MLFA

KNN & DT

Feb

1.)

ID	X_1	X_2	Y
1	−1	−1	2
2	−1	2	1
3	−1	4	1
4	0	−1	1
5	0	0	1
6	0	3	2
7	1	−1	2
8	1	0	2
9	2	0	3
10	2	3	3
11	−1	0	?

2020

WEDNESDAY
DAY (050-316)
8th Week

19

Manhattan distance =

$(X1, Y1)$ & $(X2, Y2)$

$$|X_1 - X_2| + |Y_1 - Y_2|$$

Distance matrix :→ (To calculate distance amongst samples.

	1	2	3	4	5	6	7	8	9	10	11	12
1		3	5	1	2	5	2	3	4	7	1	5
2			2	4	3	2	5	4	5	4	2	2
3				6	5	2	7	6	7	4	4	4
4					1	4	1	2	3	6	2	4
5						3	2	1	2	5	1	3
6							5	4	5	2	4	2
7								1	2	5	3	3
8									1	4	2	2
9										3	3	3
10											6	2
11												4
12												

MAR
APR
MAY
JUN

MAR '20

S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T	W	T	F	S	S	M	T
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31

For sample 11 :-

k = 1, => Nearest Neighbours = Samples 1 & 5.

=> $Y_1 = 2$, and $Y_5 = 1$.

(It's a tie)

=) Increment k, k = 2 =)

Nearest neighbours = Samples 1 & 5.

(It's a tie)

=) Increment k, k = 3 =)

Nearest neighbours = Samples 1, 5 and 2, 8

=) $Y_1 = 2$, $Y_8 = 2$ (It's a tie)

$Y_5 = 1$

$Y_2 = 1$

Increment k, k = 4 =)

Nearest neighbours = Samples 1, 5, 2, 8.
(It's a tie)

Increment k, k = 5 =)

Nearest neighbours = samples 1, 5, 2, 8, 7.

=) $Y_1 = 2$, $Y_8 = 2$
$Y_5 = 1$, $Y_7 = 1$
$Y_2 = 1$,

=) Final Y for sample 11 = 2.

For sample 12 :→

k = 1, Nearest neighbours = samples 2, 6 and 10

=) $Y_2 = 1$

$Y_6 = 2$ =) (It's a tie)

$Y_{10} = 3$

Increment k, k = 2 ⇒

Nearest neighbours = Samples 2, 6, 10, 5 and 9

$Y_2 = 1$, $Y_5 = 1$

$Y_6 = 2$, $Y_9 = 3$

$Y_{10} = 3$ (tie)

Increment k, k = 3 ⇒

Nearest neighbours = samples 2, 6, 10, 5, 9, 3, 4 and 11

$Y_2 = 1$, $Y_3 = 1$

$Y_6 = 2$, $Y_4 = 1$

$Y_{10} = 3$, $Y_{11} = 1$

$Y_5 = 1$

$Y_9 = 3$ ⇒ Final classification for sample 12 = 1

Q 2.

WEDNESDAY
DAY (064-302)
10th Week
04

Final label/
Classification in weighted K-NN :→

$$y' = \underset{v}{\operatorname{argmax}} \sum_{(x_i, y_i) \in N_K} w_i \times I(v = y_i)$$

, where N_K is the K-nearest neighbour-hood ,

x_i is the feature vector ,

y_i is the label of sample x_i

and w_i is the weight of sample i.

In our case w_i will be the inverse distance from the query point.

For sample 11 :→

$k = 1$ ⇒ Nearest neighbours = samples 6 and 10

labels	weightage
$Y_6 = C$	$Y_6 = 1$
$Y_{10} = B$	$Y_{10} = 1$

⇒ (It's a tie)

$k = 2$ ⇒ Nearest neighbours = samples 6, 10, 3, 7 and 9.

Labels	Weightage
$Y_6 = C$	$Y_6 = 1$
$Y_{10} = B$	$Y_{10} = 1$
$Y_3 = B$	$Y_3 = \frac{1}{2} = 0.5$
$Y_7 = C$	$Y_7 = \frac{1}{2} = 0.5$
$Y_9 = C$	$Y_9 = \frac{1}{2} = 0.5$

$$\Rightarrow Y_{11} = \text{Argmax}_Y \left(\left[(1) \cdot Y_6 + (0.5) Y_7 + (0.5) Y_9 \right], \left[(1) Y_{10} + (0.5) Y_3 \right] \right)$$

$$= \text{Argmax}_Y \left([2\ C], [1.5\ B] \right)$$

⇒ $Y_{11} = C$

⇒ Final classification of sample 11 = C

for sample 12 :→

k = 1 ⇒ Nearest neighbours = samples 1, 3, 5 and 9

labels	weightage
$Y_1 = A$	$\gamma_1 = \frac{1}{2} = 0.5$
$Y_3 = B$	$\gamma_3 = 0.5$
$Y_5 = B$	$\gamma_5 = 0.5$
$Y_9 = C$	$\gamma_9 = 0.5$

$$\Rightarrow Y_{12} = \underset{Y}{\text{argmax}} \Big([(0.5)\gamma_1], [(0.5)\gamma_3 + (0.5)\gamma_5], [(0.5)\gamma_9] \Big)$$

$$= \underset{Y}{\text{argmax}} \Big([0.5\ A], [1.\ B], [0.5\ C] \Big)$$

03 TUESDAY
DAY (063-303)
10th Week

Appointment | Notes | Work to do

$\Rightarrow Y_{12} = B$

$\Rightarrow$ Final classification for sample 12 - B

Q. In case of the following dataset, you want to build a decision stump. which splitting criteria will you choose?

ID	X1	X2	Y
1	1	5	A
2	6	5	B
3	2	4	A
4	6	4	B
5	1	2.5	B
6	8	2.5	A
7	5	2	A
8	3	1.5	B
9	1	1	B
10	6	0	A

(a) $X1 > 4$ (b) $X1 > 5.5$ (c) $X2 > 3$ (d) $X2 > 0.5$

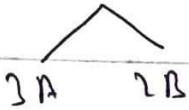
→ Solution:-

Here, we have 2 classes, A & B; out of which 5 are labelled as A & 5 as B.

$$\text{Entropy}([5A, 5B]) = -\left(\frac{5}{10}\right)\log_2\left(\frac{5}{10}\right) - \left(\frac{5}{10}\right)\log_2\left(\frac{5}{10}\right)$$

$$= 1.$$

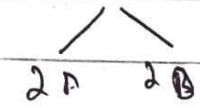
For $X_1 > 4 \Rightarrow$ 5 labels



$$\text{Entropy}([3A, 2B]) = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.9709$$

$$\text{Gain} = (1 - 0.9709) = 0.0291$$

For $X_1 > 5.5 \Rightarrow$ 4 labels



Entropy $([2A, 2B]) = 1$.

Gain $= 1 - 1$

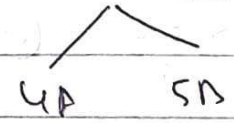
$= 0$

For $X_2 > 3 \Rightarrow$ 4 labels

Entropy $([2A, 2B]) = 0$.

Gain $= 1 - 1 = 0$.

For $X_2 > 0.5 \Rightarrow$ 9 labels



Entropy $([4A, 5B]) = 0.9910$.

Gain $= 1 - 0.9910$

$= 0.009$

Highest gain for $X_1 > 4$

$\therefore$ $X_1 > 4$.

Q. Consider the following dataset.

ID	1	2	3	4	5	6	7	8
X_1	1	1	1	1	2	2	2	2
X_2	1	1	2	2	1	1	2	2
X_3	1	2	1	2	1	2	1	2
Y	3	10	4	12	2	9	2	?

We want to build a regression stump, using variance reduction as the splitting criteria
(original variance − split1 fraction * split1 variance − split2 fraction * split2 variance).

On which attribute will you split?

(a) X_1 (b) X_2 (c) X_3 (d) All equivalent.

→ Solution:-

By observation,
When $X_3 = 1 \Rightarrow$ Y values are low.
When $X_3 = 2 \Rightarrow$ Y values are high.
However, this is not the case with X_1 & X_2.

∴ Splitting should be on X_3.

don't do this way!

Original Variance = $Var(Y) = \frac{\Sigma (Y - \bar{Y})^2}{n}$

$\bar{Y} = 6$

$Var(Y) = \frac{106}{7} = 15.14$

We have 3 attributes X_1, X_2 & X_3.

For X_1

X_1 [1, 2]

1 → [3, 10, 4, 12]

2 → [7, 9, 2]

Split 1 variance

$$\bar{X}_1(1) = \frac{29}{4}$$

$$Var(X_1[1]) = \frac{235}{16}$$

Split 2 variance

$$\bar{X}_1(2) = \frac{13}{3}$$

$$Var(X_1[2]) = \frac{98}{9}$$

$$\text{Variance Reduction } (X_1) = \frac{106}{7} - \frac{4}{7} \times \frac{235}{16} - \frac{3}{7} \times \frac{98}{9}$$

$$= \frac{25}{12} = 2.083$$

For X_2

X_2 [1, 2]

1 → [3, 10, 2, 9]

2 → [4, 12, 2]

Split 1 variance

$$\bar{X}_2(1) = 6$$

$$Var(X_2[1]) = \frac{25}{2}$$

Split 2 Variance

$$\bar{X}_2(2) = 6$$

$$Var(X_2(2)) = \frac{56}{3}$$

$$\text{Variance Reduction} = \frac{106}{7} - \frac{4}{7} \times \frac{25}{2} - \frac{3}{7} \times \frac{56}{3} = 0$$

For X_3

$X_3[1, 2]$

$1 \rightarrow [3, 4, 2, 2]$

$2 \rightarrow [14, 17, 9]$

Split 1 Variance

$\bar{X}_3(1) = \frac{11}{4}$

$\text{Var}(X_3(1)) = \frac{11}{16}$

Split 2 Variance

$\bar{X}_3(2) = \frac{31}{3}$

$\text{var}(X_3(2)) = \frac{14}{9}$

$$\text{Variance Reduction} = \frac{106}{7} - \frac{4}{7} \times \frac{11}{16} - \frac{3}{7} \times \frac{14}{9}$$

$$= \frac{169}{12} = 14.083$$

We choose the attribute with minimum variance reduction.

∴ Splitting will be done on X_2.