MCSE-642; Data Mining

Name: Shyed Shahriar Housaini; ID: 07905536

Ans. to the q. no-1 (a)

sample: [34, 21, 56, 76, 98]

Max: 98 ; New max: 20

Min: 21 ; New min: 10

Normalized values are:

For; 34:-

$$v' = \frac{34-21}{98-21}[20-10]+10$$

$= 11.69$ Ans:

For 21:

$$v' = \frac{21-21}{98-21}[20-10]+10 = 10 \text{ Ans.}$$

For 56:-

$$v' = \frac{56-21}{98-21}[20-10]+10 = 14.55 \text{ Ans.}$$

For 76:-

$$v' = \frac{76-21}{98-21}[20-10]+10 = 17.143 \text{ Ans.}$$

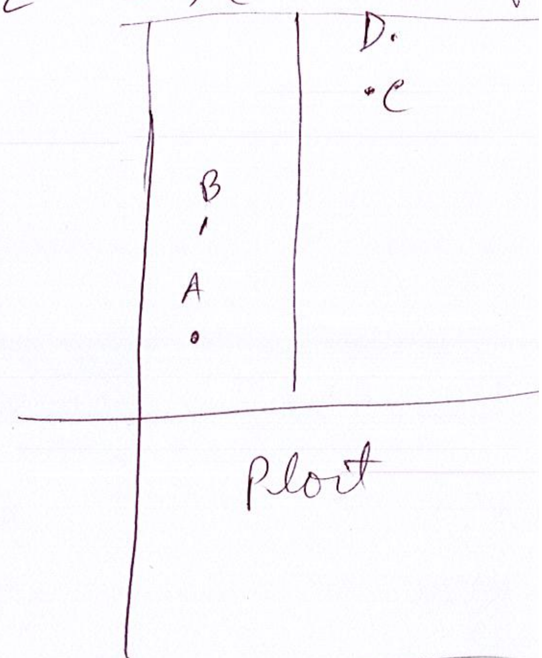
For 98:-

$$v' = \frac{98-21}{98-21}[20-10]+10 = 20 \text{ Ans.}$$

Ans. to the q. no- 1 (b)

Let $C_1 = A, (1,1)$; group
$C_2 = B, (1,2)$ group

A 1, 1
B 1, 2
C 3, 4
D 4, 5



Plot

$$D^0 = \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 1 & 3.6 & 5 \\ 1 & 1.4 & 3.16 & 4.5 \end{bmatrix} \end{array}$$

Object clustering 1:

$$G^0 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

New centroids are:

$$\left(\frac{1+1}{2}\right), \left(\frac{1+2}{2}\right) = C_1$$

$$\left(\frac{3+4}{2}\right), \left(\frac{4+5}{2}\right) = C_2$$

$$D^1 = \begin{array}{c} \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0.5 & 1.12 & 3.35 & 4.71 \\ 4.3 & 3.81 & 1.58 & 1.58 \end{bmatrix} \end{array}$$

$$G^1 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

Centroids are c_1 1, 1·5

c_2 3·5, 4·5

Point	vector	Cluster ID	Cluster Centroids
A	1,1	0	1, 1.5
B	1,2	0	1, 1.5
C	3,4	1	3·5, 4.5
D	4,5	1	3·5, 4.5

Ans. to the q. no- 3(a)

Issues with K-fold cross validation:

1) Using simple K-Fold (non-satisfied cv) on a classification problem can be tricky.

2) Since we are randomly shuffling the data and then dividing it into folds, chances are we may get highly imbalanced fold which may cause our training to be biased.

3) For example, if we have a fold that has majority belonging to one class (say positive) and only a few as negative clas. This will ruin our training.

Satified K-fold cross validation solves those ~~ifs~~ issues, by satisfied sampling.

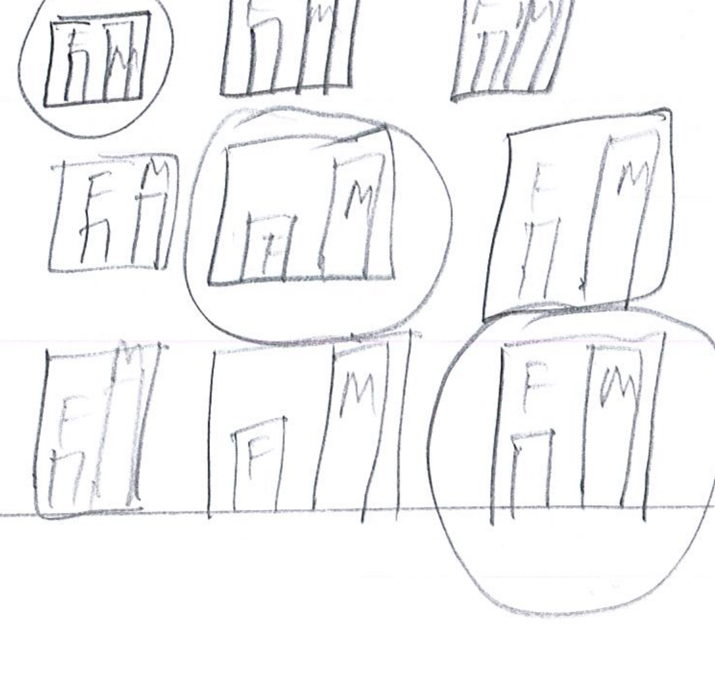
Satisfied Sampling is a sampling technique where the samples are selected in the same population, by dividing them in groups.

For example, if a population have 70% male and 30% female subjects, we divide the population into two groups and choose 30% female & 70% male sample from respective groups.

Satisfied K-fold cross validation

K = 3

class distribution



Ans. To q. no-3 (b)

TP = 4

TN = 3

FP = 1

FN = 2

Predictive level \ Actual Level	1	0
1	TP (4)	FP (1)
0	FN(2)	TN(3)

$$\text{Accuracy} = \frac{TN+TP}{TN+FP+TP+FN} = \frac{3+4}{3+1+4+2}$$

$$= \frac{7}{10} = 0.7$$ Ans

~~For score~~ $\text{Precision} = \frac{TP}{TP+FP} = 0.8$

$\text{Recall} = \frac{TP}{TP+FN} = 0.6667$

$$F_1 \text{ score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$= 2 \times \frac{0.8 \times 0.6667}{0.8 + 0.6667}$$

$$= 0.7273$$ Ans

Ans. to the q. no 4 (a)

Total net input of h_1

$$net\ h_1 = i_1 w_1 + i_2 w_3 + i_3 w_5$$

$$= .8 \times .4 + .09 \times .07 + .49 \times .89$$

$$= 0.4744$$

$$Out\ h_1 = \frac{1}{1+e^{-net\ h_1}} = \frac{1}{1+e^{-0.4744}}$$

$$= \frac{1}{1+0.62} = 0.62$$

$$net\ h_2 = i_1 \times w_2 + i_2 w_4 + i_3 w_6$$

$$= .8 \times .5 + .09 \times .63 + .49 \times .12$$

$$= 0.52$$

$$Out\ h_2 = \frac{1}{1+e^{-0.52}} = \frac{1}{1.6} = 0.63$$

$$\therefore net\ O_1 = w_7 \times Out_{h_1} + w_9 \times Out_{h_2}$$

$$= 0.34 \times .62 + .44 \times 0.63$$

$$= 0.49$$

$$Out_{O_1} = \frac{1}{1+e^{-net_{O_1}}} = \frac{1}{1+e^{-0.49}} = \frac{1}{1+0.61}$$

$$= 0.62$$

$$net_{O_2} = w_8 \times Out_{h_1} + w_{10}\, Out_{h_2}$$

$$= .03 \times .62 + .59 \times .63$$

$$= 0.39$$

$$\therefore Out_{O_2} = \frac{1}{1+e^{-0.39}} = \frac{1}{1+0.68} = 0.595$$

Ans

Ans. to the q. no-4 (b)

We have to update w_7.

$O_1 = 0.8$ | $\eta = 0.5$

$O_2 - 0.2$

$$\frac{\partial E_{total}}{\partial w_7} = \frac{\partial E_{total}}{\partial out_1} \times \frac{\partial Out_{O_1}}{\partial net_{O_1}} \times \frac{\partial net_{O_1}}{\partial w_7}$$

$$E_{total} = \cancel{\quad} - \frac{1}{2}(target_{O_1} - Out_{O_1})^2 + \frac{1}{2}(target_{O_2} - Out_{O_2})^2$$

$$\frac{\partial E_{total}}{\partial Out_{O_1}} = 2 \times \frac{1}{2}(target_{O_1} - Out_{O_1})^2 \times -1 + 0$$

$$= -(.8 - .62) = -0.18$$

$$\therefore Out_{O_1} = \frac{1}{1+e^{-net_{O_1}}}$$

$$\frac{\partial Out_{O_1}}{\partial net_{O_1}} = Out_{O_1}(1 - Out_{O_1}) = 0.2356$$

$net_{o1} = w_7 \times Out_{h1} + w_9 \times Out_{h2}$

$$\frac{\partial net_{o1}}{\partial w_7} = 1 \times Out_{h1} \times w_7^{(1-1)}$$

$$= Out_{h1} = \cdot 62$$

$$\therefore \frac{\partial E_{total}}{\partial w_7} = (-0{\cdot}18)(0{\cdot}2356)(\cdot 62)$$

$$= -0{\cdot}026$$

$\eta = 0{\cdot}5$

$$w_7^+ = w_7 - \eta \frac{\partial E_{total}}{\partial w_7}$$

$$= \cdot 34 - 0{\cdot}5 \times (-0{\cdot}026)$$

$$= 0{\cdot}357$$

Ans.

Ans. to q. no-2 (a)

Transactin ID	Item Purchased
1	A, B, E
2	B, D
3	A, B, D
4	B, C
5	A, B, C E

Support = 30% = $\frac{30}{100} \times 5$ = ~~1~~'5 ≈ 2

Confidence = 65%

Item set	Support
{A}	3
{B}	5
{C}	2
{D}	2
{E}	2

Item set	Support
{A, B}	3
{A, C}	1
{A, D}	1
{A, E}	2
{B, C}	2
{B, D}	2
{B, E}	2
{C, D}	0
{C, E}	1
{D, E}	0

Item Set	Support
{A, B}	3
{A, E}	2
{B, C}	2
{B, D}	2
{B, E}	2

Item Set	Support
{A, B, E}	2
{B, C, D}	0
{B, D, E}	0
{B, C, E}	1
{A, B, C}	1
{A, B, D}	1

Item Set	Support
{A, B, E}	2

(P. T. O)

Rule	Support	Confidence	Confidence (%)
$\{A,B\} \rightarrow \{E\}$	2	$2/3 = 0.67$	$67\% > 65\%$
$\{A,E\} \rightarrow \{B\}$	2	$2/2 = 1$	$100\% > 65\%$
$\{B,E\} \rightarrow \{A\}$	2	$2/2 = 1$	$100\% > 65\%$
$\{E\} \rightarrow \{A,B\}$	2	$2/2 = 1$	$100\% > 65\%$
$\{B\} \rightarrow \{A,E\}$	2	$2/5 = 0.4$	$40\% < 65\%$
$\{A\} \rightarrow \{B,E\}$	2	$2/3 = 0.67$	$67\% > 65\%$

~~Foff~~ Following rules are found

$\{A,B\} \rightarrow \{E\}$

$\{A,E\} \rightarrow \{B\}$

$\{B,E\} \rightarrow \{A\}$

$\{E\} \rightarrow \{A,B\}$

$\{A\} \rightarrow \{B,E\}$

Ans.

Ans. to the q. no- 2 (b)

Future-x	Class		Total
	Yes	NO	
High	2	2	4
Medium	4	2	6
Low	3	1	4

$$Gain = Entropy(P) - \left(\sum_{i=1}^{k} \frac{n_1}{n} Entropy(i)\right)$$

$$Entropy(t) = -\sum_{j} P(j/t) \log_2 (j/t)$$

~~Entropy (High)~~ $= \frac{2}{4} = .5 = P_1$

$= \frac{2}{4} = .5 = P_2$

$$Entropy(class) = -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= 0.94$$

$$Entropy(x, High) = \frac{-2}{4}\log_2\left(\frac{2}{4}\right) - \left(\frac{2}{4}\right)\log_2\frac{2}{4}$$

$$= 1$$

Entropy(x, Medium)

$$= -\frac{4}{6}\log_2\left(\frac{4}{6}\right) - \frac{2}{6}\log_2\left(\frac{2}{6}\right)$$

$$= \cancel{0.25}\ 0.83$$

Entropy (x, low)

$$= -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\left(\frac{1}{4}\right)$$

$$= 0.82$$

Information gain (class, future)

$$= E\ (\text{class} - E\ (\text{class}, \text{future})$$

$$= 0.94 - 0.91$$

$$= .03$$

$$\therefore \text{Gain Ratio} = \frac{0.03}{0.91} = 0.033$$

Ans