

→ Commonly, a document is represented as a bag of words. ie we predefine a lexicon of N words & each document is represented as an N-dimensional binary vector whose element i is 1 if word i appears in a document. suffixes "-s" & "-ing" are removed to avoid duplicats & word such as "of" and which are not informative are not used. Documents are then grouped depending on number of shared words.

v) Reinforcement Learning :-

- In some applications the o/p of the system is a sequence of actions. In such case single action is not Important. What is Important is the policy that is the sequence of correct actions to reach goal.
- An action is good if it is a part of good policy. In such case ML program should be able to access the goodness of policies & learn from past good action sequences to be able to generate a policy.
- Such methods are called reinforcement learning.

Eg: 1) Game playing → Single Move is not Important. It is the sequence of right moves that is good.

2) Robot Navigation.

Questions

1) What are the Key tasks of Machine Learning? 5M (May-16)
2) Explain the steps required for selecting the right ML algorithm 8M – (May-16)
3) Application of ML (16) (May-16) (Dec16) Issues in ML (5) (Dec16)

Chap 2 : learning with Regression

- Contents :-
  - Linear Regression
  - Logistic Regression

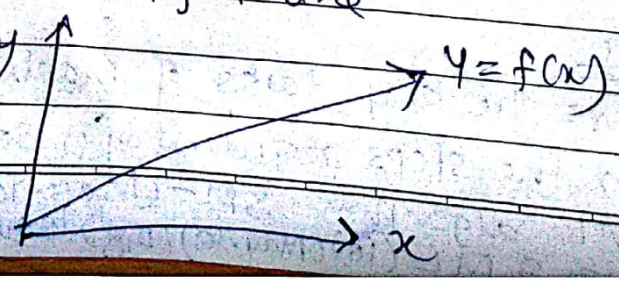
→ **What is Regression:**
- Analyze the Relationship between two or more variables.
- Give a value of x & y, find a function (f(x)) that can predict value of y for given x

→ **Simple Linear Regression:**
- In simple linear regression, we predict the scores on one variable from the scores on a second variable.
- Two type of variables are used: Dependent & Independent variables.

→ Independent variables are Input to the system, on which we are basing our prediction. It is also called as predictor variable denoted as "x".

→ Dependent variables – those change as a consequence of change in other values in the s/m. The variable which we are predicting. It is also called as Criterion variable. denoted as "y".

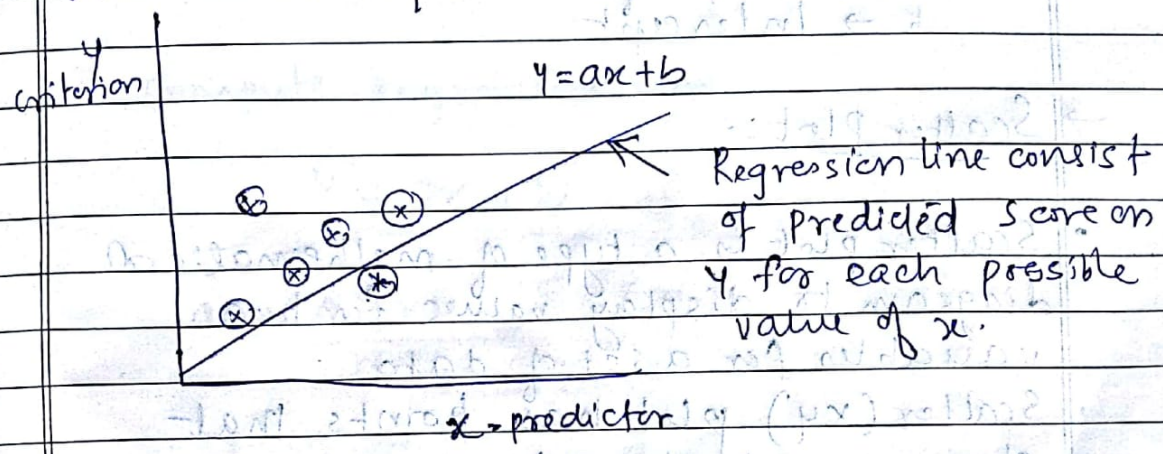
- In simple linear regression the prediction of y when plotted as a function of x forms a straight line



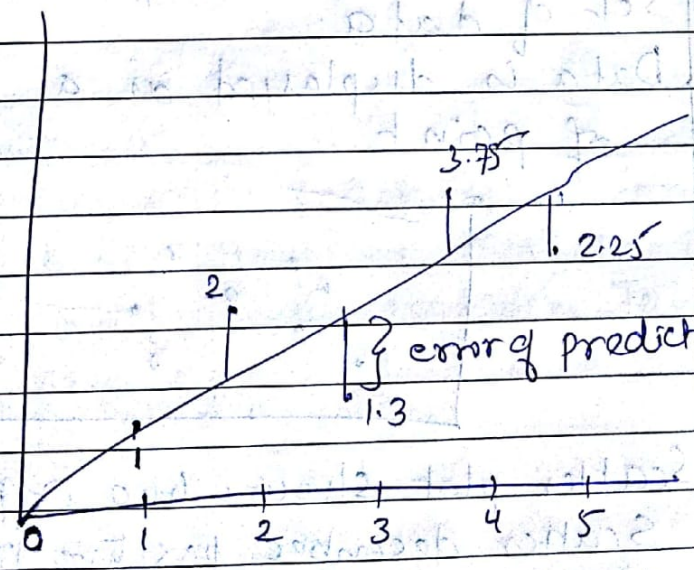
Scanned by CamScanner

→ Regression line :

- Linear Regression consist of finding the best fitting straight line through the points.
- Best fitting is the line that is closest to all the data points.
- It is called as Regression line.
- Also called as "trend line", or "line of best fit" or "least square line".



X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25



Error of Prediction = (value of Point - Predicted value)

The most commonly used criterion for best fitting line that minimizes sum of squared errors of prediction"

Regression line has an equation of the form

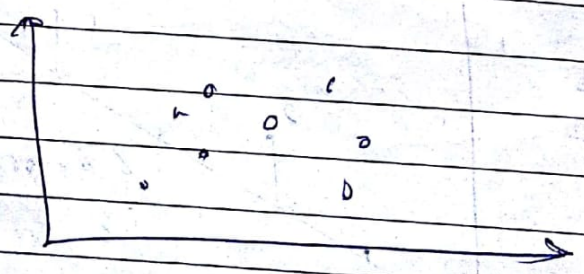
$$\boxed{y = ax + b}$$

where

$x \rightarrow$ Predictor variable
$y \rightarrow$ Criteria variable
$a \rightarrow$ slope of line
$b \rightarrow$ Intercept.

→ Scatter plot :-

- Scatter plot is a type of mathematical diagram to display values for two variables for a set of data
- Scatter (xy) plot has points that show the relationship between two set of data.
- Data is displayed as a collection of point

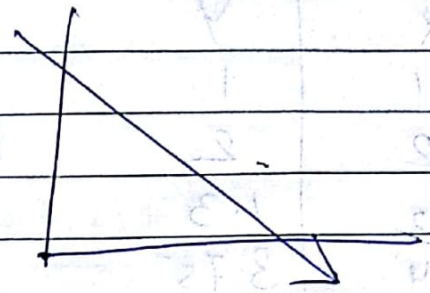


- Scatter plot shows two type of trend :
  Scatter describes positive trend.
  If one set of value increases ⇒ other set tends to increase.

x	y
1	1
2	2
3	2.45
4	3.75

- Scatter describes negative trend.
  If one set of value increases → other tends to decrease

x	y
1	6
2	3.75
3	3
4	2



→ To compute regression line

$$y' = ax + b$$

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$b = \frac{1}{n}\left[\Sigma y - a\Sigma x\right]$$

MU Questions:-

1. Explain Regression line, Scatter plot, Error in prediction & Best fitting line (5M) (Dec 16)
2. Explain in brief linear Regression Technique (5M) (May-16)

Scanned by CamScanner

→ Problems on Linear Regression :-

1)

X	Y
1	1
2	2
3	1.3
4	3.75
5	2.25

Find Regression Line

Solution :-

To find

$$y' = ax + b$$

x	y	xy	x^2
1	1	1	1
2	2	4	4
3	1.3	3.9	9
4	3.75	15	16
5	2.25	11.25	25
$\Sigma x = 15$	$\Sigma y = 10.3$	$\Sigma xy = 35.15$	$\Sigma x^2 = 55$

$$\boxed{n = 5}$$

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{5 \times 35.15 - 15 \times 10.3}{5 \times 55 - (15)^2}$$

$$\boxed{a = 0.425}$$

Scanned by CamScanner

$b = 1/n\,[\Sigma y - a\Sigma x]$

$b = \frac{1}{5}[10.3 - 0.425 \times 15]$

$= \frac{1}{5}[3.925] = 0.785$

$\boxed{b = 0.785}$

$y' = ax + b$

$\boxed{y' = 0.425x + 0.785}$

To calculate sum of squared error:-

x	y	y'	Error(y-y')	Squared error(y-y')²
1	1	1.21	-0.21	0.044
2	2	1.635	0.365	0.133
3	1.3	2.06	-0.760	0.578
4	3.75	2.485	1.265	1.6
5	2.25	2.910	-0.660	0.436
				2.791

(2) (1,10) (2,11) (3,15) (4,12) (5,17) (6,18) (7,20) find Regression line.

x	y	xy	x^2
1	10	10	1
2	11	22	4
3	15	45	9
4	12	48	16
5	17	85	25
6	18	108	36
7	20	140	49
$\Sigma x = 28$	$\Sigma y = 103$	$\Sigma xy = 458$	$\Sigma x^2 = 140$

Scanned by CamScanner

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2} = \frac{7\times 458 - 28\times 103}{7\times 140 - (28)^2}$$

$$a = 1.642$$

$$\boxed{a = 1.642}$$

$$b = \frac{1}{n}\left[\Sigma y - a\Sigma x\right]$$

$$= \frac{1}{7}\left[103 - 1.642\times 28\right]$$

$$b = 8.147$$

Regression line

$$y' = ax + b$$

$$\boxed{y' = 1.64x + 8.147}$$

3) The sales of a company (in million dollars) for each year are shown in table below

x (year)	2005	2006	2007	2008	2009
y (sales)	12	19	29	37	45

(a) find the least square regression line $y = ax + b$

(b) use line as a model to estimate the sales of a company in 2012.

Sol^n:-

let us consider

$$t = x - 2005$$

Scanned by CamScanner

t (year after 2005)	0	1	2	3	4
y	12	19	29	37	45

t	y	ty	t^2
0	12	0	0
1	19	19	1
2	29	58	4
3	37	111	9
4	45	180	16
$\Sigma t = 10$	$\Sigma y = 142$	$\Sigma ty = 368$	$\Sigma t^2 = 30$

$\boxed{n = 5}$

$$a = \frac{n\Sigma ty - \Sigma t \Sigma y}{n\Sigma t^2 - (\Sigma t)^2} = \frac{5 \times 368 - 10 \times 142}{5 \times 30 - (10)^2}$$

$$= \frac{5 \times 360 - 10 \times 142}{5 \times 30 - 100}$$

$$= 8.4$$

$$\boxed{a = 8.4}$$

$$b = \frac{1}{n}\left[\Sigma y - a\Sigma t\right]$$

$$= \frac{1}{5}\left[142 - 8.4 \times 10\right]$$

$$\boxed{b = 11.6}$$

$$y = ax + b$$

$$\boxed{y = 8.4t + 11.6}$$

Scanned by CamScanner

b) In 2012

$t = x - 2005$

$= 2012 - 2005$

$= 7$

$t = 7$

$y = 8.4t + 11.6$

$= 8.4 \times 7 + 11.6$

$\boxed{y = 70.4}$ - million dollars.

4)

x	y	xy	x^2
55	52	2860	3025
60	54	3240	3600
65	56	3640	4225
70	58	4060	4900
80	62	4960	6400
330	282	18760	22150

$n = 5$

$y = ax + b$

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{5 \times 18760 - 330 \times 282}{5 \times 22150 - 330 \times 330}$$

$= 0.42$

$\boxed{a = 0.42}$

Scanned by CamScanner

$$b = \frac{1}{n}[\Sigma y - a\Sigma x] = \frac{1}{5}[282 - 0.42 \times 330]$$

$$= 28.68$$

$$b = 28.68$$

$$y = ax + b$$

$$\boxed{y = 0.42x + 28.68}$$

MU - May 17

(5) The following table shows the midterm & final exam grades obtained for a students in a database course.

Midterm (x)	final exam (y)
72	84
50	63
81	77
74	78
94	90
82	75
59	49
85	79
83	77
33	52
83	74
81	50

Use the method of least square regression to predict the final exam grade of a student who received 86 on Midterm exam.

Soln:-

$n = 12$

Scanned by CamScanner

x	y	xy	x^2	x^2
72	84	6048	5184	5184
50	63	3150	2500	2500
81	77	6237	6561	6561
74	78	5772	5476	5476
94	90	8460	8836	8836
82	75	6150	6724	6724
59	49	2891	3481	3481
85	79	6715	7225	7225
83	77	6391	6889	6889
33	52	1716	1089	1089
83	74	6142	~~6889~~ 6889	6889
81	50	4050	6561	6561
877	848	63722		67415

$$a = \frac{n\Sigma xy - \Sigma x \Sigma y}{n\Sigma x^2 - (\Sigma x)^2}$$

$$= \frac{12 \times 63722 - 877 \times 848}{12 \times 67415 - 877 \times 877}$$

$$\boxed{a = 0.526}$$

$$b = \frac{1}{n}\left[\Sigma y - a\Sigma x\right]$$

$$= \frac{1}{12}\left[848 - 0.526 \times 877\right]$$

$$\boxed{b = 32.225}$$

$$y = ax + b$$

$$\boxed{y = 0.526x + 32.22}$$

Scanned by CamScanner

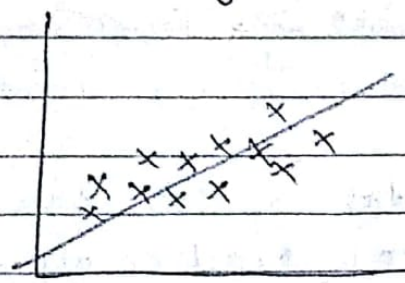
for $x = 86$

$y = 0.526 \times 86 + 32.22$

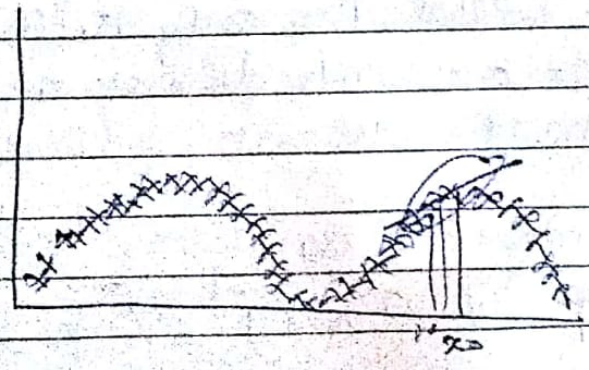
$\boxed{y = 77.456}$

→ **locally weighted Regression:-**

In simple linear regression, we have some data points & we try to find approximation function to generalize these data points.



One of the problem with linear regression is it tries to fit a constant line to your data once the model was created.
Such behaviour is OK when your data follows linear pattern & does not have much noise. However when data set is not ~~not~~ linear, linear regression tends to underfit the training data.



Scanned by CamScanner

Locally weighted Regression tries to solve this problem by assigning weight to your training data. Weights are bigger for the data points that are closer to the data you are trying to predict, thus local in the name.

Locally → local means using nearby po
Weight → we value points based on how far they are.
Regression → approximating a function

How it works?
1) build a local model of the function
2) Use the model to predict o/p value
3) Throw away the model

→ LWR requires entire dataset everytime you try to make prediction making it much more computationally expensive compared to linear regression.

→ LWR is similar to simple linear regression model, the only different is that we are introducing a weight matrix W.

model Parameter as

$$\beta = (X^T W X)^{-1} X^T W Y$$

Scanned by CamScanner

To get prediction we need to multiply betas with our input X_0

$$y' = \beta X_0$$

→ To find weights Kernel Smoothing is used. Kernel Smoothing allows us to smooth out our data points of interest (X_0) using nearby data points (X).

→ The Kernel is usually defined as D or W. Whose value are decreasing as the distance between X_0 & X increases.

$$D = \frac{\| X - X_0 \|}{h_\lambda(X_0)}$$

→ To construct W we need to evaluate Kernel for each training i/p (X) & the value we are trying to predict, X_0. W is diagonal where all non-diagonal cells are 0.

→ After Running the Kernel on our dataset we will essentially create a W where weights are decreasing as the distance betn X_0 & X increases. This Kernel property will make nearby points more important when solving linear regression.

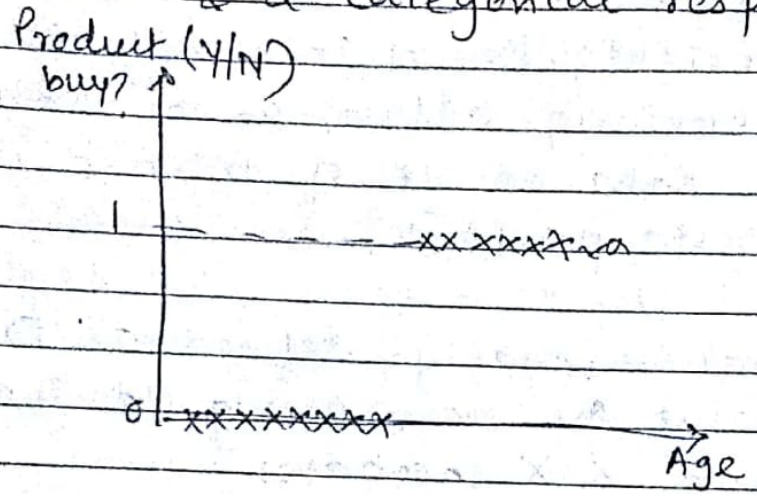
Gaussian Kernel

$$D = a \cdot e^{-\frac{|x - x_0|}{2k^2}}$$

Scanned by CamScanner

→ Logistic Regression :-

→ It models the relationship between predictor variables & a categorical response variable



In LR we have bunch of data & with the data we try to build an equation to do classification.
LR helps us to estimate a probability of falling into a certain level of the categorical response given a set of predictors.

Eg: We could use Logistic Regression to model the relationship betn various measurements of a manufactured specimen (such as dimensions & chemical composition) to predict if crack > 10 mils will occur or not.

Scanned by CamScanner

Types of LR :-

i) Binary LR : Used when response is binary.
The cracking eg above.

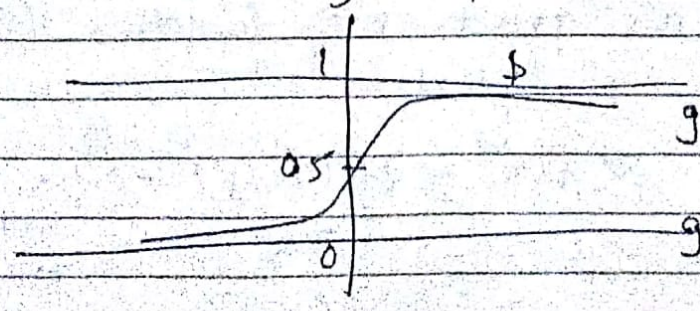
ii) Nominal LR : used when there are 3 or more categories with no natural ordering to the levels.
Eg : department at business (Marketing, HR, accounts)

iii) Ordinal LR : Used when there are 3 or more categories with a natural ordering to the levels, but the ranking of level do not necessarily means the interval between them are equal.
Eg : How student rate effectiveness of a college on scale 1-5.

→ For us Regression aspect means that we try to find a best-fit set of parameter.
Logistic Regression is finding bet-fit parameter to a non-linear function called sigmoid.

$$\sigma(z) = \frac{1}{1+e^{-z}}$$

It have following shape



$g(z) \to 1$ as $z \to \infty$

$g(z) \to 0$ as $z \to -\infty$

Scanned by CamScanner

for logistic regression classifier we will take our features & multiple each one by a weight & then add them up

$$z = w_0 x_0 + w_1 x_1 + \cdots + w_n x_n$$

$$\boxed{z = w^T x}$$

This z value will be put into sigmoid & we will get no. between 0 & 1. Anything above 0.5 will be classified as 1 & anything below 0.5 will be classified as 0.

In $z = w^T x$
vector x is our input data. & we want to find the best coefficient w, so that this classifier will be as succesful as possible.
In order to do that, we need to consider the ideas from optimization theory.

Gradient ascent :- Based on idea that if we want to find maximum point on a function the best way to move is in the direction of gradient.

we write the gradient function with symbol ∇ & the gradient of function $f(x,y)$ is given by equation

$$\nabla f(x) = \frac{\partial f(x,y)}{\partial x} \rightarrow \text{we will move in x direction by amt } \frac{\partial f(x,y)}{\partial x}$$

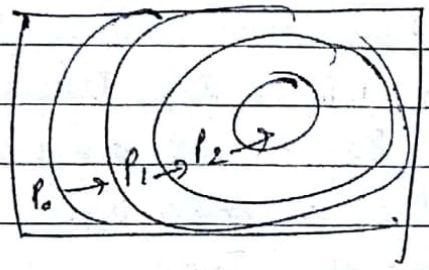
$$= \frac{\partial f(x,y)}{\partial y} \rightarrow \text{move in y dir.}$$

Scanned by CamScanner

The gradient Operator will always point in the direction of greatest increase.

$$w = w + \alpha * \nabla f(w)$$

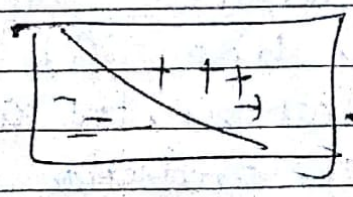
This step is repeated until we reach stopping condition



Gradient ascent algo moves in the direction of gradient evaluated at each point starting with P_0 then P_1 then P_2 until stop condition. Gradient operator alway ensure that we move in the best direction.

Algorithms :-

Start with the weights all set to 1
Repeat R no. of times
calculate the gradient of entire dataset
update the weight vector by α * gradient
Return the weight vector.



Gradient descent :-

$$w = w - \alpha \nabla f(w)$$

with gradient descent we are trying to minimize some function.

Chap 8 : Learning with Trees

- Contents –
  - Using Decision Trees
  - Construction of decision Trees
  - Classification & Regression Trees (CART)

→ Using Decision Trees
→ What is Decision Tree
→ Appropriate problems for Decision Tree learning
→ Decision Tree Representation
→ Attribute Selection Measure
→ Issues in Decision Tree
→ Strength & weakness of DT.

1) <u>What is Decision Trees :-</u>

→ Decision Trees can be defined as Tree-based classifier.
→ DT are important & powerful tools for classification & prediction.
→ The attractiveness of DT is due to the fact that decision trees can be represented in the form of Rules. Rules can be readily be expressed so that humans can understand them or even directly used in a database access language like SQL.
→ In some applications, the accuracy of a classification or prediction is the only thing that matters in such situations we doesn't necessarily care how or why the model work. In other situation, the ability to explain the reason for a decision is crucial.

Scanned by CamScanner

→ DT is a type of supervised learning where the data is continuously split according to a certain parameter.

In marketing one has described the customer segments to marketing professional. So that they can utilize this knowledge in launching a successful marketing campaign.

2) Decision Tree Representation :-
Decision Tree has two type of nodes.

i) Decision node :- specifies some test to be carried out on a single attribute value with one branch & subtree for each possible outcome of test

ii) Leaf node: Indicates the value of the target attribute (class) of example

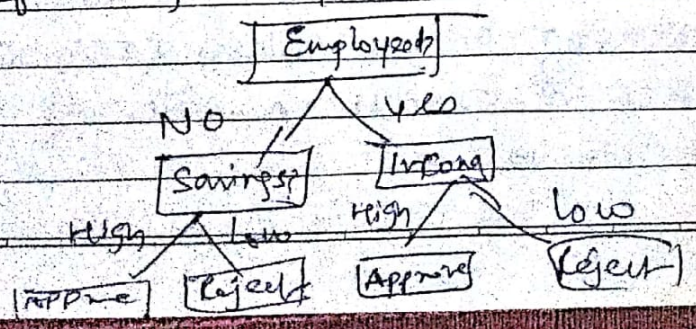
Decision Tree can be used for both classification & Regression.

1) classification - decision variable is categorical for eg : Yes / No.

2) Regression : decision or outcome variable is continuous eg : 123

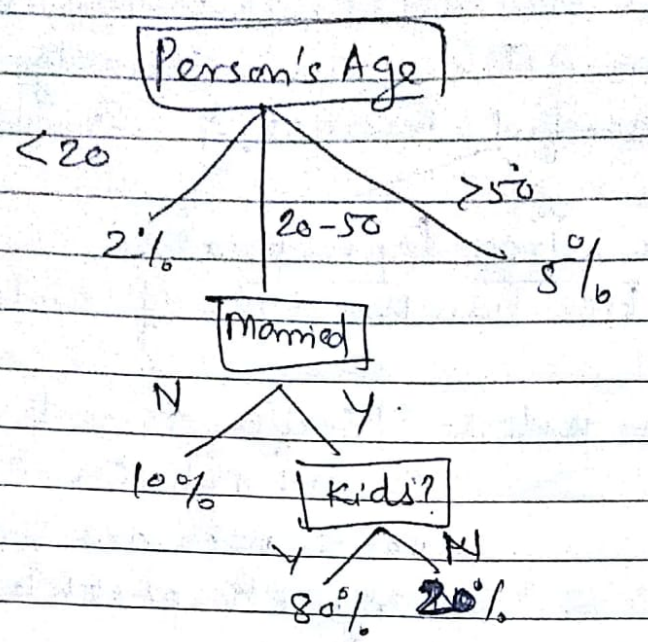
However most popularly used for classification

classification eg :- DT about whether to approve loan or not



Scanned by CamScanner

Regression eg :- Marketing of ~~Loan~~ Loan



3) Appropriate problems for Decision Tree Learning :

→ Decision tree Learning is generally best suited to problems with the following characteristics

~~i) Instances are described by a fixed set of attributes and the~~

i) Instances are represented by a attribute-value pair

- Instances are described by a fixed set of attributes and their values.
  for eg: temprature, value => hot/mild/cool.

Attr - value pair

- The easiest situation for decision tree learning occurs when each attribute takes on small number of disjoint possible values.

Scanned by CamScanner

ii) The target function has discrete o/p values.
- A DT assigns a classification to each example.

iii) Simplest case exist when there are only two possible classes (Boolean classification)

iv) DT method can also be easily extended to learning function with more than two possible o/p values.
- Real valued O/Ps.

v) The training data may contains errors.
- DT methods are robust to error - both errors in classificatn of the training examples & errors in the attr value that describe these examples.

vi) Disjunctive ~~Repre~~ description may be required
- DT naturally represent disjunctive expressions.

vii) Training data may contain missing values.
- DT methods can be used even when some traing examples have unknown values

4) Attribute Selection Measure :-

i) Gini Index :-
- All attributes are assumed continuous valued
- Assume there exist several possible split values for each attribute
- May need other tools such as clustering to get the possible split values
- Can be modified for categorical attributes
- Gini is used in CART.

Scanned by CamScanner

If a dataset T contains eg from n classes giniindex, gini(T) is defined as

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2$$

where p_j → relative frequency of class j in T.

After spliting T into two subsets T_1 & T_2 with sizes N_1 & N_2. gini index of split data is

$$gini_{split}(T) = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

The attribute providing smallest $gini_{split}(T)$ is chosen to split the node.

Eg:-

Outlook	Temp	Humidity	Windy	Play?
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
overcast	Hot	High	False	Yes
Rainy	mild	High	False	Yes
Rainy	cool	normal	false	Yes
Rainy	cool	normal	~~Fal~~ True	NO
Overcast	cool	normal	true	yes
sunny	mild	high	false	No
sunny	cool	normal	false	yes
Rainy	mild	normal	false	yes
sunny	mild	normal	True	yes
overcast	mild	High	true	yes
overcast	Hot	normal	false	yes
Rainy	mild	High	true	yes

Scanned by CamScanner

$$gini(D) = 1 - \left(\left(\frac{9}{14}\right)^2 + \left(\frac{5}{14}\right)^2\right)$$

$$= 0.459$$

$$split(D) = \frac{N_1}{N} gini(D_1) + \frac{N_2}{N} gini(D_2) + \cdots$$

Eg:- Outlook split into 5, 4, 5

$$Split = \frac{5}{14} gini(sunny) + \frac{4}{14} gini(overcast) + \frac{5}{14} gini(rainy)$$

$$sunny = 1 - \left(\left(\frac{2}{5}\right)^2 + \left(\frac{3}{5}\right)^2\right) = 0.48$$

$$overcast = 1 - \left(\frac{4}{4}^2 + \frac{0}{4}^2\right) = 0$$

$$rainy = sunny = 0.48$$

$$Split = \frac{5}{14} * 0.48 * 2 = 0.342$$

The attr that generates smallest gini split value is chosen to split the node on.

Scanned by CamScanner

ii) Information Gain (ID3) :-

→ All attributes are assumed to be categorical

→ Can be modified for continuous valued attributes.

→ Select the attr with highest information gain.

→ assume there are 2 classes P & N. Let the set of examples S contain p elements of class P & n elements of class N

→ The amount of Information needed to decide if an arbitary example in S belongs to P or N is defined as

$$I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$$

Assume that using attributes A, a set S will be partitioned into set $\{S_1, S_2, \ldots S_n\}$ if S_i contains p_i examples of P & n_i examples of N. The entropy or the expected information needed to classify objects in all subtrees S_i is.

$$E(A) = \sum_{i=1}^{v} \frac{p_i+n_i}{p+n} I(p_i, n_i)$$

Entropy (E) : It measures the amount of uncertainty or randomness in data.

E is expected amt of informati

Scanned by CamScanner

needed to assign a class to a randomly drawn object in S under the optimal, shortest-length code.

Gain (A) : Measures the reduction in Entropy acheived because of Split. Choose Split that acheives most reduction i.e maximum Gain.

$$Gain(A) = I(p,n) - E(A).$$

5) Issues in Decision Tree Learning :-

Practical Issues in learning DT includes determing how deeply to grow the decision tree, handling continuos attributes, choosing an appropriate attribute selection Measure, Handling training data with missing attribute & handling ~~training data~~ attributes with different cost.

1) Avoid Overfitting data :-

- The ID3 algorithms grows each branch of the tree just deeply enough to perfectly classify the training examples. While this is sometimes a reasonable strategy infact it can lead to difficulties when there is noise in data, or when the number of training example is too small to produce a representative sample of the target function. This simple algorithm can produce tree that overfits the training example.

Scanned by CamScanner

Two approaches to avoid Overfitting,
prune the tree so that it is not too specific.

i) Pre-pruning (prune while building tree):
Halt tree construction early - do not split a node if this would result in the goodness measure falling below threshold

ii) post-pruning (prune after building tree):
"fully grown" tree get a sequence of progressively pruned trees.

ii) Incorporating Continuos-Valued Attributes
→ Initial defination of ID3 is restricted to attributes that takes on discrete set of values. First the target attribute whose value is predicted by the learned tree must be discrete valued. Second, the attribute tested in decision node of the tree must also be discrete valued. This second restriction can easily be removed so that continuos valued decision attr can be incorporated into learning tree.
→ This can be accopmlished by dynamically defining new discrete valued attributes that can partition the continuos attr value into discrete set of intervals.
→ Threshold will defined
Eg:-

Temp	40	48	60	72	80	90
Play tennis:-	No	No	Yes	Yes	Yes	NO

We would like to pick a threshold c that produces greatest Information gain.

$(48+60)/2 = 54$ $(80+90)/2 = 85$

The IG can then be computed for each of the candidate attributes. Temp ≥ 54 and Temp ≥ 85. The best can be selected. Temp ≥ 54

iii) Attribute Measures for Selecting Attributes:

There is a natural bias in the information gain measure that favours attributes with many values over those with few values.

As an eg: consider Date, which has a very large number of possible values. If we were to add this attr to the data, it would have the highest information gain of any of the attributes.

→ This is because Date alone perfectly predicts the target attribute over the training data.

→ It will have very high information gain, despite being very poor predictor of the target function.

→ One way to avoid this difficulty is to select decision attributes based on some measure other than IG. One alternative measure that has been used successfully is the gain ratio.

→ The gain ratio penalizes attr such as Date by incorporating a term called Split Information ie: sensitive to how broadly & uniformly the attr splits the data:

$$\text{SplitInformation}(S,A) = -\sum_{i=1}^{c} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$$

$$\text{Gain ratio}(S,A) = \frac{\text{Gain}(S,A)}{\text{SplitInformation}(S,A)}$$

Scanned by CamScanner

iv) Handling Training examples with missing attribute values :-
In certain cases, the available data may be missing values for some attributes.
There are two ways to deal with missing values :
i) To assign it the value that is most common among training examples.
ii) Assign a probability to each of possible value of attribute rather than simply assigning common values.

v) Handling attributes with Different costs :-
→ In some learning task attributes may have associated cost.
→ Eg :- Classify medical disease.
attr → Temp, Biopsy Result, Pulse, Blood Test Result
These attr vary significantly in their costs.
→ In such task, we would prefer decision tree that use low-cost attr where possible relying on high-cost attr only when need to produce reliable classifications.
→ We might devide the Gain by the cost of attr.

$$\frac{Gain(S, A)}{cost\ A}$$

6) Strength & Weakness of DT :-
Strengths :-
1) Easy to Understand - able to generate understandable rule
2) perform classification without requiring much computation
3) Provides clear Indication of which fields are most imp for prediction & classification.

Scanned by CamScanner

(i) less data cleaning is required - can handle missing values.

Weakness: i) Not suitable for prediction of cont attr.
ii) Perform poorly with many classes & small data set.
iii) Sometimes does overfitting of data.

* Constructing Decision Trees :-

→ ID3 (Iterative Dichotomiser 3) algorithm is used to build a decision tree.

→ Given a set of non-categorical attributes C_1, C_2, C_n, the categorical attribute C & training set T of records:

function ID3 (R: a set of non-categorical attr
C: the categorical attr,
S: a training set) returns a decision tree;

begin
If S is empty, return a single node with value failure;
If S consists of records all with the same value for categorical attr, return a single node with that value;
If R is empty, then return a single node with value the most frequent of the values of the categorical attr that are found in records of S;

Let D be the attribute with largest Gain(D,S) among attributes R;
Let $\{d_j \mid j=1,2,\ldots m\}$ be the value of attr D;
Let $\{S_j \mid j=1,2\ldots m\}$ be the subset of S consisting respectively of records with value d_j for attribute D;

Return a tree with root labeled D and arcs labeled $d_1, d_2 \ldots d_m$ going respectively to the trees.

ID3 (R − {D}, C, S1) , ID3 (R − {D}, C, S2), C

end ID3;

Example 1:-

Target classification should we play baseball which can be yes or No.

Day	outlook	Temp	Humidity	windy	Play
D1	sunny	Hot	High	weak	No
D2	sunny	Hot	High	strong	No
D3	overcast	Hot	High	weak	Yes
D4	Rainy	mild	High	weak	Yes
D5	Rainy	cool	normal	weak	Yes
D6	Rainy	cool	normal	Strong	No
D7	overcast	cool	normal	strong	Yes
D8	sunny	mild	High	weak	No
D9	sunny	cool	Normal	weak	Yes
D10	Rainy	mild	Normal	weak	Yes
D11	Sunny	mild	Normal	strong	Yes
D12	Overcast	mild	High	strong	Yes
D13	Overcast	hot	Normal	weak	Yes
D14	Rainy	mild	High	strong	No

Scanned by CamScanner

Solution :-

class P : playball = "yes"
class N : playball = "No"
Total No of records = 14

no. of records with yes = 9
no = 5

Information gain $I(p,n) = -\frac{p}{p+n}\log_2\frac{p}{p+n} - \frac{n}{p+n}\log_2\frac{n}{p+n}$

$I(9,5) = -\frac{9}{14}\log_2\frac{9}{14} - \frac{5}{14}\log_2\frac{5}{14}$

$I(9,5) = 0.940$

Step 1 :- Compute Gain for attributes outlook, Temp, Humidity & windy.

i) compute entropy for outlook : Sunny, overcast, Rainy.

for Outlook = Sunny.
$p_i = 2$
$n_i = 3$ total = 5

$I(p_i, n_i) = I(2,3) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.971$

for outlook

Outlook	p_i	n_i	$I(p_i,n_i)$
Sunny	2	3	0.971
overcast	4	0	0
Rainy	3	2	0.971

Scanned by CamScanner

$$E(A) = \sum_{i=1}^{v} \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

$$E(\text{outlook}) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,$$

$$= \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971$$

$E(\text{outlook}) = 0.694$

$$\text{Gain}(S, \text{outlook}) = I(p,n) - E(\text{outlook})$$
$$= 0.940 - 0.694$$
$$= 0.246$$

ii) Entropy for Temp: Hot, Mild, Cool.

Temp	p_i	n_i	$I(p_i, n_i)$
Hot	2	2	1
Mild	4	2	0.920
Cool	3	1	0.81

$E(\text{Temp}) = 0.911$

$\text{Gain}(S, \text{Temp}) = 0.029$

Humidity	p_i	n_i	I
high	3	4	
normal	6	1	

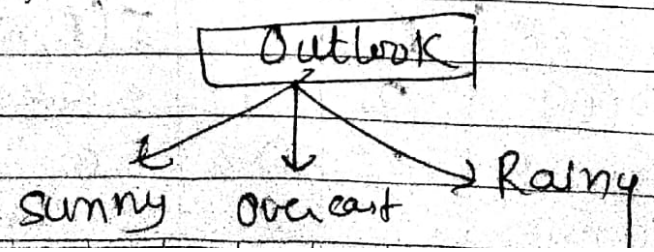
Ily

iii) $\text{Gain}(S, \text{Humidity}) = 0.151$

iv) $\text{Gain}(S, \text{wind}) = 0.048$

Wind	p_i	n_i	$I(p_i, n_i)$
strong	3	3	1
weak	6	2	0.811

Outlook has the highest Gain, Hence used as root node.



Scanned by CamScanner

Step 2 :-

As attribute outlook at root, we have to decide on the remaining ~~three~~ attributes for Sunny branch node.

Consider Outlook = Sunny.

Day	Outlook	Temp	Hum	Wind	Play ball
D1	Sunny	Hot	High	weak	No
D2	Sunny	Hot	High	Strong	No
D8	Sunny	mild	high	weak	No
D9	Sunny	Cool	non	weak	Yes
D11	Sunny	Mild	nom.	Strong	Yes

Total no. of record = 5

$p = 2$

$n = 3$

$$I(p,n) = I(2,3) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2 3/5$$

$$\boxed{I(2,3) = 0.971}$$

i) Compute entropy for Temp :- (Hot, Mild, Cool)

Temp	P_i	n_i	$I(P_i, n_i)$
Hot	0	2	0
Mild	1	1	1
Cool	1	0	0

$$E(\text{Temp}) = \frac{2}{5} I(0,2) + \frac{2}{5} I(1,1) + \frac{1}{5} I(1,0)$$

$$\boxed{E(\text{Temp}) = 0.4}$$

Scanned by CamScanner

$= I(p,n) - E(\text{Temp})$

$\text{Gain}(\text{sunny}, \text{temp}) = 0.971 - 0.4$
$= 0.571$

ii) Compute Entropy for Humidity (High, Normal)

Humidity	p_i	n_i	$I(p_i, n_i)$
High	0	3	0
normal	2	0	0

$E(\text{Humidity}) = \frac{3}{5} \times I(0,3) + \frac{2}{5} I(2,0)$
$= 0$

$\text{Gain}(\text{sunny}, \text{Humidity}) = I(p,n) - E(\text{Hu})$
$= 0.971 - 0$
$= 0.971$

iii) Compute Entropy for Wind, weak, strong.

Wind	p_i	n_i	$I(p_i, n_i)$
weak	1	2	0.918
strong	1	1	1

for weak
$p = 1 \quad n = 2$
$I(p_i, n_i) = I(1,2) = -\frac{1}{3} \log_2 \left(\frac{1}{3}\right) - \frac{2}{3} \log_2 \frac{2}{3}$
$= 0.918$

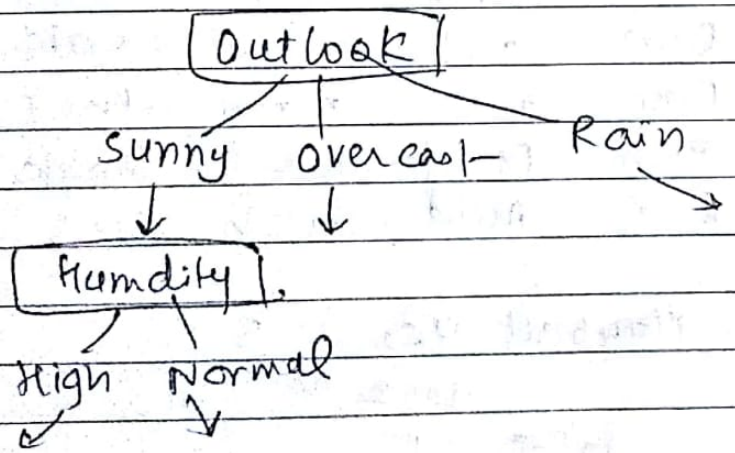
Scanned by CamScanner

$$E(\text{wind}) = \frac{3}{5} I(1,2) + \frac{2}{5} I(1,1)$$
$$= 0.951$$

$$\text{Gain}(\text{sunny}, \text{wind}) = I(p,n) - E(\text{wind})$$
$$= 0.971 - 0.951$$
$$= 0.02$$

Gain(sunny, Temp) = 0.57
Gain(sunny, Hum) = 0.97
Gain(sunny, wind) = 0.02

∴ Humidity having Highest Gain.
∴ below outlook = sunny.

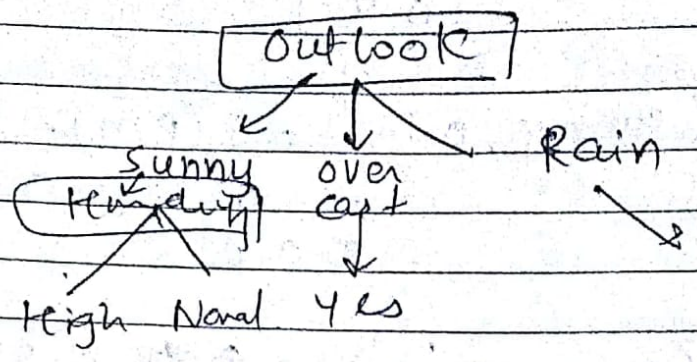


Step3 :-
Consider now only temp & wind for outlook = overcast.

Day	Outlook	Temp	Hum	Wind	Play golf
D3	Overcast	Hot	High	weak	Yes
D7	overcast	Cool	normal	strong	Yes
D12	overcast	mild	High	strong	Yes
D13	overcast	Hot	normal	weak	Yes

Scanned by CamScanner

Since for the attributes & wind play ball = so assign class 'yes' to Overcast



Step 4 :-

Consider Temp & Wind for Outlook = Rain

Day	Outlook	Temp	Hum	Wind	Playball
D4	Rain	Mild	High	weak	Yes
D5	Rain	cool	normal	weak	Yes
D6	Rain	cool	normal	strong	No
D10	Rain	Mild	Normal	weak	Yes
D14	Rain	Mild	High	strong	Yes

P : Playball Yes = 3
no = 2
total = 5

$$I(p,n) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}$$

$$I(3,2) = 0.970$$

1) Compute Entropy for Wind.

Wind	p_i	n_i	$I(p_i, n_i)$
weak	3	0	0
strong	0	2	0

Scanned by CamScanner

$E(\text{wind}) = \frac{3}{5} I(3,0) + \frac{2}{5} I(0,2) = 0$

$\text{Gain}(\text{rain}, \text{wind}) = I(p,n) - E(\text{wind})$
$= 0.97 - 0$
$= 0.97$

ii) Compute Entropy for Temp (hot, Mild, cool)

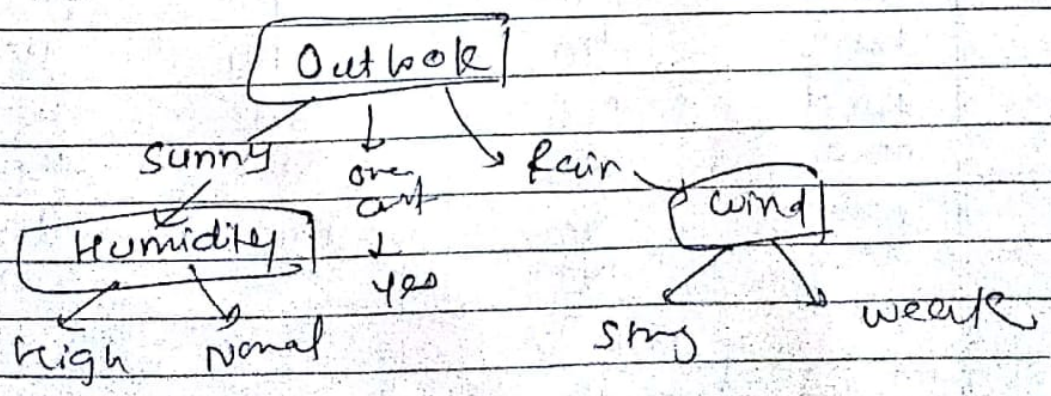
Temp	p_i	n_i	$I(p_i, n_i)$
Hot	0	0	0
mild	2	1	0.918
cool	1	1	1

$E(\text{Temp}) = \frac{0}{5} * I(0,0) + \frac{3}{5} I(2,1) + \frac{2}{5} I(1,1)$
$= 0.951$

$\text{Gain}(S_{rain}, \text{Temp}) = 0.97 - 0.951$
$= 0.019$

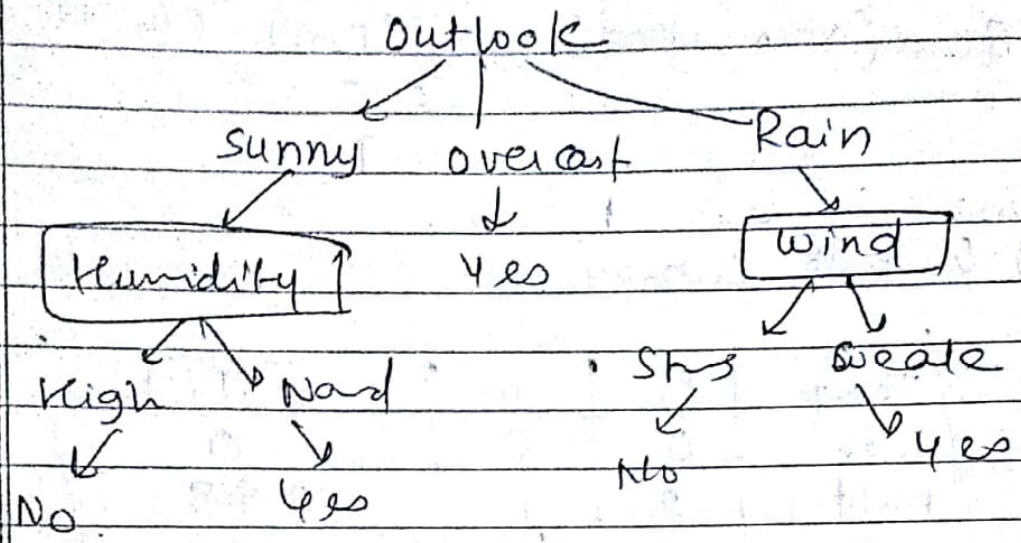
$G(\text{rain}, \text{Temp}) = 0.019$
$G(\text{rain}, \text{wind}) = 0.971.$

Highest Gain



Scanned by CamScanner

final DT.



Example 2 :-

Stock Market involving only discrete ranges has profit as Categorical value (up, down)

Age	Competition	Type	Profit
old	yes	SW	Down
old	No	SW	Down
old	No	HW	Down
mid	yes	SW	Down
mid	yes	HW	Down
mid	No	HW	up
mid	No	SW	up
new	yes	SW	up
new	No	HW	up
new	No	SW	up

$n = 10$

Scanned by CamScanner

	P_1	P_2	Total
	up	down	10
	5	5	

$$Gini(T) = 1 - \Sigma p_j^2$$

$$= 1 - \left(\frac{5}{10}^2 + \frac{5}{10}^2\right)$$

$$Gini(T) = 0.5$$

Step 1:- Split for all attr age, comp & Type.

Age :-

$$Split_{(age)} = \frac{N_1}{N} gini(T_1) + \frac{N_2}{N} gini(T_2)$$

	p	n
old	0	3
mid	2	2
new	3	0

$$= \frac{3}{10} gini(old) + \frac{4}{10} gini(mid) + \frac{3}{10} gini(new)$$

$$= \frac{3}{10}\left[1 - \left(\frac{0}{3}^2 + \frac{3}{3}^2\right)\right] + \frac{4}{10}\left[1 - \left(\frac{2}{4}^2 + \frac{2}{4}^2\right)\right] + \frac{3}{10}\left[1 - \left(\frac{3}{3}^2 + \frac{0}{3}^2\right)\right]$$

$$Split_{(age)} = 0.2$$

$$Split(comp) = \frac{4}{10}\left[1 - \left(\frac{1}{4}^2 + \frac{3}{4}^2\right)\right] + \frac{6}{10}\left[1 - \left(\frac{4}{6}^2 + \frac{2}{6}^2\right)\right]$$

	p	n
yes	1	3
no	4	2

$$= 0.42$$

$$Split(comp) = 0.42$$

$$Split(type) = \frac{6}{10}\left[1 - \left(\frac{3}{6}^2 + \frac{3}{6}^2\right)\right] + \frac{4}{10}\left[\left(1 - \frac{2}{4}^2 + \frac{2}{4}^2\right)\right]$$

$$= 0.86$$

type

	p	n
so	3	3
ho	2	2

Smallest is split(age)

Scanned by CamScanner

Age

old mid new

Step 2:-

for Age = old compute for Comp, Type

Age	Comp	Type	profit
old	Yes	SW	Down
old	No	SW	Down
old	No	HW	Down

∴ Down for all

Age

old mid new

down

Step 3:-

Age = mid

$$Split(comp) = \frac{2}{4}\left[1-\left(\frac{0^2}{2}+\frac{2^2}{2}\right)\right]+\frac{2}{4}\left[1-\left(\frac{2^2}{2}+\frac{0^2}{2}\right)\right]$$

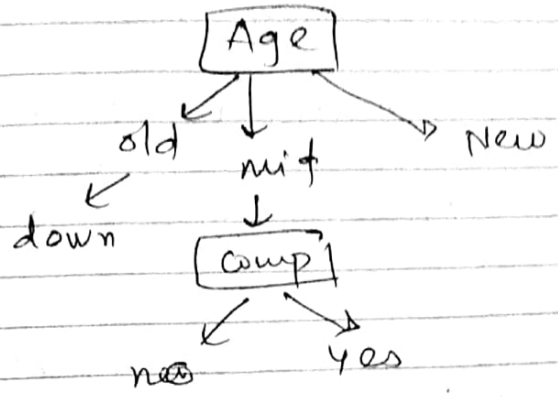
$$= 0$$

$$Split(type) = \frac{2}{4}\left[\left(1-\frac{1^2}{2}+\frac{1^2}{2}\right)\right]+\frac{2}{4}\left[\left(1-\left(\frac{1^2}{2}+\frac{1^2}{2}\right)\right)\right]$$

$$= 0.5$$

Scanned by CamScanner

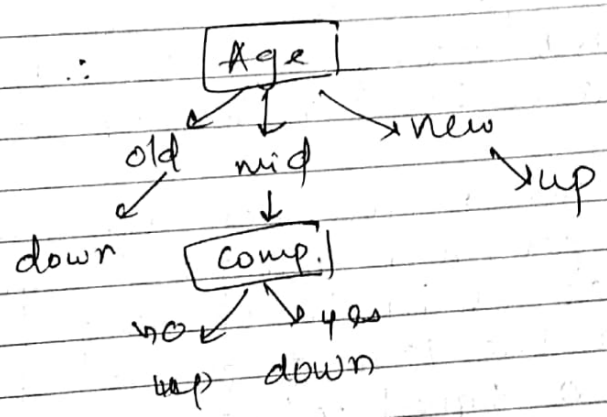
Smallest is Comp.



Step 4 :- Age = new.

all up.

∴



Scanned by CamScanner

Classification	Regression
1. O/P - class	1. O/P - real no
2. Used: Respo- -catego-	2. used - RV - numeric or continuous
3. ~~Spl~~ class	3. Prediction
4. Split	4. Split

Classification & Regression Tree :-

Classification tree → predicted O/P is class to which data belongs

Regression tree → predicted O/P is a real no.

Classification trees :-

- CT are used to separate the dataset into classes belonging to response variable. usually the response variable has two classes : Yes or No (1/0)
- CT are used when response or target variable is categorical in nature.

In a standard classification, the idea is to split the dataset based on homogeneity of data.

for eg: we have 2 variables, age & wt that predict if a person is going to sign up for a gym or not. In our training data if it showed the 90% of people who are older than 40 signed up, we split the data here & age becomes a top node in tree.

- Entropy or GiniIndex is used in Classification Tree.

Regression Tree :-

- Regression tree are needed when the response variable is numeric or continuous. Eg: predicted price of a consumer good.
- The Regression Tree are applicable for prediction type of problem as opposed to classification.

Scanned by CamScanner

- In a regression tree the idea is, since the target variable does not have classes, we fit a regression model to the target variable using each of the Independent variables.
- Then for each independent variable, the data is split at several split points.
- At each split point 'error' between predicted value & actual value is squared to get a sum of squared Error. The split point errors across variables are compared & the variable / point yeilding the lowest SSE is chosen as root node. This process is recursively continued.

Eg:-

Outlook	temp	Humidity	Windy	Hours Played
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
overcast	Hot	High	False	46
sunny	mild	High	F	45
sunny	cool	normal	F	52
sunny	cool	normal	T	23
overcast	cool	normal	T	43
Rainy	mild	High	F	35
Rainy	cool	normal	F	38
sunny	mild	normal	F	46
Rainy	mild	normal	T	48
overcast	mild	High	T	52
overcast	Hot	normal	F	44
sunny	mild	High	T	30

Scanned by CamScanner

- Standard deviation → A decision tree is built up top-down from root node & involves partitioning the data into subsets that contain instances with similar value. We use SD to calculate homogeneity of numerical sample.

$$SD \rightarrow S = \sqrt{\frac{\Sigma(x-\mu)^2}{n}}$$

- SD Reduction → It is based on the decrease in SD after a dataset is split on an attribute. Constructing a tree is all about finding att that returns highest SDR.

$$SDR = S(D) - S(A)$$

~~Calc~~ Soln :-

Step 1 :-

SD for Hours played.

$$SD_{(\text{Hours Play})} = \sqrt{\frac{\Sigma(x-\mu)^2}{n}}$$

$$\mu = \frac{557}{14} = 39.78$$

$$\Sigma(x-\mu)^2 = 218.44 + 95.65 + 38.69 + 27.24 + 149.33 + 281.57 + 10.37 + 22.84 + 3.17 + 38.69 + 67.57 + 149.33 + 17.81 + 95.65$$

$$= 1216.35$$

$$SD_{(\text{Hours Play})} = \sqrt{\frac{1216.35}{14}} = \sqrt{86.88} = 9.32$$

Scanned by CamScanner

Outlook) Rainy → 25, 30, 35, 38, 48 (5) 35.2
Sunny → 45, 52, 23, 46, 30 (5) 41.5
overcast → 46, 43, 52, 44 (4) – 46.5

Temp ⇒) ~~Hot → 25, 30, 46, 44 (4) – 36.25~~
Mild → 45, 35, 46, 48, 52, 30 (6) – 42.66
Cool → 52, 23, 43, 38 – 39

Step 2 :- The dataset is then split on the different attribute.
SD for each branch is calculated.
The resulting SD is subtracted from SD before Split = SDR.

i) SDR for outlook

		Hrs played (SD)
outlook	overcast	3.49
	Rainy	7.78
	sunny	10.87

$$S(\text{overcast}) = \sqrt{\frac{(46-46.25)^2 + (43-46.25)^2 + (52-46.25)^2 + (44-46.25)^2}{4}}$$

μ(overcast) = 46.25

$= 3.49$

$$S(\text{Hrs, Outlook}) = P(\text{sunny}) \times S(\text{sunny}) + P(\text{overcast}) \times S(\text{overcast}) + P(\text{Rainy}) \times S(\text{Rainy})$$

$$= \frac{5}{14} \times 10.87 + \frac{4}{14} \times 3.49 + \frac{5}{14} \times 7.78$$

$= 7.66$

$$SDR(\text{outlook}) = S(\text{Hrs}) - S(\text{Outlook})$$
$$= 9.32 - 7.66 =$$
$$= 1.66$$

ii) SDR for Temp

Temp		SD(Hrs Play)
	Hot	18.95
	cool	10.51
	Mild	7.65

$$S(\text{Hrs, Temp}) = \frac{4}{14} \times 10.51 + \frac{6}{14} \times 7.65 + \frac{4}{14} \times 18.95 = 8.83$$

SDR = 0.49

High → 25, 30, 46, 45, 35, 52, 30 → 37.5 (7)
normal → 52, 23, 43, 38, 46, 48, 44 → 42 (7)

PAGE No. DATE

Hrs

Humidity	High	9.36
	normal	8.37

SDR = 0.28

Windy
True → 25, 46, 45, 52, 30, 23, 43, 38 (37.5) (41.3)
false → 25, 46, 45, 52, 35, 38, 46, 44 (41.3)

$S(\text{Humidity}) = \frac{7}{14} \times 9.36 + \frac{7}{14} \times 8.37 =$

Windy	False	7.87
	True	10.59

SDR = 0.29

Outlook → highest SDR → root node.

Outlook
- Sunny
- Overcast
- Rainy

Step 2:-
Outlook = sunny.

temp	Humidity	windy	Hrs
mild	High	F	45
cool	normal	F	52
mild	normal	F	46
cool	normal	T	23
mild	High	T	30

SD = 10.87

Windy	F	3.09
	T	3.50

SDR = 7.62

$10.87 - \left(\frac{3}{5} \times 3.09 + \frac{2}{5} \times 3.0\right)$

= 7.62

Temp			Humidity	
	mid	7.31	high	
	cool	14.5	normal	

PAGE No. 75
DATE 12.4.19

65

SDR = 0.084

Outlook

SDR = 0.38

sunny overcast Rainy

⇒ highest SDR for windy.

windy

false true

47.7 26.5

Step3 :-

for overcast : 46, 43, 52, 44 → % diff

Outlook

sunny overcast Rainy

46.3

windy

false true

47.7 26.5

Step4 :- for Rainy we consider Humidity & temp

$\mu = 35.2$

SD = 7.78

		SD
Humidity	High	5
	normal	5

$\Rightarrow \left(7.78 - \frac{3}{5} \times 5 + \frac{2}{5} \times 5\right)$

SDR = 2.28

$= 2.28$

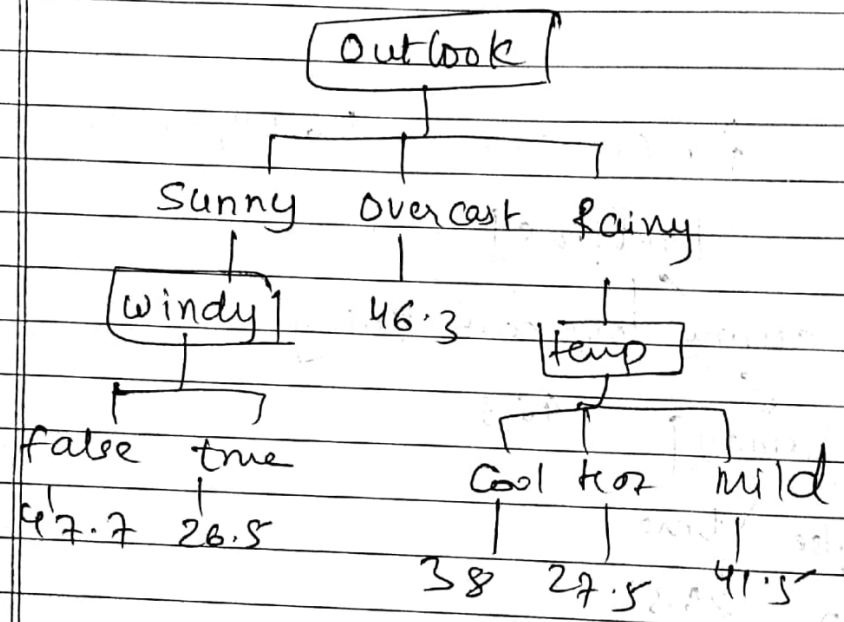
Scanned by CamScanner

Temp		
	Hot	2.5
	mild	6.5
	cool	0
SDR = 4.18		

$\Rightarrow 7.78 - \left(\frac{2}{5} \times 2.5 + \frac{2}{5} \times 6.5 + \frac{2}{5} \times 0\right)$

Highest $\Rightarrow$ Temp

So temp is drawn below Rainy



Scanned by CamScanner