Unit - 1

Introduction to machine learning

Enable the machine to automatically learn from data improve performance from experience and predict things without being explicitly program.

Features of machine learning

- Machine can learn itself from past data & automatically improve.
- Machine learning is data driven technology. Large amount of data generated by organizations on daily basis. So by notable relationship in data organization makes better decisions.
- From the given dataset it detects various patterns on data.
- It is similar to data mining because it is also deals with the huge amount of data.

Machine Learning life Cycle involve 7 major steps

1. Gathering Data
2. Data Preparation
3. Data Wrangling
4. Analyze data
5. Train the model
6. Test the model
7. Deployment

(Q.) Intense based learning or Memory learning or Lazy learning Model based learning.

Sol^n **Bias and Variance**

Bias is simply defined as the inability of model because of that there is some difference or error occurring between the model predicate values.

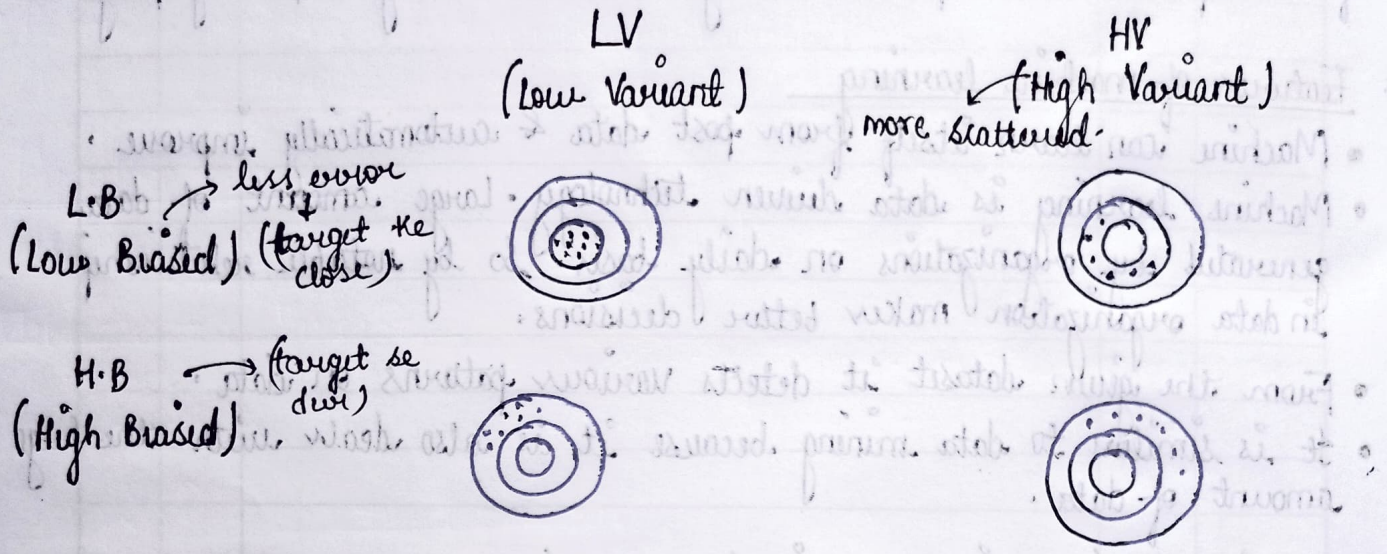
These difference between actual or expected values & predicated values are known as error or biased error or error due to biased.

Low Bias

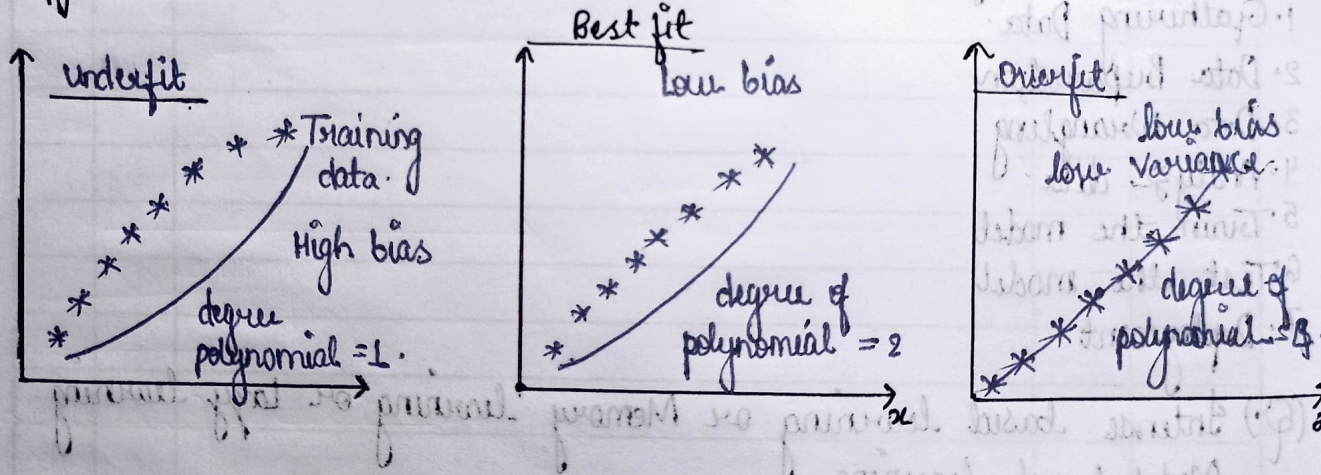
Low bias values means assumption are taken to build the target function In these case the model will closely match the training dataset.

High Bias

High Bias Value means assumption are taken to build the target function. In these cases the model will not closely match the training dataset.

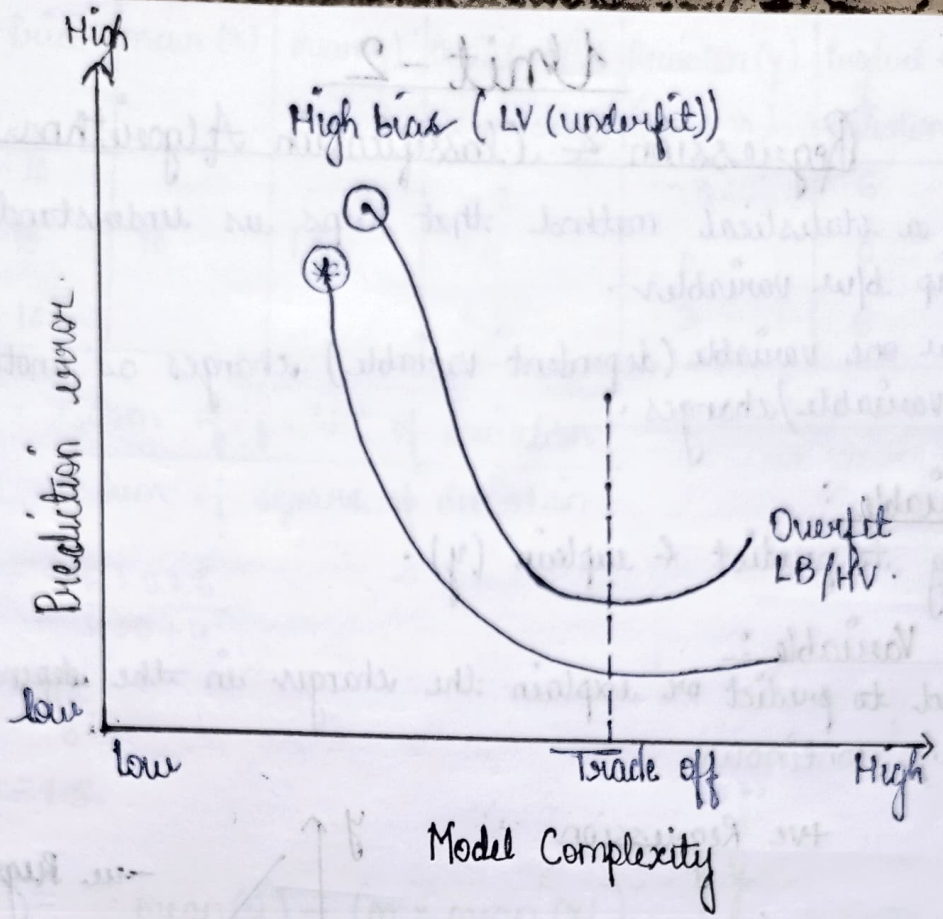


Regression



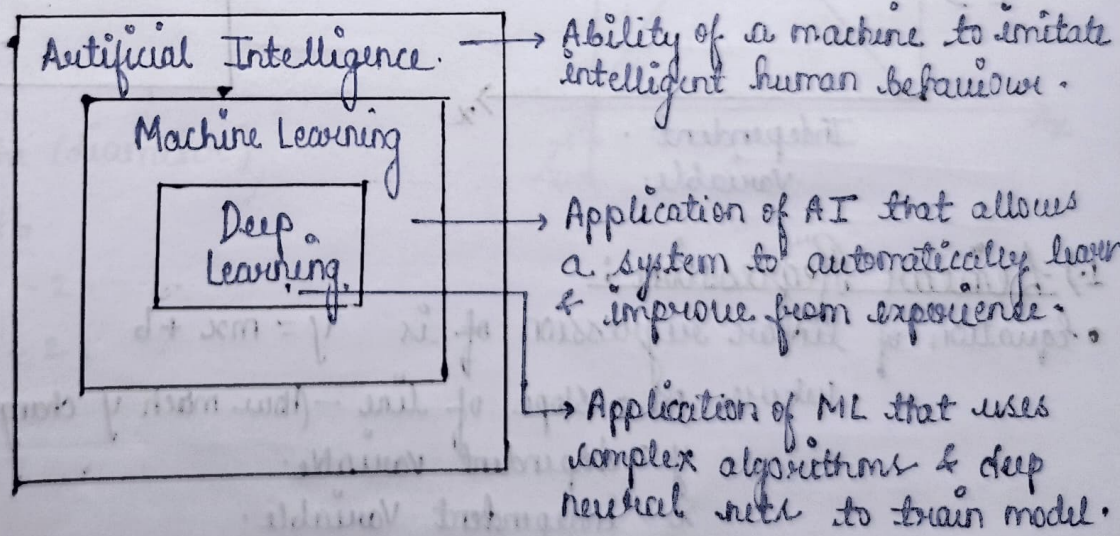
Classification problem Specification

Model 1	Model 2	Model 3
Training error 1%. Testing error 20%.	Training error 25%. Testing error 26%.	Training error = <10%. Testing error = <10%.
Low Bias High Variant.	High Bias High Variant	Low Bias Low Variant.
Overfitting	Underfitting	Optimal fitting.



13/02/2024

Difference between AI / ML / Deep learning.



Unit - 2

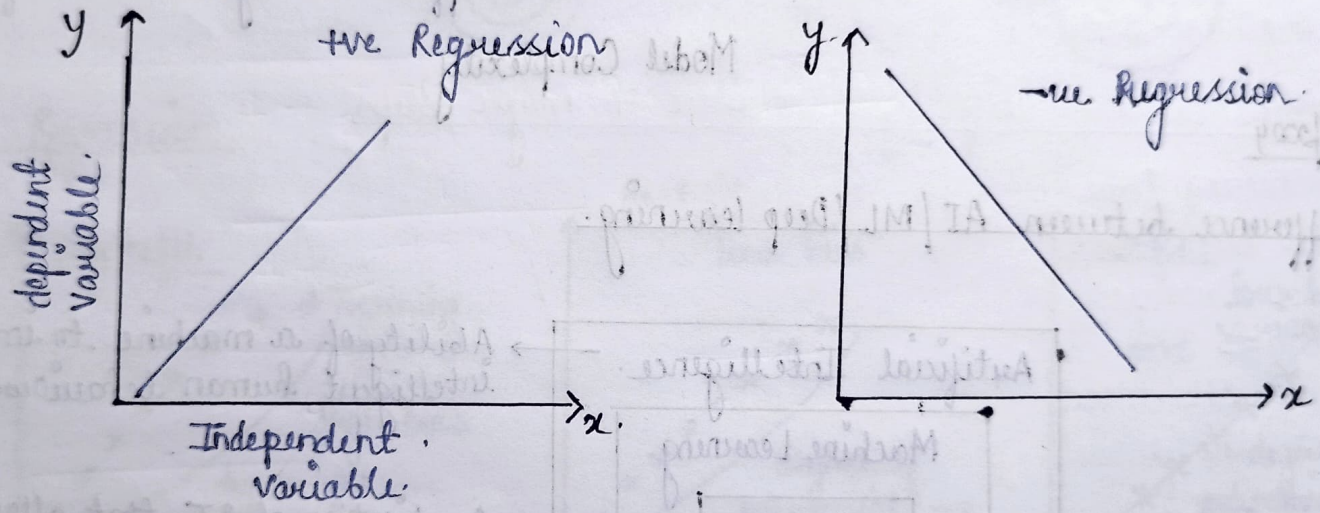
Regression & Classification Algorithm

Regression

- Regression is a statistical method that helps us understand & predict the relationship b/w variables.
- Describes how one variable (dependent variable) changes as another variable (independent variable) changes.

Dependent Variable:
We are trying to predict & explain (y).

Independent Variable:
That are used to predict or explain the changes in the dependent variable (x). → always: continous: shoes!



1.) Linear Regression:

- Equation of linear regression of is $y = mx + b$

where m = slope of line - (how much y changes for unit change in x)
y = dependent variable.
x = independent variable.
b = intercept - (when x = 0 then what is the value of y)

(Q.) Predicting Pizza prices

Diameter (x)	Price (y) → dependent : continous means linear regression.
8	10
10	13
20	16.

Sol

Diameter	Price	mean (x)	mean (y)	Deviation (x) (x − mean(x))	Deviation (y) (y − mean)	Product of deviation	Sum of product of deviation	Square of deviation x
8	10			−2	−3	6		4
10	13	10	13	0	0	0	12	0
12	16			+2	3	6		4

i) Calculate $m = \frac{\text{Sum of product of deviation}}{\text{Sum of square of deviation}}$

$= \frac{6+0+6}{4+0+4}$

$= \frac{12}{8}$

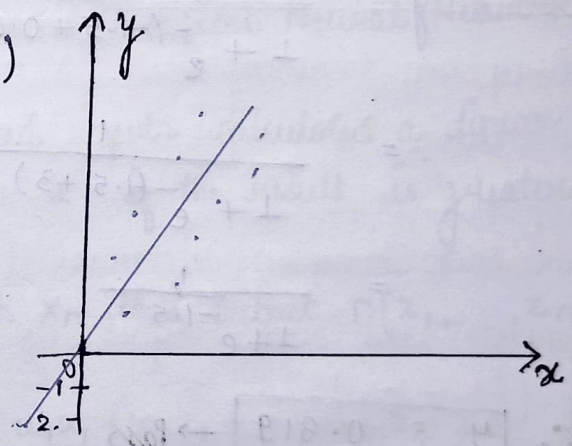
$m = 1.5$

ii) Calculate b = mean(y) − (m x mean (x))

$= 13 - (1.5 \times 10)$

$= 13 - 15$

$= -2$



(ii) let $x = 20$ inche (diameter)

$y = mx + b$.

$y = 1.5 \times 20 - 2$.

$= \frac{15}{10} \times 20 - 2$

$= 30 - 2$

$y = 28$

17/02/2024

2.) Logistic Regression :

(Q.)

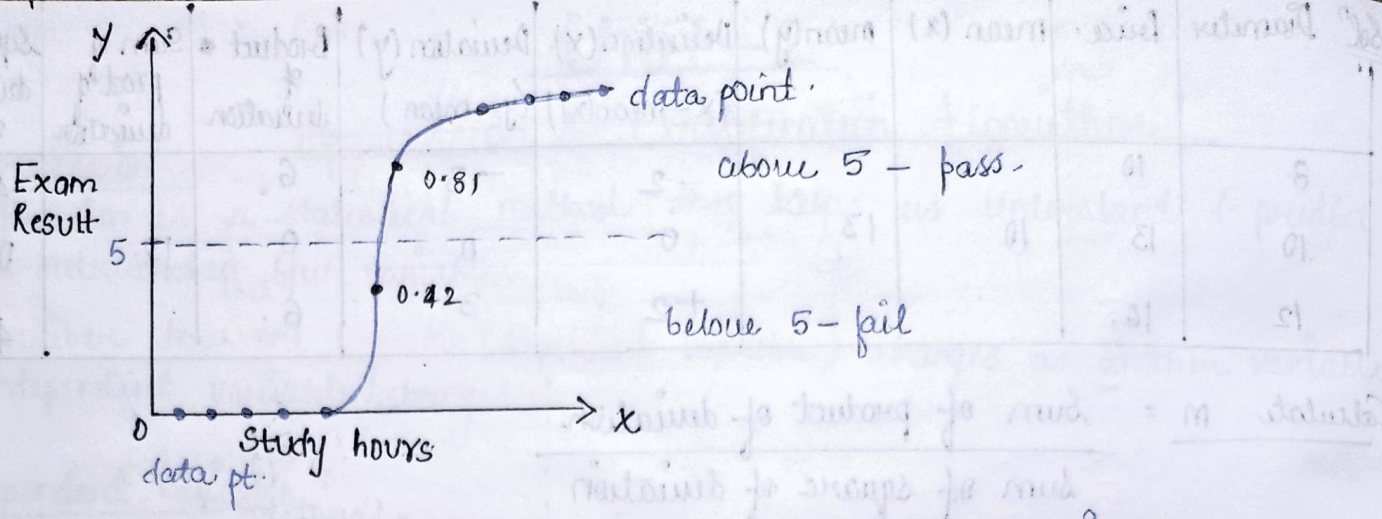
Study Hours	Exam Result
2	0
3	0
4	1
5	1
6	1
7	1
8	1

Given $a_0 = -1.5$

$a_1 = 0.6$

$x =$

dependent variable : categorical means logistic regression



- $$y = \frac{1}{1 + e^{-(a_0 + a_1 x)}}$$

sigmoid function.

where x = independent variable
y = dependent variable.
a_0 = intercept
a_1 = slope

Sol^n

$$y = \frac{1}{1 + e^{-(-1.5 + 0.6 \times 5)}} \quad \text{(i) } \boxed{x = 5}$$

$$= \frac{1}{1 + e^{-(1.5 + 3)}}.$$

$$= \frac{1}{1 + e^{-1.5}} = \frac{1}{1 + 0.22}$$

$\therefore \boxed{y = 0.819} \rightarrow$ Pass

$= 81.9\%$

(ii) $x = 2$.

$$y = \frac{1}{1 + e^{-(-1.5 + 0.6 \times 2)}} = \frac{1}{1 + e^{-(-1.5 + 1.2)}} = \frac{1}{1 + e^{-(-0.3)}}$$

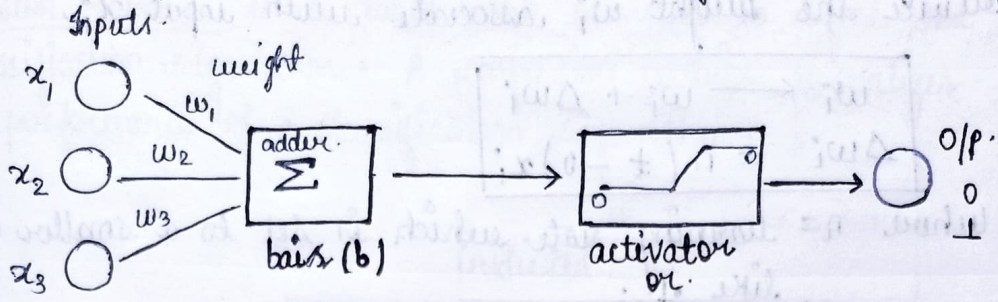
$$= \frac{1}{1 + 1.349} = \frac{1}{2.349}$$

$\therefore \boxed{y = 0.42} \rightarrow$ fail

15/02/2024

Perceptron : In machine learning the perceptron or Mc Culloch – Pitts neuron is an algorithm for supervised learning of binary classifiers.
A binary classifier is a func^n which can decides whether ← not an input

represented by a vector of numbers belongs to some specific class.



$= x_1 w_1 + x_2 w_2 + x_3 w_3 + \cdots + x_i w_i + b$

$$n = \sum_{i=1}^{n} x_i w_i + b$$

or

$$n = X \cdot w + b$$

$$f(x) = \frac{1}{1+e^{-x}}$$

- A perceptron unit is used to build the ANNS (Artificial Neural Network System).
- A perceptron takes a vector of real valued inputs calculates a linear combination of these inputs then outputs a '1' if the result is greater than some threshold & −1 otherwise.
- Most precisely given i/p from x_1 through x_n then output $O(x_1, \ldots, x_n)$ computed by the perceptron is

$$O(x_1, \ldots, x_n) = \begin{cases} 1 & \text{if } w_1 x_1 + w_2 x_2 + \cdots + w_n x_n > 0 \\ -1 & \text{otherwise.} \end{cases}$$

- Where each w is a real valued constant or weight that determines the contribution of x input x to the perceptron output.

Perceptron Training Rule

1.) One way to learn an acceptable weight vector is to begin with random weights then iteratively apply the perceptron to each training example, modifying the perceptron weights whenever it mis classify an example.

2.) This process is repeated iteratively through the training examples as many times as needed until the perceptron classifies all the training examples correctly.

3.) Weights are modified at each step according to the perceptron training rule which revise the weight w_i associate with input x_i.

$$w_i \leftarrow w_i + \Delta w_i$$
$$\Delta w_i = \eta (t - o) x_i$$

where η = learning rate which is set to a smaller value like .1.

t = target output

o = actual o/p

x_i = I/p associated with w_i.

* Using this equation we will modify a weight ~~until~~ in this iteration until all the examples are classified correctly.

Perceptron Algorithm

Perceptron training rule (X, η)

initialize w ($w_i \leftarrow$ an initial (small) random value).

repeat

for each training instance $(x, tx) \in X$: (tx → target)

Compute the real output Ox = Activation (Summation $(w.x)$)

if $(tx \neq Ox)$

for each w_i

$w_i \leftarrow w_i \; \Delta w_i$

$\Delta w_i \leftarrow \eta (tx - Ox) x_i$

end for

end if

end for.

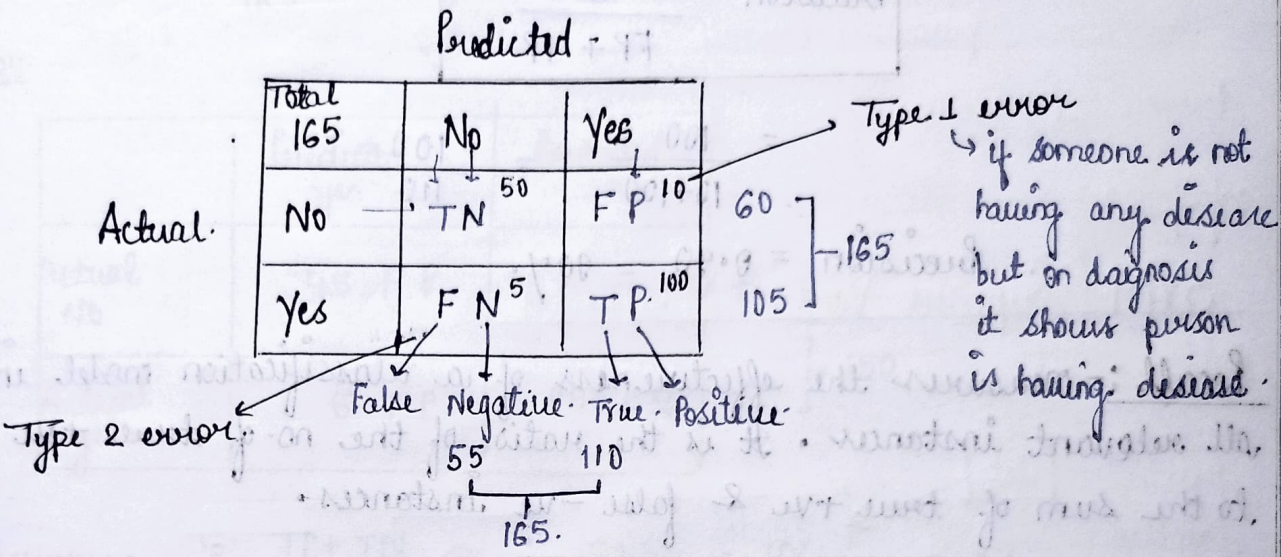
Until all the training instances in X are correctly classified

return w

Once all the training instances are classified correctly final weight are returned as the learned parameter by the perceptron network.

16/02/2024

Confusion Matrix : is a table that is used to define the performance of a classification algorithm. A confusion matrix visualizes & summarize the performance of a classification algorithm.



- The matrix displays the no. of instances produced by the model on the test data.

True +ve (TP) :- Occurs when the model accurately predicts a +ve data-points.

True -ve (TN) :- Occurs when model accurately predicts a -ve datapoint.

False +ve (FP) :- Occurs when model predicts a +ve data point incorrectly.

False -ve (FN) :- Occurs when a model miss predicts a -ve data point.

Accuracy :- used to measure the performance of the model. It is the ratio of total correct instance to the total instances.

$$\text{Accuracy} = \frac{TP + TN}{\text{Total}}$$

$$= \frac{100 + 50}{165}$$

$$\therefore \text{Accuracy} = \frac{150}{165} = 0.91 = 91\%$$

Precision :- is a measure of how accurate a models true predictions are. It is defined as the ratio of true +ve predictions to the total no. of +ve predictions made by the model.

$$\text{Precision} = \frac{TP}{FP + TP}$$

$$= \frac{100}{10+100} = \frac{100}{110}$$

$$\therefore \text{Precision} = 0.90 = 90\%.$$

Recall :- measures the effectiveness of a classification model in identifying all relevant instances. It is the ratio of the no. of true +ve instances to the sum of true +ve & false −ve instances.

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$= \frac{100}{100+5} = \frac{100}{105}$$

$$\therefore \text{Recall} = 0.95 = 95\%.$$

F1 score :- is used to evaluate the overall performance of a classification model. It is the harmonic mean of precision & recall.

$$F_1 \text{ score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$$

$$= \frac{2 \times 0.90 \times 0.95}{0.90 + 0.95}$$

$$= \frac{1.71}{1.85}$$

$$\therefore F_1 \text{ score} = 0.92 = 92\%.$$

Error :-

$$\text{error} = \frac{FP + FN}{\text{Total}}$$ → Type 1.

or → Type 2

$$\text{error} = 1 - \text{accuracy}.$$

or

$$\text{error} = \frac{10+5}{165} = 0.09 \quad \text{or} \quad \text{error} = 1 - 0.91 = 0.09.$$

17/02/2024

(Q.)

	Predicted yes	Predicted No	
Actual No	45 FP	5 TN	50
Actual yes.	5 TP	95 FN	100

→ 150.

Sol: Accuracy $= \frac{TP + TN}{\text{Total}} = \frac{5+5}{150} = \frac{10}{150} = 0.06$

Precision $= \frac{TP}{FP + TP} = \frac{5}{45+5} = \frac{5}{50} = 0.1.$

Recall $= \frac{TP}{TP + FN} = \frac{5}{5+95} = \frac{5}{100} = 0.05.$

F_1 Score $= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0.1 \times 0.05}{0.1 + 0.05} = \frac{0.01}{0.15} = 0.06$

Error $= 1 - \text{accuracy} = 1 - 0.06 = 0.94$

(Q.)

	Predicted ~~yes~~ No	Predicted ~~No~~ yes
Actual ~~yes~~ No	45 TN	5 FP
Actual yes.	5 FN	95 TP

Sol: Accuracy $= \frac{TP + TN}{\text{Total}} = \frac{45+95}{150} = \frac{140}{150} = 0.90 = 90\%.$

Precision = $\frac{TP}{TP+FP} = \frac{95}{95+95} = \frac{0.67 \cdot 95}{100} = 0.95$.

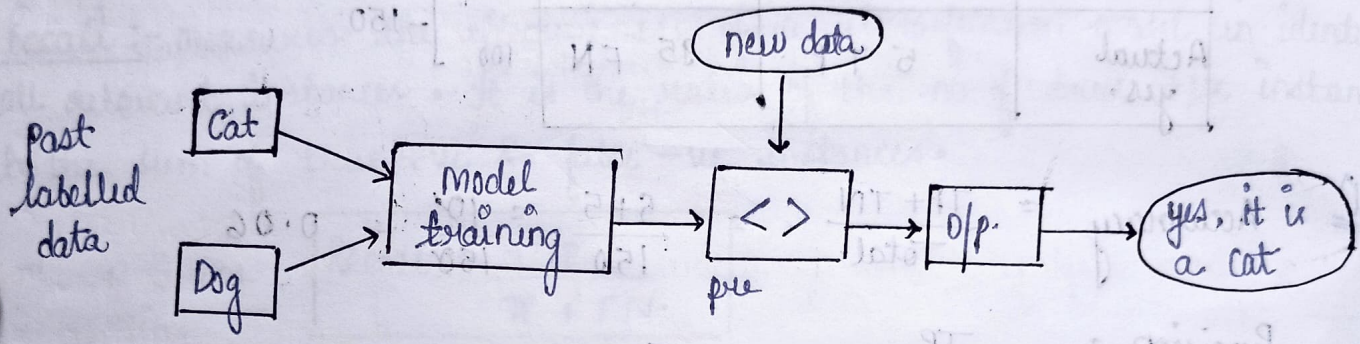
Recall = $\frac{TP}{TP+FN} = \frac{95}{95+5} = \frac{95}{100} = 0.95$.

F_1 score = $\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0.95 \times 0.95}{0.95+0.95} = \frac{1.805}{1.92} = 0.95$.

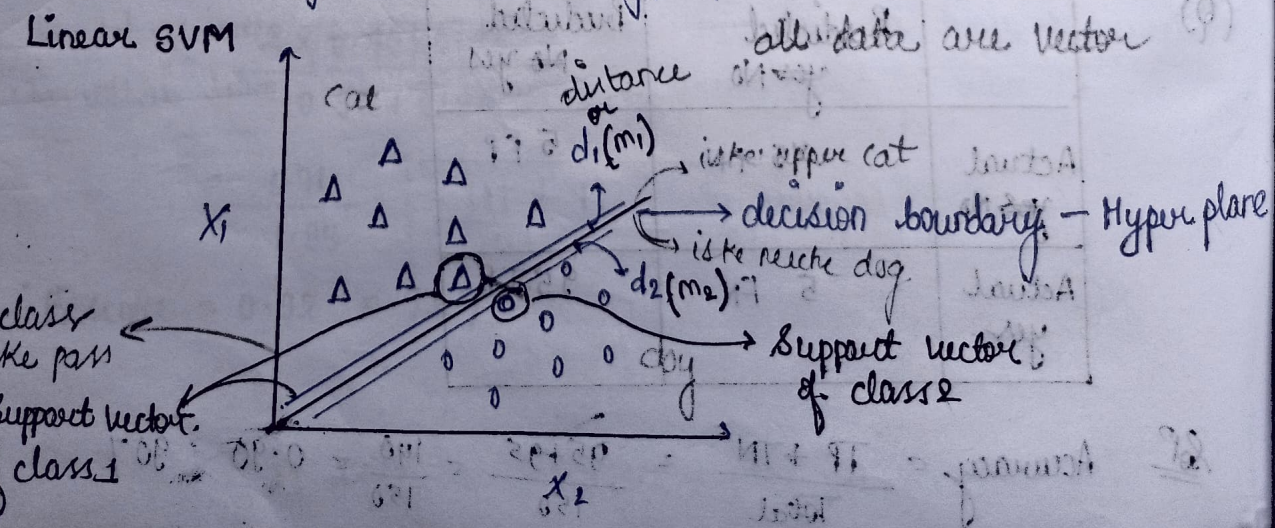
Error = 1 − accuracy = 1 − 0.9 = 0.1 = 10 %.

20/02/2024

SVM (Support Vector Machine)



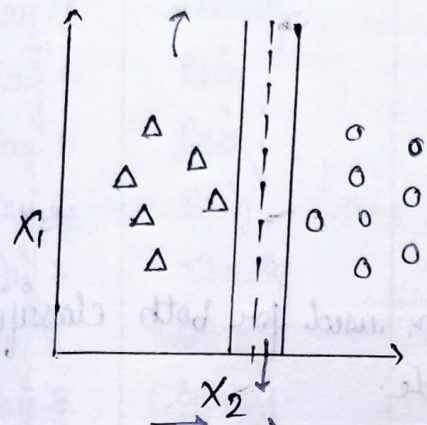
- A support vector m/c is a type of supervised m/c algo. used in m/c learning to solve a classification problem.
- SVM are particularly good at solving binary classification problem which require classifying the elements of a dataset into two groups.
- The goal of the SVM algorithm is to create the best line or decision boundary that can seggregate n-dimension space into classes so that we can easily put the new data point in the correct category in future.
- This best decision boundary is called a hyperplane.



- Compare m_1 or d_1 & m_2 or d_2 which will having a minimum distance or margin from support vector will consider best line (hyperplane) for these data.

- Hyper plane can be shown mathematically as —

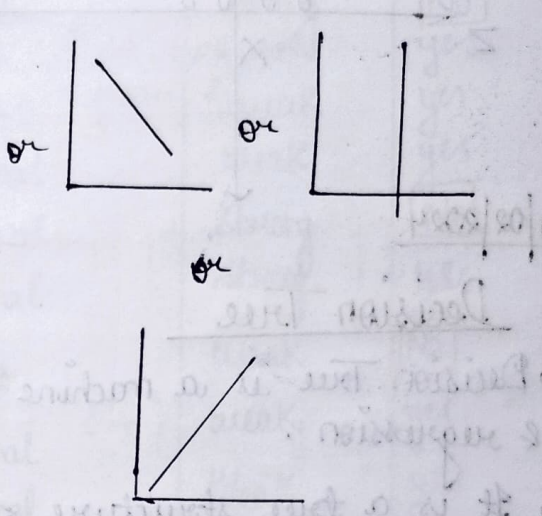
$\vec{w} \cdot \vec{x} + b = -1$ (intercept)



$\vec{w} \cdot \vec{x} + b = 1$.

t. weight.

$\vec{w} \cdot \vec{x} + b = 0$



$$\text{max} = \frac{2}{(\text{max distance})\,|w|}.$$

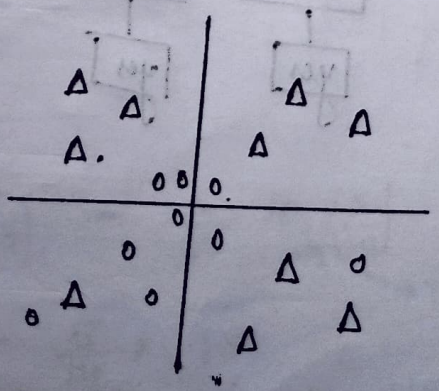
or $(w \cdot x + b) \geq 1 \;\forall\; x$ of class 1

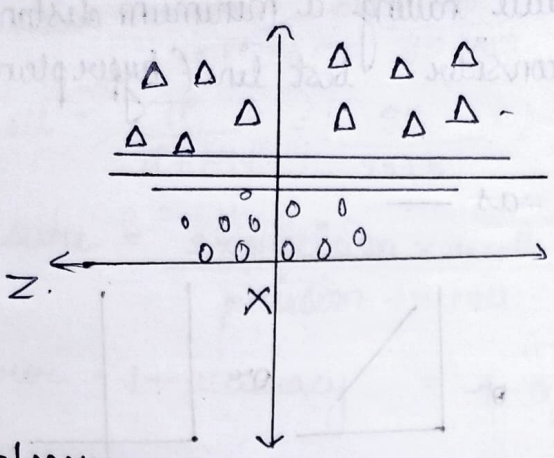
$(w \cdot x + b) \leq 1 \;\forall\; x$ of class 2.

Non Linear SVM

- If data is linearly arrange than we can separate it by using straight line but for non-linear data we can't draw a single straight line.
- So to separate those data points we need to add one more dimension.
- For linear data we have ~~too~~ used two dimension x & y so for non-linear data we will add 3rd dim^n z.
- It can be calculated as

$$z = x^2 + y^2$$



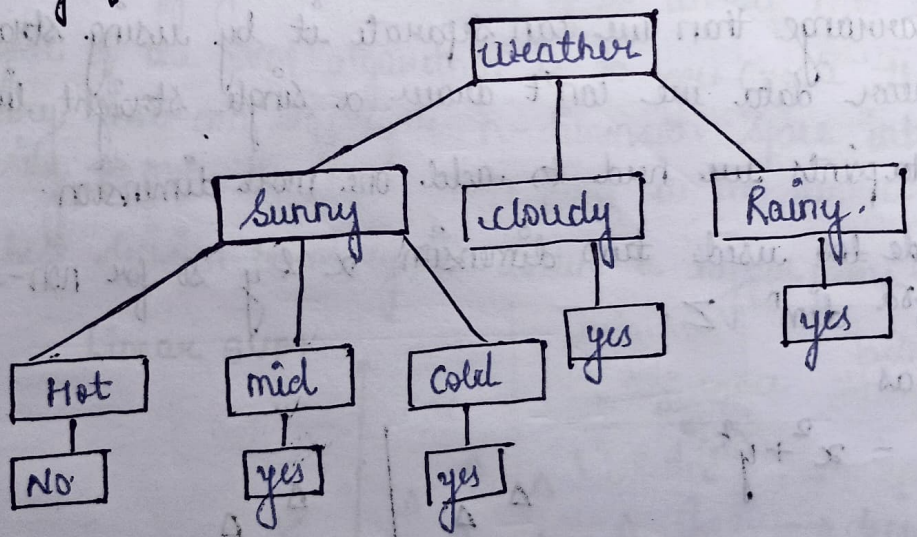


21/02/2024

Decision Tree

- Decision Tree is a machine learning algorithm used for both classification & regression.
- It is a tree structure < nodes < root node, leaf node
  edges.
- Decision nodes.
- Leaf nodes.
- Splitting.
- Entropy & information gain
- Pruning.

eg :- Play football outside



ID-3 Algorithm for Decision Tree

Day	Weather	Temperature	Humidity	Wind	Play
Day 1	Sunny	Hot	High	weak	No
Day 2	Sunny	Hot	High	Strong	No
Day 3	cloudy	Hot	High	weak	yes
Day 4	Rain	mild	High	weak	yes
Day 5	Rain	Cold	normal	weak	yes
Day 6	Rain	cold	normal	Strong	No
Day 7	cloudy	Cold	normal	Strong	yes
Day 8	Sunny	mild	high	weak	No
Day 9	Sunny	Cold	normal	weak	yes
Day 10	Rain	mild	normal	Weak	yes
Day 11	Sunny	mild	normal	Strong	yes
Day 12	cloudy	mild	high	Strong	yes
Day 13	cloudy	hot	normal	weak	yes
Day 14	Rain	mild	high	Strong	No

Entropy:

Entropy is a matric to measure the impurity in a given attribute. It specifies randomness in data. Entropy can be represented as :-

$$\text{Entropy}(S) = -P(\text{yes}) \log_2 P(\text{yes}) - P(\text{NO}) \log_2 P(\text{NO})$$

where S = Total no. of samples.

P(yes) = prob. of yes.

P(NO) = prob. of NO.

Step 1 : Entropy of entire dataset.

$$S\{+9, -5\} = -\frac{9}{14} \log_2\left(\frac{9}{14}\right) - \frac{5}{14} \log_2\left(\frac{5}{14}\right)$$

→ Total no. sample (rows)

$$= -\frac{9}{14} \times -0.63 - \frac{5}{14} \times -1.48$$

$$= 0.405 + 0.528 = 0.933 \cong 0.94$$

Step 2 Entropy of all attributes. (i) Entropy of weather.

- Entropy of Sunny = {+2 (yes), −3 (no)}

$$= -\frac{2}{5}\log_2\left(\frac{2}{5}\right) - \frac{3}{5}\log_2\left(\frac{3}{5}\right)$$

(5 ← Total no. of Sunny)

$$= 0.97.$$

- Entropy of Cloudy {+4, −0} $= -\frac{4}{4}\log_2\left(\frac{4}{4}\right) - 0$

$$= 0.$$

- Entropy of Rain {+3, −2} $= -\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$

$$= -\frac{3}{5} \times 0.73 - \frac{2}{5} \times -1.32$$

$$= 0.438 + 0.528$$

$$= 0.966$$

$$= 0.97$$

Information Gain weather

= Entropy (whole dataset) − $\frac{5}{14}$ Ent(S) − $\frac{4}{14}$ Ent(C) − $\frac{5}{14}$ Ent(R)

($\frac{5}{14}$: total no. sunny; $\frac{4}{14}$: total no of cloudy; $\frac{5}{14}$: # rain; all → Total no. of dataset)

$$= 0.94 - \frac{5}{14} \times 0.97 - \frac{4}{14} \times 0 - \frac{5}{14} \times 0.97.$$

$$= 0.246.$$

Step 3 (ii) Entropy of Temperature.

- Entropy of hot {+2, −2} $= -\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right).$

$$= 1$$

- Entropy of cold {+3, −1} $= -\frac{3}{4}\log_2\left(\frac{3}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right).$

$$=$$

- Entropy of mild {+4, −2} $= -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right) = 0.91.$

(ii) Entropy of humidity

Information gain of Temp

$$= \text{Entropy}(WD) - \frac{\text{Total \# hot}}{\text{Total \# dataset}} \times \text{Ent}(H) - \frac{\text{Total \# cold}}{\text{Total \# dataset}} \times \text{Ent}(C)$$

$$- \frac{\text{Total \# mild}}{\text{Total \# dataset}} \times \text{Ent}(M).$$

$$= 0.94 - \frac{4}{14}\text{Ent}(H) - \frac{6}{14}\text{Ent}(M) - \frac{4}{14}\text{Ent}(C)$$

$$= 0.029.$$

(iii) Entropy of humidity

(i) Entropy of high $\{+3, -4\} = -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)$

$= 0.5 + 0.46 = 0.96.$

Entropy of normal $\{+5, -2\} = -\frac{5}{7}\log_2\left(\frac{5}{7}\right) - \frac{2}{7}\log_2\left(\frac{2}{7}\right).$

$= 1$

Information gain = 0.15

(iv) Entropy of wind

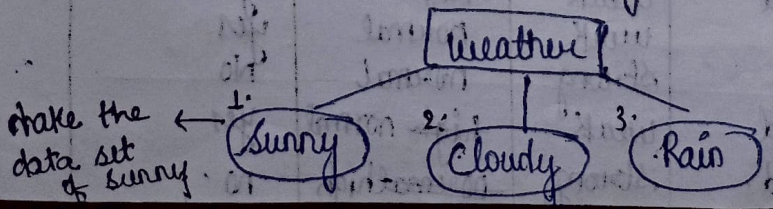
Entropy of weak $\{-6, -2\} = -\frac{6}{8}\log_2\left(\frac{6}{8}\right) - \frac{2}{8}\log_2\left(\frac{2}{8}\right) = 0.81$

Entropy of strong $\{-3, -3\} = -\frac{3}{6}\log_2\left(\frac{3}{6}\right) - \frac{3}{6}\log_2\left(\frac{3}{6}\right) = 1$

Information gain $= 0.94 - \frac{8}{14} \times \text{Entro}(W) - \frac{6}{14}\text{Ent}(S).$

$= 0.0478.$

∴ Information gain of weather is higher so it will become root node



Now, for sunny

Day	Weather	Temp	Wind	Play	Humidity
Day 1	Sunny	Hot	Weak	No	High
Day 2	Sunny	Hot	Strong	No	high
Day 8	Sunny	Mild	Weak	No	High
Day 9	Sunny	Cold	Weak	yes	normal
Day 11	Sunny	Mild	Strong	yes	normal

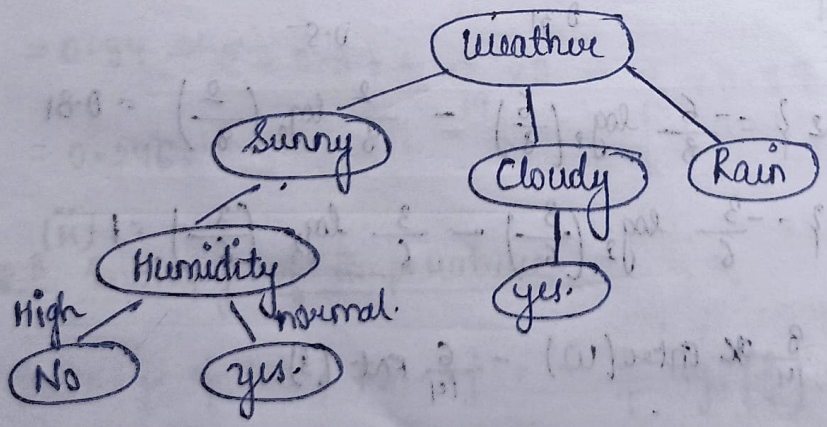
Step 1: Entropy of Sunny $\{+2, -3\} = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\left(\frac{3}{5}\right) = 0.97$.

Step 2: Entropy of all attributes.

Entropy of hot $\{0, -2\} = 0$.
Entropy of mild $\{+1, -1\} = 1$
Entropy of cold $\{1, 0\} = 0$.

Information Gain

Gain (S_{sunny}, Temp) = 0.57.
Gain (S_{sunny}, Humidity) = 0.97
Gain (S_{sunny}, wind) = 0.019.



Now for Rain

Day	Weather	Temp	Wind	Humidity	Play
4	Rain	mild	weak	high	yes
5	Rain	Cold	weak	normal	yes
6	Rain	Cold	Strong	normal	No
10	Rain	mild	weak	~~high~~ normal	yes
14	Rain	mild	Strong	~~normal~~ high	no

Step 1: Entropy of rain = 0.97

Step 2: (i) Entropy of temp.

Entropy of cold $= -\frac{1}{2}\log\left(\frac{1}{2}\right) - \frac{1}{2}\log_2\left(\frac{1}{2}\right) = 0.5 + 0.5 = 1$

Entropy of mild $= -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.38 + 0.52 = 0.90$

(ii) Entropy of wind.

Entropy of weak $\{+3, 0\} = 0$ Entropy of strong $\{0, -2\} = 0$

(iii) Entropy of humidity.

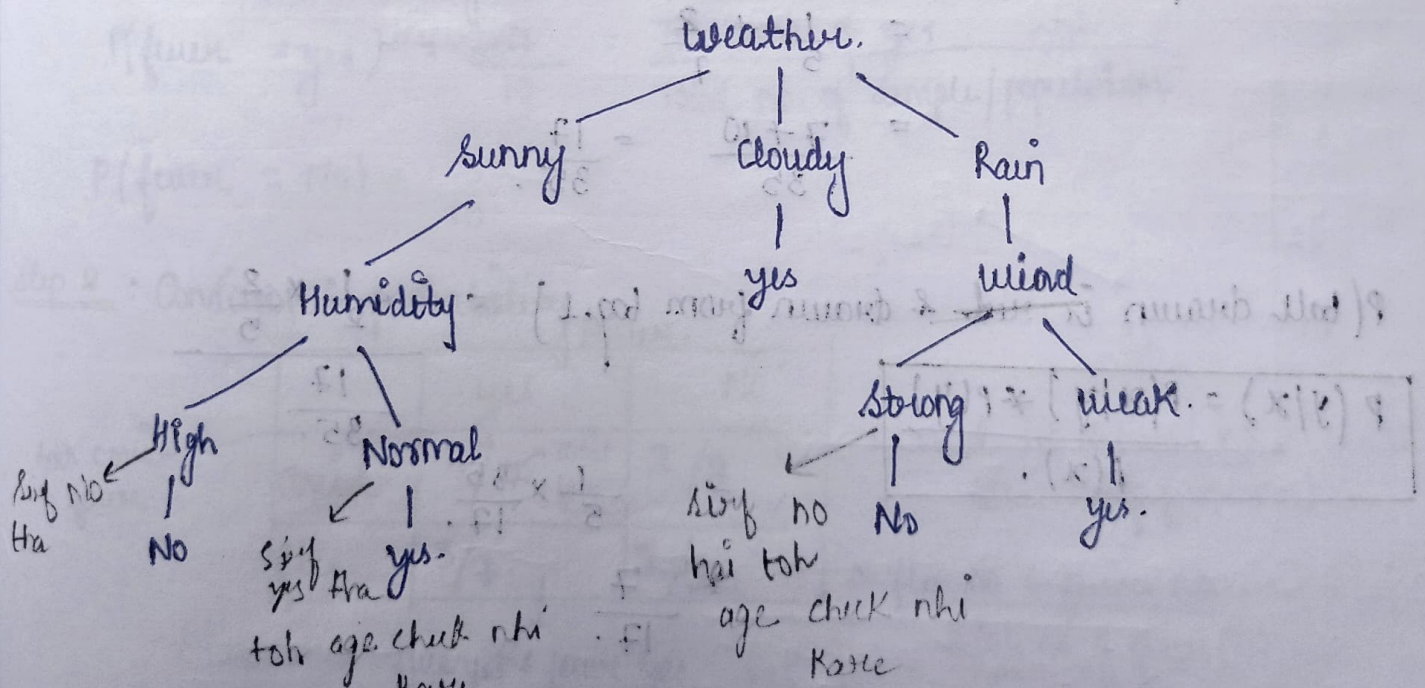
Entropy of high $\{+1, -1\} = 1$

Entropy of normal $= \{+2, -1\} = -\frac{2}{3}\log_2\left(\frac{2}{3}\right) - \frac{1}{3}\log_2\left(\frac{1}{3}\right) = 0.38 + 0.52$
$= 0.90.$

Information Gain (temp) $= 0.97 - \frac{2}{5} \times 1 - \frac{3}{5} \times 0.90 = 0.97 - 0.36 - 0.4$
$= 0.03$

Information Gain (Wind) $= 0.97 - \frac{3}{5} \times 0 - 0 = 0.97.$

Information Gain (humidity) $= 0.97 - \frac{2}{5} \times 1 - \frac{3}{5} \times 0.90$
$= 0.97 - 0.4 - 0.54 = 0.03$



26/02/2024

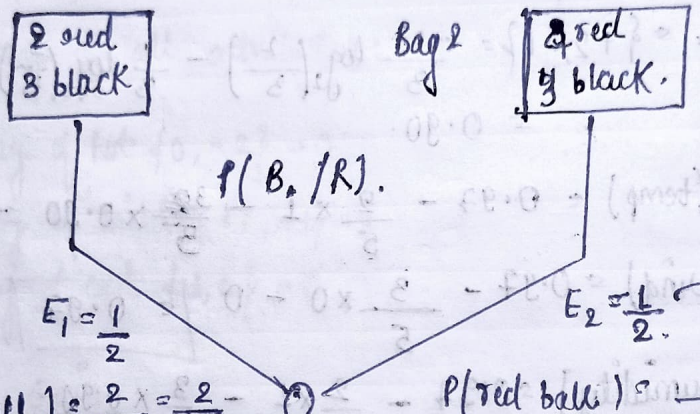
Naive - Bayes - Classification

- It is a supervised machine learning.
- It is based on applying Bayes theorem.

Bayes Theorem

Q1.) Bag 1 contains 2 red & 3 black balls. Bag 2 contain 4 red & 3 black balls. One ball is drawn at random from one of these bags & its red. Find the probability that it is drawn from bag 1.

Sol"



Total probability (P(x)) $= \frac{1}{2} \times \frac{2}{5} + \frac{1}{2} \times \frac{4}{7}$.

$= \frac{1}{5} + \frac{2}{7}$

$= \frac{7+10}{35} = \frac{17}{35}$.

P(ball drawn is red & drawn from bag 1) $= \frac{\frac{1}{2} \times \frac{2}{5}}{\frac{17}{35}}$.

$$P(y/x) = \frac{P(x/y) * P(y)}{P(x)}$$

$= \frac{1}{5} \times \frac{35}{17}$.

$= \frac{7}{17}$.

If more than one variable:

$$P(y/x_1, x_2 \cdots x_n) = \frac{P(x_1/y) * P(x_2/y) \cdots P(x_n/y) * P(y)}{P(x_1) * P(x_2) * \cdots * P(x_n)}$$

$$P(N/x) = \frac{P(x/N) * P(N)}{P(x)}$$ [Probability of no of x].

$$P(N/x) = \frac{P(x_1/N) * P(x_2/N) * \cdots * P(x_n/N) * P(N)}{P(x_1) * P(x_2) * \cdots * P(x_n)}.$$

(Q.)

Person	Covid yes/No	Flue yes/No	Fever yes/No.
1	yes	No	yes
2	No	yes	yes
3	yes	yes	yes
4	No	no	No
5	yes	no	yes
6	No	no	yes
7	yes	no	yes
8	yes	no	No
9	No	yes	yes
10	No.	yes	No.
11	yes	yes	yes

if new person is added it will be yes, beag yes has higher probability

Sol:

Step 1 : Prior Probability

$$P(\text{fever} = \text{yes}) = \frac{7}{10} = \frac{\text{Total no. of yes}}{\text{Total no. of sample/population}}$$

$$P(\text{fever} = \text{No}) = \frac{3}{10}$$

Step 2 : Conditional probability

	Fever. yes	No
Covid.	4/7 → Total fever.	2/3
flu	3/7.	2/3

both covid & fever.

Covid no & fever no / Total no. of fever (no)

flu no & fever no / Total no. of fever (no)

flue yes & fever yes / Total no. of fever (yes)

$$P(\text{yes}/\text{flu}, \text{covid}) = P(\text{flu}/\text{yes}) * P(\text{covid}/\text{yes}) * P(\text{yes})$$

$$= \frac{3}{7} \times \frac{4}{7} \times \frac{7}{10} = \frac{12}{70} = 0.17$$

$P(NO / flu, cold) = P(flu / NO) * P(cold / NO) * P(NO)$.

$= \frac{2}{3} \times \frac{2}{3} \times \frac{3}{10} 5$

$= \frac{2}{15}$

$P(NO / flu, cold) = 0.13$.

27/02/2024

KNN - K-nearest neighbor algorithm (Classification)

Data Set

K always given in question.

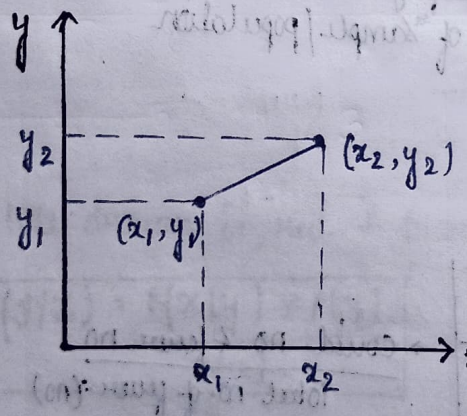
If not given generally K is taken odd value,

~~by default K=5.~~

mostly, K=5 is used.

Height (cm) (x)	Weight (Kg.) (y)	Class	Dist.	Rank
167	51	Underweight	6.7	5
182	62	Normal	13	9
176	67	Normal	~~11.6~~	8
173	64	Normal	7.6	6
172	65	Normal	8.2	7
174	56	Underweight	4.1	4
169	58	Normal	1.4	1
173	57	Normal	3	3
170	55	Normal	2	2
170	57	Normal		

Sol



Step 1 : Calculate the distance & sort:

$$d_1 = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

x_1, y_1 – variable.

x_2, y_2 – value for which we have to calculate.

$d_1 = \sqrt{(170-167)^2 + (57-51)^2} = 6.7$

$d_2 = \sqrt{(170-182)^2 + (57-62)^2} = 13$

$d_3 = \sqrt{(170-176)^2 + (57-67)^2} = 11.6$

$d_4 = \sqrt{(170-173)^2 + (57-64)^2} = 7.6$

$d_5 = \sqrt{(170-172)^2 + (57-65)^2} = 8.2$

$d_6 = \sqrt{(170-174)^2 + (57-56)^2} = 4.1$

$d_7 = \sqrt{(170-169)^2 + (57-58)^2} = 1.4$

$d_8 = \sqrt{(170-173)^2 + (57-57)^2} = 3$

$d_9 = \sqrt{(170-170)^2 + (57-55)^2} = 2$

Step 2 (rank ke nearest)

If K = 1 = Normal = new value can be classify as normal.

If K = 3 = Normal = new value can be classify as normal

If K = 5 = 3 normal & 2 underweight = consider the majority

∴ new value can be classify as normal

Step 3 : Majority voting.

K = 5 : 3 normal & 2 underweight = 3 normal

∴ consider majority.

Eg 2 Predicting movie genre.

IMDb Rating	Duration	Genere	Distance	Rank
8.0 (Mission Impassible)	160	Action	46.0	2
6.2 (Gadar 2)	170	Action	56.01	4
7.2 (Rocky & Rani)	168	Comedy	54.0	3
8.2 (OMG 2)	155	Comedy	41	1

Now predict the genere of barbie movie with IMDb rating = 7.4 & duration 114

Soln Step 1 : Calculate the distance.

$d_1 = \sqrt{(7.4-8.0)^2 + (114-160)^2} = 46.0$

$d_2 = \sqrt{(7.4-6.2)^2 + (114-170)^2} = 56.0$

$d_3 = \sqrt{(7.4-7.2)^2 + (114-168)^2} = 54.0$

$d_4 = \sqrt{(7.4-8.2)^2 + (114-155)^2} = 41$

Step 2 If K=3

2 comedy & 1 action.

If K=1. = comedy.

Step 3 Barbie movie comes under comedy genere (majority voting by).

29/02/2024

(KNN) K-means Clustering

$A_1(2,10)$, $A_2(2,5)$, $A_3(8,4)$

$B_1(5,8)$, $B_2(7,5)$, $B_3(6,4)$

$C_1(1,2)$, $C_2(4,9)$.

Sol' Suppose initially we assign A_1, B_1, C_1 as the center of each cluster respectively.

Datapoints		Distance to						cluster	New cluster
x_2	y_2	2	10	5	8	1	2	Smallest	
2	10	0		3.61		8.06		1	
2	5	5		4.24		3.16		3	
8	4	8.49		5.00		7.28		2	
5	8	3.61		0		7.21		2	
7	5	7.07		3.61		6.71		2	
6	4	7.21		4.12		5.39		2	
1	2	8.06		7.21		0		3	
4	9	2.24		1.41		7.62		2	
		C_1		C_2		C_3			

Step 1 Calculate the distance (C_1)

$$d = \sqrt{(x_2-x_1)^2 + (y_2-y_1)^2}$$

$d = \sqrt{(2-2)^2 + (10-10)^2} = 10$ $\quad d = \sqrt{(2-2)^2 + (5-10)^2} = 5$

C_2 (cluster 2) : $d = \sqrt{(x_2-5)^2 + (y_2-8)^2}$

Step 2 : Under cluster column, find smallest value among 3 clusters - & write the cluster no. having smallest value.

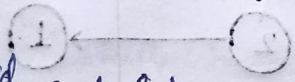
Step 3 : Calculate new Centroid

1st centroid : $x = \frac{2}{1} = 2$

$y = \frac{10}{1} = 10$

2nd Centroid : $y = \frac{4+8+5+4+9}{5} = 6.$

$xy = \frac{8+5+7+6+4}{5} = 6$



3rd Centroid :

$x = \frac{2+1}{2} = 1.5$

$y = \frac{5+2}{2} = 3.5.$

Initial Centroid	New Centroid
A_1 : (2, 10)	A_1 : (2, 10)
B_1 : (5, 8)	B_1 : (6, 6)
C_1 : (1, 2)	C_1 : (1.5, 3.5)

(from prev. table)

Datapoints		Distance to						Cluster	New Cluster
		2	10	6	6	1.5	3.5		
2	10	0		5.68		6.52		1	1
2	5	5		4.12		1.58		3	3
8	4	8.49		2.83		6.52		2	2
5	8	3.61		2.24		5.70		2	2
7	5	7.07		1.41		5.70		2	2
6	4	7.21		2.00		4.53		2	2
1	2	8.06		6.41		1.58		3	3
4	9	2.24		3.61		6.04		2	1

→ C_1 ki value small est hai

The datapoint is moved to one cluster to another cluster here so this is not converged here so we need to calculate new centroid.

New Centroid

A_1 : [3, 9.5]

B_1 : [6.5, 5.25]

C_1 : [1.5, 3.5].

	Datapoints		Distance to						Cluster (value of prev new cluster)	New Cluster
			3	9.5	6.5	5.25	1.5	3.5		
A_1	2	10	1.12		6.54		6.52		1	1
A_2	2	5	4.61		4.51		1.58		3	3
A_3	8	4	7.43		1.95		6.52		2	2
B_1	5	8	2.50		3.13		5.70		2 →	1
B_2	7	5	6.02		0.56		5.70		2	2
B_3	6	4	6.26		1.35		4.53		2	2
C_1	1	2	7.76		6.39		1.58		3	3
C_2	4	9	1.12		4.51		6.04		1	1

New Centroid

→ from new cluster

A_1 : $x = \frac{2+4+5}{3} = 3$. $= 3.67$.

$y = \frac{10+8+9}{3} = \frac{27}{3} = 9$

B_1 : $x = \frac{8+7+6}{3} = 7$.

$y = \frac{4+5+4}{3} = 4.33$

C_1 : [1.5, 3.5].

Datapoint			Distance 1 to 1		Distance 2 to 2		Distance 3 to 3		(Old) new Cluster	New Cluster
			3.67	9	7	4.33	1.5	3.5	Cluster	
A_1	2	10	1.94		7.56		6.52		1	1
A_2	2	5	4.33		5.04		1.58		3	3
A_3	8	4	6.62		1.05.		6.52		2	2
B_1	5	8	1.67		4.18		5.70		1	1
B_2	7	5	5.21		0.67		5.70		2	2
B_3	6	4	5.52		1.05		4.53		2	2
A_1	1	2	7.49		6.44.		1.58		3	3
C_2	4	9.	0.33		5.55.		6.04		1	1.

1st cluster : Data point A_1, B_1, C_2
2nd cluster : Data point A_3, B_2, B_3
3rd cluster : Data point A_2, C_1

} time 1 : 1st cluster
time 2 : 2nd

01/03/2024

(Q.)

Datapoint	Coordinates
A_1	(2, 10)
A_2	(2, 6)
A_3	(11, 11)
A_4	(6, 9)
A_5	(6, 4)
A_6	(1, 2)
A_7	(5, 10)
A_8	(4, 9)
A_9	(10, 12)
A_{10}	(7, 5)
A_{11}	(9, 11)
A_{12}	(4, 6)
A_{13}	(3, 10)
A_{14}	(3, 8)
A_{15}	(6, 11)

Centroids

Centroid 1 : (2, 6)
Centroid 2 : (5, 10)
Centroid 3 : (6, 11)

Sol^n

Datapoints		Distance to						Clusters
		x_2	y_2					
x_1	y_1	2	6	5	10	6	11	
2	10	4		3		4.12		2
2	6	0		5		6.40		1
11	11	10.2		6.08		5		3
6	9	5		1.41		2		2
6	4	4.47		6.08		7		1
1	2	4.12		8.94		10.2		1
5	10	5		0		1.41		2
4	9	3.6		1.41		2.82		2
10	12	10		5.38		4.12		3
7	5	5.09		5.38		6.08		1
9	11	8.60		4.12		3		3
4	6	2		4.12		5.38		1
3	10	4.12		2		3.16		2
3	8	2.23		2.82		4.24		1
6	11	6.40		1.41		0		3

$\therefore A_1 = \sqrt{(2-2)^2 + (6-10)^2} = \sqrt{0+16} = 4$

New Centroid

C_1: $x = \frac{2+6+1+7+4+3}{6} = 3.85$

$y = \frac{6+4+2+5+6+8}{6} = 5.16$

C_2: $x = \frac{2+5+3+4}{4} = 3.5$ $\quad x = \frac{2+6+5+4+3}{5} = 4$

$y = \frac{10+10+10+9}{4} = 9.75$ $\quad y = \frac{10+9+9+10+10}{5} = 9.6$

C_3: $x = \frac{11+6+10+11+6}{5} = 8.4$ $\quad x = \frac{11+10+9+6}{4} = 9$

$y = \frac{9+12+11+11+11}{5} = 10.8$ $\quad y = \frac{11+12+11+11}{4} = 11.25$

Datapoints		Distance to 3·8	5·16	4·9.6	9 11·25	Cluster	New Cluster
2	10	5·06	2·03	7·11	2	2
2	6	1·79	4·11	8·75	1	1
11	11	9·5	7·13	2·01	3	3
6	9	4·5	2·08	3·75	2	2
6	4	2·75	5·94	7·84	1	1
1	2	4·02	8·17	12·22	1	1
5	10	5·06	1·07	4·19	2	2
4	9	3·87	0·6	5·48	2	2
10	12	9·43	6·46	1·25	3	3
7	5	3·5	5·49	6·56	1	1
9	11	8·02	5·19	0·25	3	3
4	6	0·97	3·6	7·25	1	1
3	10	4·86	1·07	6·12	2	2
3	8	2·86	1·88	6·82	1	2
6	11	6·35	2·44	3·01	3	2

New centroid

C_1: $x = \frac{2+6+1+7+4}{5} = 4$

$y = \frac{6+4+2+5+6}{5} = 4.6$

C_2: $x = \frac{2+6+5+4+3+3+6}{7} = 4.14$

$y = \frac{10+9+10+9+10+8+11}{7} = 9.57$.

C_3: $x = \frac{11+10+9}{3} = 10$

$y = \frac{11+12+11}{3} = 11.33$.

Datapoints		Distance to 4	4.6	4.14	9.57	10	11.33	Cluster	New Cluster
A_1	2 10	5.75		2.18		8.10		2	2
A_2	2 6	2.44		4.16		9.61		1	1
A_3	11 11	9.48		7.00		1.05		3	3
A_4	6 9	4.83		1.94		~~8.35~~ 4.62		2	2
A_5	6 4	2.08		5.87		8.35		1	1
A_6	1 2	3.96		8.19		12.96		1	1
A_7	5 10	5.49		0.96		5.17		2	2
A_8	4 9	4.4		0.58		6.43		2	2
A_9	10 12	9.52		6.34		0.67		3	3
A_{10}	7 5	3.02		5.39		7.00		1	1
A_{11}	9 11	8.12		5.06		1.05		3	3
A_{12}	4 6	1.4		3.57		3.02		1	1
A_{13}	3 10	5.49		1.21		7.12		2	2
A_{14}	3 8	3.54		1.94		7.75		2	2
A_{15}	6 11	6.70		2.34		4.01		2	2

∴ 1st cluster data point $A_2, A_5, A_6, A_{10}, A_{12}$

$\neq$ ~~5~~

2nd cluster data point $A_1, A_4, A_7, A_8, A_{13}, A_{14}, A_{15}$

3rd cluster data point A_3, A_9, A_{11}

04/03/2024

KNN - K-Medoid Clustering.

(Q.) Apply K-medoid clustering algorithm to form two cluster.

i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
x	2	3	3	4	6	6	7	7	8	7
y	6	4	8	7	2	4	3	4	5	6

Sol: **Step 1**: Select two medoids (Random).

$C_1 = (3,4)$.

$C_2 = (7,4)$

Step 2:

- Use the manhattan distance to find distance between datapoint & medoid
- (x_1, y_1) and (x_2, y_2) are datapoint.
- Manhattan distance $= |x_1 - x_2| + |y_1 - y_2|$.
- $M dist_{C_1}[(2,6),(3,4)] = |2-3| + |6-4| = 3$.

$M dist_{C_1}[(3,4),(3,4)] = |3-3| + |4-4| = 0$.

$M dist_{C_1}[(3,8),(3,4)] = |3-3| + |8-4| = 4$

⋮

Similarly $M dist_{C_2}[(2,6),(7,4)] = |2-7| + |6-4| = 7$.

$M dist_{C_2}[(3,4),(7,4)] = |3-7| + |4-4| = 4$.

⋮

i	x	y	C_1 (3,4)	C_2 (7,4)	Smaller Cluster
x_1	2	6	3	7	C_1
x_2	3	4	0	4	C_1
x_3	3	8	4	8	C_1
x_4	4	7	4	6	C_1
x_5	6	2	5	3	C_2
x_6	6	4	3	1	C_2
x_7	7	3	5	1	C_2
x_8	7	4	4	0	C_2
x_9	8	5	6	2	C_2
x_{10}	7	6	6	2	C_2

C_1: {(2,6), (3,4), (3,8), (4,7), ~~(6,2)~~}

C_2: {(6,4), (7,3), (7,4), (8,5), (7,6), (6,2)}

Step 3: Calculate the total cost.

$Cost = \sum_i |C_i - x_i|$

Total cost = {cost [(3,4), (2,6)] + cost [(3,4), (3,8)] + cost [(3,4), (3,4)]
+ cost [(3,4), (4,7)] + cost [(7,4), (6,2)] + cost [(7,4), (6,4)] +
cost [(7,4), (7,3)] + cost [(7,4), (7,4)] + cost [(7,4), (8,5)] +
cost [(7,4), (7,6)].

= 3+4+4+2+3+1+1+2.

Cost = 20.

Step 4: Randomly select one non medoid point & recalculate the cost.

$C_1 = (3,4)$ & $C_2 = (7,4)$

Now $C_2 = O = (7,3)$.

Now, Calculate the distance.

M $dist_0 = [(2,6), (7,4)] = |2-7| + |6-4| = 8$

i	x	y	C_1 (3,4)	O (7,4)	cluster
x_1	2	6	3	8	C_1
x_2	3	4	0	5	C_1
x_3	3	8	4	9	C_1
x_4	4	7	4	7	C_1
x_5	6	2	5	2	C_2
x_6	6	4	3	2	C_2
x_7	7	3	5	0	C_2
x_8	7	4	4	1	C_2
x_9	8	5	6	3	C_2
x_{10}	7	6	6	3	C_2

New cluster are:

C_1 (3,4) = {(2,6), (3,4), (3,8), (4,7)}

O (7,3) = {(6,2), (6,4), (7,3), (7,4), (8,5), (7,6)}

Calculate the new cost = Cost((3,4), (2,6)) + Cost((3,4), (3,4)) + cost((3,4), (3,4)) +
cost((3,4), (4,7)) + cost((7,3), (6,2)) + cost((7,3), (6,4)) +
Cost((7,3), (7,3)) + cost((7,3), (7,4)) + cost((7,3), (8,5)) +
Cost((7,3), (7,6)).

new cost = 22

Based on this new & prev. total cost we have to decide.

S = Current total cost − previous total cost

= 22 − 20 = 2 > 0

Cost S is greater than 0, so C_2 replacing with O is not a good idea bcoz prev. cost was less than new cost.

So the final medoids are C_1 (3,4) & C_2 (7,4) i.e the prev. medoid in this case