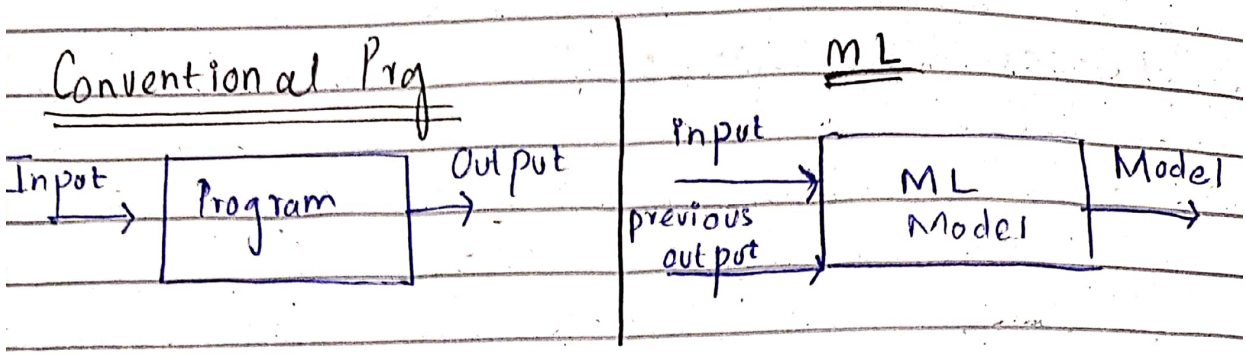
to transfer Learning capability to machine without explicit program.

- UnSeen data will be process to classify
- hard coded ko replace kr dia.



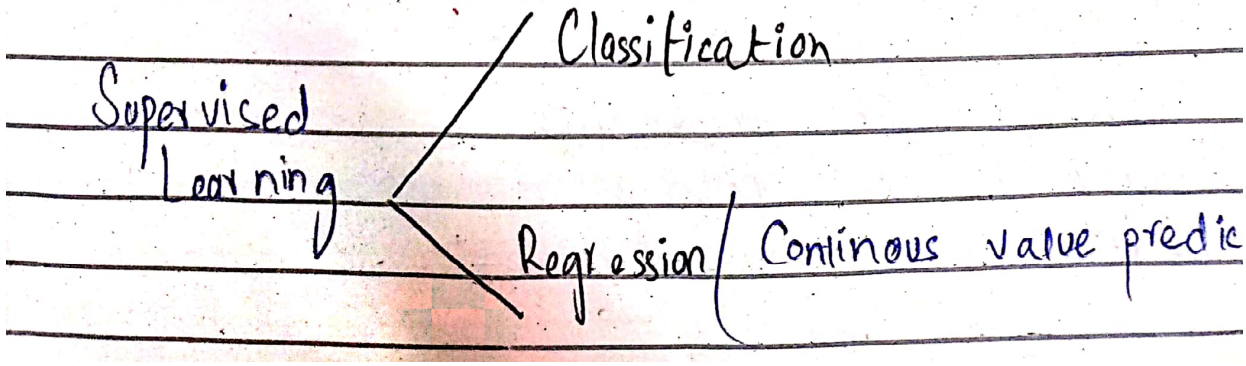
Categories of ML

1) Supervised Learning
2) UnSupervised Learning
3) Reinforcement Learning

1. SuperVised learning:

class Labels are available, you will train your model on the bases of previous data. and predict the output of future data.

- Training data has class Label
- We want to predict that Label for future instance.



Regression ex
Krta ha like predict the price of a house, car

UnSupervised Learning

class Label are not available
So we group up things that are similar

Class Label ni hoty but feature ki base py group krty like fruit ko color, shape ki feature py check krty.

jitny zyda feature ~~hoin~~ utna zyda bhtr grouping.

Reinforcement :

Supervised py train kia or ~~test~~ test kia galt answere aya humny us ko bataya galat ha us ny apny knowledge base mn us ko add kr lia, Learn kr lia us data ko

- Learning with expercience
- Learning with passage of time

Optimization : maximum benefit

Applications:

- To check wether loan should be awarded or not on the bases of credit card score. — Supervised
- Real Time Stock Value Prediction — Supervised
- Spam OR NOT SPAM / FakeNews or Real — Unsup
- Computer vision to identify a fertile Land and untertile Land — Unsup
- Like Self Driving Car — Reinforcement
-

Decision Tree

Inductive
- Inductive means recursive
- derive classes from (x,c)
- ID3 is greedy (best ki taraf jata ha) but guarantee ni ha

Concept Learning System
jiska name rakh skaien (e.g Game, Animal etc) agy us k sub concept
Algo

1)

Decision Tree is attribute-based description
- phly ek universal object ly lo (Animal etc)
- then sub-concept is defined

Resturant Problem

Should we wait for a table or not

Input Attribute
Alternate resturant, etc

Output Will Wait

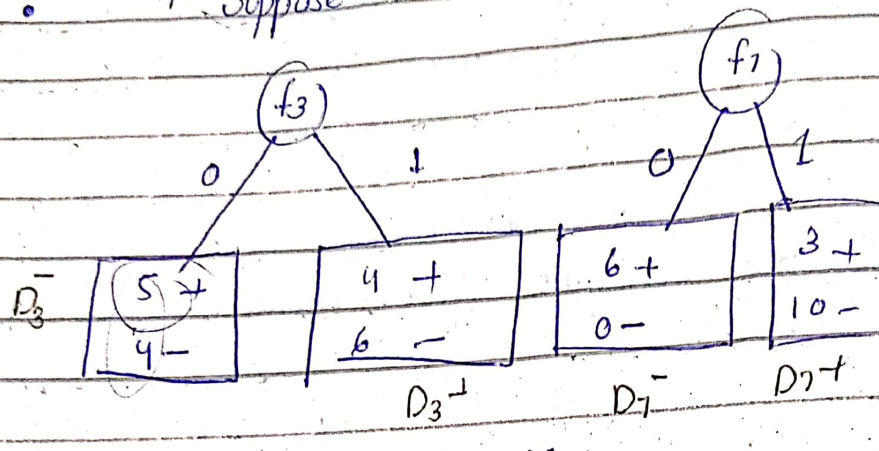
How to Construct Decision Tree

Task

Measure of Node Impurity

Entropy

- Total 19 instances
- 9 posi & 10 neg
- Suppose it has 7 features



p := proportion of +ive

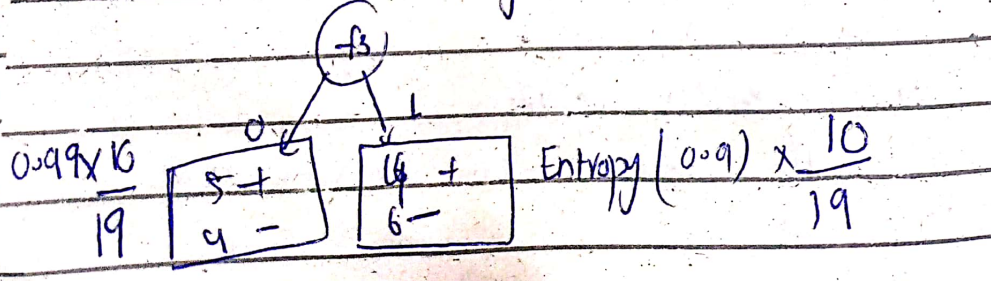
$$H = -p\log_2 p(1-p)\log_2(1-p)$$

$\log_2 0 = 0$ $\log_2 1$ = undefined

$$D_3^{+-} = \frac{-5}{9}\log_2\frac{5}{9} - \left(\frac{4}{9}\right)\log_2\left(\frac{4}{9}\right) = 0.6551 - 0.37$$

$$D_3^{+} = \frac{-4}{10}\log_2\frac{4}{10} - \left(\frac{6}{10}\right)\log_2\left(\frac{6}{10}\right)$$

Calculate Average Entropy



$\frac{9}{19} \times 0.99 + \frac{10}{19} \times 0.97$

$= AE$

jiski AE woh better split hoga

Class Entropy.
info Gain ka max

⇒ Stopping ?

1) Stop if entropy is below some threshold.
2) Number of element in data set is below threshold
3) Agly split is entropy decrease nh hor

overfit kisi ko classify he nh kr sky ga

Example

Step 1 : class level ka H (Entropy)

Total = 6

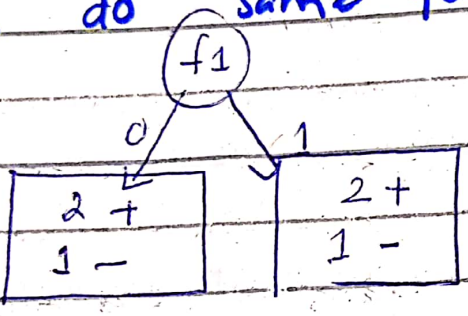
+3 , -3

Class Level Entropy:

$$H = -P\log_2(P) - (1-P)\log_2(1-P)$$

$$= -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \left(\frac{3}{6}\right)\log_2\left(\frac{3}{6}\right)$$

f1 do same for all



Calculate AE

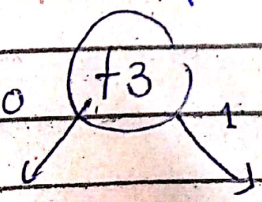
$AE_1 = 0.92$

$AE_2 = 0.92$

$AE_3 = 0.8$

$AE_4 = 1$

Choose f3 & split

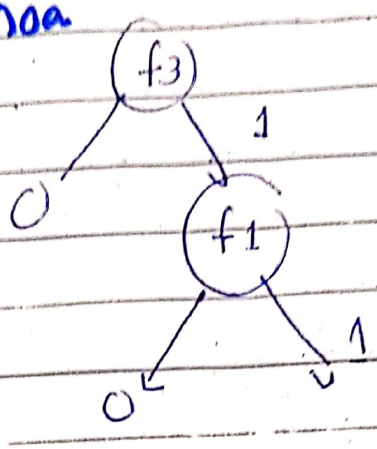


0 K instance 1 K sary instance

ap f3 ko chor dia f1, f2, f4 ko call gaya
kare

AE1 = 0.58
AE2 = 0.55
AE3 = 0.95
} koi b ek select kr lo

F1 Choose hoa



STopping Criteria ajaye toh stop kr do
ni toh baki ko f2, f4
split krty jao

Pruning

Pre Pruning jaisy stopping criteria
ka zikr kia.

Post Pruning para kr k phir prune
kr lo

Leaf Node ka content o

What is over fitting

performing good on some training data but

It can't be generalized so won't work in real envo.

- Pruning is the technique to avoid over fitting

ID3

Pruning :

Statistical method is used to prune the least reliable

Two common stra

- Pre Prune (by halting)
- Statistical Significance (between attribute significant

Sometime practitioner combine both tech pre + post (hybrid) approach eg lagai hoi pre ha but agr error ajaye toh post prune kar lo simple.

Post Pruning is divided into

1) Training sample
2) 10 → 20% for valid (Training Test)
3) 5 → 10% for testing

Regression tree

- target column could be a number
- Use stand deviation

X = 25, 30, 46, 45, 52, 23, 43, 35, 38, 46, 48, 52, 44, 30

Average = 39.78

Standard deviation

$$\sqrt{\frac{(25-39.78)^2 + (30-39.78)^2 + \dots}{N}}$$

Ex 2 25, 30, 35, 38, 48

$$\text{Avg} = \frac{25+30+35+38+48}{5}$$

$$= 35.2$$

$$\text{Stand} = \sqrt{\frac{(25-35.2)^2 + (30-35.2)^2 + (35-35.2)^2 + (38-35.2)^2 + (48-35.2)^2}{5}}$$

$$= \sqrt{\frac{(10.2)^2 + (-5.2)^2 + (0.2)^2 + (2.8)^2 + (12.8)^2}{5}}$$

$$\sqrt{\frac{104.04 + 27.04 + 0.04 + 7.84 + 163.84}{5}}$$

$$\sqrt{\frac{302.8}{5}}$$

$$= \sqrt{60.56}$$

$$= 7.78$$

Sunny

Overcast = 3.49

Rainy = 10.87

Summarized

	Std	Instan
Overcast	3.49	4
Rain	10.87	5
Sunny	7.78	5
		14

Weighted Standard

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

$$= 7.68$$

Naive Bayes

Spam Classification Example

First we built a histogram

$P(\text{Dear} \mid \text{Normal}) = \frac{8}{17}$

8 times dear occured in normal message.

Class Prior Probability

$P(S) \times P(\text{Dear} \mid S) \times P(\text{Friend} \mid N)$

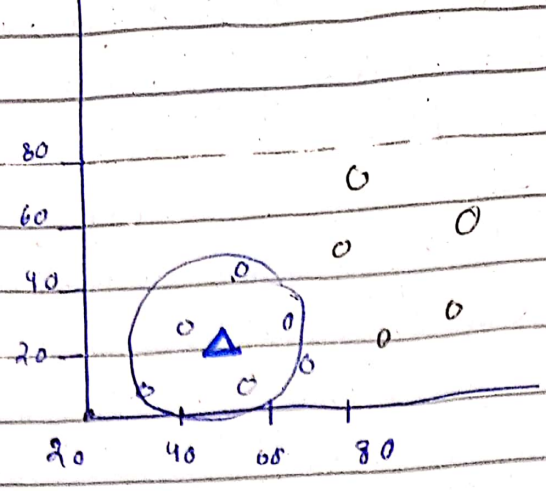
KNN

- Supervised Learning
- non-parametric classification
- lazy learning algo :
  phly sy train ni krta,
  jab query dy jaye tab train krta.
- active learning
  already prepared

KNN is Lazy

~~Example~~ :

- Sensitive to imbalanced class



near by mn 5 viral hain
toh viral classify

Step 1 Measure the distance for the new data point thats why it is considered Lazy.

Step 2

Step 3 Sort distance

Step 4 identify closet depend on k determine majority class

k shows the number of closl point to select.

Leave One Out Cross Validation

How good is the classifier ?

- 1 method is Leave One out Cross validation
  - One by one check algo ki row ko algo kya predict krta phir us ki acutual or prediction match kr lo.
- Each tuppe ps check across all training da

Accuracy = Correct Instance = 0.833
Total instance

$ Traing test

Problem of leave One Out
when we have 1000+ datapoint
than calculating distance
is time consuming.
Solution
Split the data and into 2
parts
- Large training data
- Small testing data

80, 90 training
20, 10 % testing

Problem with imbalanced Dataset

agr strength bhar jaye toh
probablity bhar jaye ge

How to find optimal value of k
- Try k even na ho
- Should not be too high & low
- k should be odd and >1

$$k = \sqrt{n}$$

More than 2 groups standarizat

- jiska gap kam ho uska impact o ho jaa
- highest gap will influence more

perform standarization (assign weight)

$$Z_1 = \frac{X_i - \bar{X}}{SD}$$

equal impact & result will be less biased.

Advantage

simple & easy

Disadvantage

- all training data is used
- sensitive to imbalanced data.