Homework - 2

Yogesh Kalapala.
1001879640.
yxk9640.

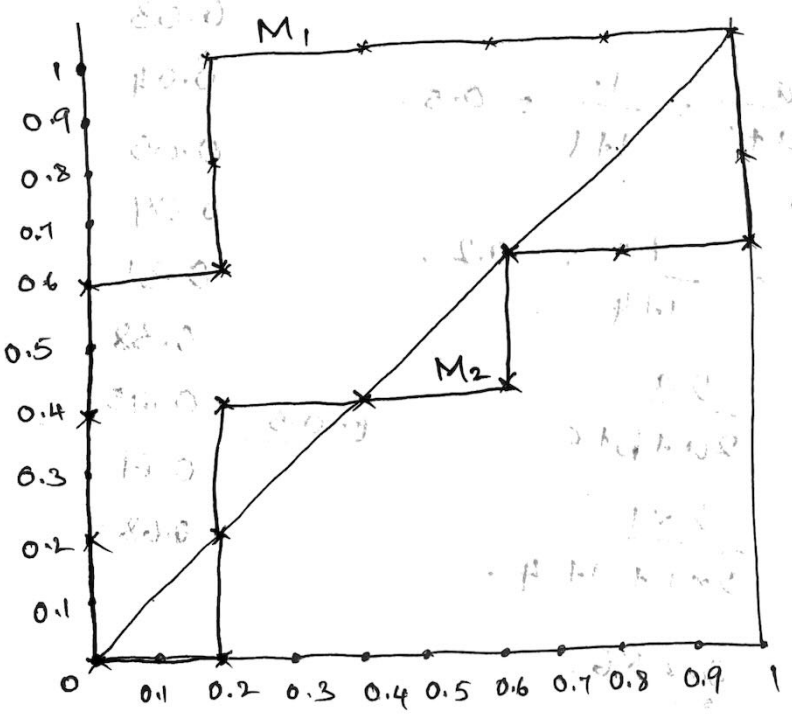
1) Classifier Evaluation:

a) ROC curve:

Number of positive points = 5 · = 1/5 = 0.2.

Number of negative points = 5 = 1/5 = 0.2 .

Model - 1	Class.	Model - 2	Class.
0.08	−	0.01 .	+
0.15	−	0.03	+
0.35	−	0.04 .	−
0.44	−	0.05	−
0.45	+	0.09	+
0.47 .	+	0.31	−
0.55	−	0.38	−
0.67	+	0.45	+
0.69	+	0.61 .	+
0.73	+	0.68	−



ROC Curve

∴ M_1 is better than M_2 as Area under M_1 is more.

b). Model - 1. (threshold - 0.5).

TP = 3; FN = 2; FP = 1; TN = 4.
(a) (b) (c) (d)

$$\text{Precision} = \frac{a}{a+c}$$

$$= \frac{3}{3+1} = 0.75.$$

$$\text{Recall} = \frac{a}{a+b}$$

$$= \frac{3}{3+2} = 0.6.$$

$$\text{F-measure} = \frac{2a}{2a+b+c}$$

$$= \frac{2\times 3}{2\times 3+2+1} = 0.667.$$

Model 1	Class	Updated.
0.08	−	−
0.15	−	−
0.35	−	−
0.44	−	−
0.45	+	−
0.47	+	−
t = 0.5		
0.55	−	+
0.67	+	+
0.69	+	+
0.73	+	+

c) Model - 2. (threshold = 0.5)

TP(a) = 1; FN(b) = 4; FP(c) = 1;
TN(d) = 4.

$$\text{Precision} = \frac{a}{a+c} = \frac{1}{1+1} = 0.5.$$

$$\text{Recall} = \frac{a}{a+b} = \frac{1}{1+4} = 0.2.$$

$$\text{F-measure} = \frac{2a}{2a+b+c}$$

$$= \frac{2\times 1}{2\times 1+1+4}.$$

$$= 0.286.$$

Model-2	class	Updated.
0.01	+	−
0.03	+	−
0.04	−	−
0.05	−	−
0.09	+	−
0.31	−	−
0.38	−	−
0.45	+	−
t = 0.5		
0.61	+	+
0.68	−	+

2) K-Means Clustering:

Given. $K=3$ $c_1 \rightarrow A_1(2,10)$. Distance $a(x_1, y_1)$ $b(x_2, y_2)$.

$c_2 \rightarrow A_2(5,8)$.

$c_3 \Rightarrow A_7(1,2)$. $dist(a,b) = |x_2 - x_1| + |y_2 - y_1|$.

The points belong to cluster whose distance is minimum.

Iteration-1.

$A_1(2,10) \rightarrow c_1(2,10) = 0+0 = 0$.

$c_2(5,8) = |5-2| + |8-10| = 3+2 = 5$.

$c_3(1,2) = |1-2| + |10-2| = 1+8 = 9$. $\therefore A_1 \in c_1$

$A_2(2,5) \rightarrow c_1(2,10) = |2-2| + |10-5| = 5$.

$c_2(5,8) = |5-2| + |8-5| = 3+3 = 6$.

$c_3(1,2) = |1-2| + |2-5| = 1+3 = 4$. $\therefore A_2 \in c_3$.

$A_3(8,4) \rightarrow c_1(2,10) = |2-8| + |10-4| = 6+6 = 12$.

$c_2(5,8) = |5-8| + |8-4| = 3+4 = 7$. $\therefore A_3 \in c_2$

$c_3(1,2) = |1-8| + |2-4| = 7+2 = 9$.

$A_4(5,8) \rightarrow c_1(2,10) = |2-5| + |10-8| = 3+2 = 5$.

$c_2(5,8) = |5-5| + |8-8| = 0$. $\therefore A_4 \in c_2$

$c_3(1,2) = |1-5| + |2-8| = 4+6 = 10$

$A_5(7,5) \rightarrow c_1(2,10) = |2-7| + |10-5| = 5+5 = 10$.

$c_2(5,8) = |5-7| + |8-5| = 2+3 = 5$. $\therefore A_5 \in c_2$

$c_3(1,2) = |1-7| + |2-5| = 6+3 = 9$.

$A_6(6,4) \rightarrow c_1(2,10) \rightarrow |2-6| + |10-4| = 4+6 = 10$.

$c_2(5,8) \rightarrow |5-6| + |8-4| = 1+4 = 5$. $\therefore A_6 \in c_2$

$c_3(1,2) \rightarrow |1-6| + |2-4| = 5+2 = 7$.

$A_7(1,2) \rightarrow c_1(2,10) \rightarrow |2-1| + |10-2| = 1+8 = 9$.

$c_2(5,8) \rightarrow |5-1| + |8-2| = 4+6 = 10$ $\therefore A_7 \in c_3$

$c_3(1,2) \rightarrow |1-1| + |2-2| = 0+0 = 0$.

$A_8(4,9) \rightarrow C_1(2,10) \rightarrow |2-4| + |10-9| = 2+1 = 3$

$C_2(5,8) \rightarrow |5-4| + |8-9| = 1+1 = 2$ $\therefore A_8 \in C_2$

$C_3(1,2) \rightarrow |1-4| + |2-9| = 3+7 = 10.$

Updated clusters:

$C_1 = \{A_1\}.$

$C_2 = \{A_3, A_4, A_5, A_6, A_8\}$

$C_3 = \{A_2, A_7\}$

Updated cluster centroids:

$C_1 = (2,10).$

$C_2 = \left(\frac{8+5+7+6+4}{5}, \frac{4+8+5+4+9}{5}\right)$

$C_2 = (6,6)$

$C_3 = \left(\frac{2+1}{2}, \frac{5+2}{2}\right) = (1.5, 3.5).$

Iteration-2.

$A_1(2,10) \rightarrow C_1(2,10) = |2-2| + |10-10| = 0.$

$C_2(6,6) = |6-2| + |6-10| = 4+4 = 8.$ $\therefore A_1 \in C_1$

$C_3(1.5,3.5) = |1.5-2| + |3.5-10| =$

$A_2(2,5) \rightarrow C_1(2,10) = |2-2| + |10-5| = 5.$

$C_2(6,6) = |6-2| + |6-5| = 4+1 = 5.$ $\therefore A_2 \in C_3$

$C_3(1.5,3.5) = |1.5-2| + |3.5-5| = 0.5 + 1.5 = 2.$

$A_3(8,4) \rightarrow C_1(2,10) = |2-8| + |10-4| = 6+6 = 12.$

$C_2(6,6) = |6-8| + |6-4| = 2+2 = 4.$ $\therefore A_3 \in C_2$

$C_3(1.5,3.5) = |1.5-8| + |3.5-4| = 6.5 + 0.5 = 7.$

$A_4(5,8) \rightarrow C_1(2,10) = |2-5| + |10-8| = 3+2 = 5.$

$C_2(6,6) = |6-5| + |6-8| = 1+2 = 3$ $\therefore A_4 \in 3$

$C_3(1.5,3.5) = |1.5-5| + |3.5-8| = 3.5 + 3.5 + 4.5 = 8.$

$A_5(7,5) \rightarrow C_1(2,10) \rightarrow |2-7| + |10-5| = 5+5 = 10$

$C_2(6,6) \rightarrow |6-7| + |6-5| = 1+1 = 2$ $\therefore A_5 \in C_2$

$C_3(1.5,3.5) \rightarrow |1.5-7| + |3.5-5| = 5.5 + 1.5 = 7.$

$A_6(6,4) \rightarrow c_1(2,10) \rightarrow |2-6| + |10-4| = 4+6 = 10$

$c_2(6,6) \rightarrow |6-6| + |6-4| = 0+2 = 2$

$\therefore A_6 \in c_2$

$c_3(1.5,3.5) \rightarrow |1.5-6| + |3.5-4| = 4.5+0.5 = 5$

$A_7(1,2) \rightarrow c_1(2,10) \rightarrow |2-1| + |10-2| = 8+1 = 9.$

$c_2(6,6) \rightarrow |6-1| + |6-2| = 5+4 = 9.$

$c_3(1.5,3.5) = |1.5-1| + |3.5-2| = 0.5+1.5 = 2.$

$\therefore A_7 \in c_3$

$A_8(4,9) \rightarrow c_1(2,10) \rightarrow |2-4| + |10-9| = 2+1 = 3$

$c_2(6,6) \rightarrow |6-4| + |6-9| = 2+3 = 5$

$c_3(1.5,3.5) \rightarrow |1.5-4| + |3.5-9| = 2.5+5.5 = 8.$

$\therefore A_8 \in c_1$

Updated clusters:

$c_1 = \{A_1, A_8\}.$

$c_2 = \{A_3, A_4, A_5, A_6\}.$

$c_3 = \{A_2, A_7\}.$

Updated centroids:

$c_1 = \left\{\frac{2+4}{2}, \frac{10+9}{2}\right\}.$

$c_1 = (3, 9.5)$

$c_2 = \left\{\frac{8+5+7+6}{4}, \frac{4+8+5+4}{4}\right\}.$

$c_2 = (6.5, 2.5)$

$c_3 = \left\{\frac{2+1}{2}, \frac{5+2}{2}\right\}.$

$c_3 = (1.5, 3.5)$

After 2 iterations.

Cluster centroids are

$c_1 (3, 9.5)$

$c_2 (6.5, 2.5).$

$c_3 (1.5, 3.5).$

3. Hierarchical Clustering:

~~Single link / MIN~~: Complete link / Max:

Given similarity matrix. for distance to be min the similarity should be max.

distance = 1 − similarity.

First cluster to combine will be P_2, P_5 with 0.97.

$P_2 \cup P_5 = 0.97$:

	P_1	$P_2 \cup P_5$	P_3	P_4
P_1	1	0.09	0.40	0.54
$P_2 \cup P_5$		1	0.63	0.46
P_3			1	0.43
P_4				1

Next combine P_1 & P_4 as

$P_1 \cup P_4 = 0.54$.

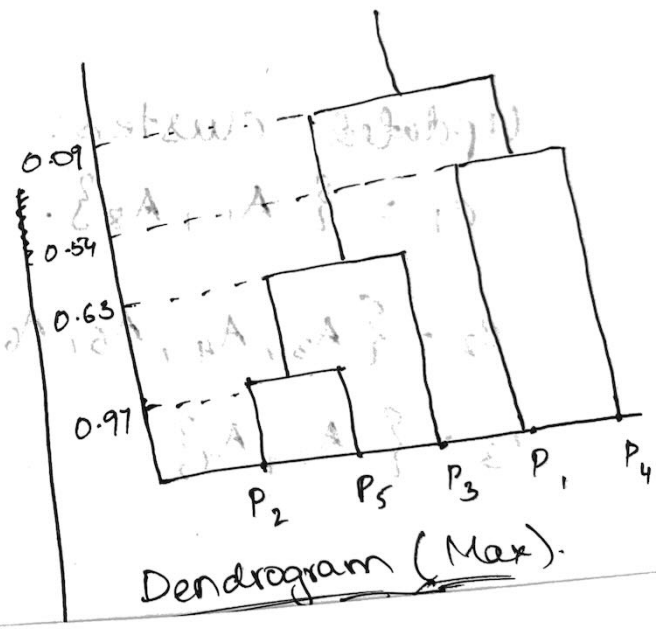
	$P_1 \cup P_4$	$P_2 \cup P_5 \cup P_3$
$P_1 \cup P_4$	1	0.09
$P_2 \cup P_5 \cup P_3$		1

Next combine P_1 & P_4 with $P_2 \cup P_5 \cup P_3$ as

$P_1 \cup P_4 \cup P_2 \cup P_5 \cup P_3 = 0.09$.

Next combine P_3 with $P_2 \cup P_5$ as $P_2 \cup P_5 \cup P_3 = 0.63$

	P_1	$P_2 \cup P_5 \cup P_3$	P_4
P_1	1	0.09	0.54
$P_2 \cup P_5 \cup P_3$		1	0.43
P_4			1



Dendrogram (Max).

~~Complete~~ ~~MAX~~ / ~~MAX~~ : Single Link / Min :

step ⓐ First cluster to combine is P_2 & P_5 as.

$P_2 \cup P_5 = 0.97$.

	P_1	$P_2 \cup P_5$	P_3	P_4
P_1	1	0.34	0.40	0.54
$P_2 \cup P_5$		1	0.84	0.75
P_3			1	0.43
P_4				1

step ⓑ Next cluster to combine P_3 with $P_2 \cup P_5$ as

$(P_2 \cup P_5) \cup P_3 = 0.84$.

	P_1	$P_2 \cup P_5 \cup P_3$	P_4
P_1	1	0.40	0.54
$P_2 \cup P_5 \cup P_3$		1	0.75
P_4			1

step: ⓒ Next, combine P_4 with $(P_2 \cup P_5 \cup P_3)$ as:

$P_4 \cup (P_2 \cup P_5 \cup P_3) = 0.75$.

	P_1	$P_2 \cup P_5 \cup P_3 \cup P_4$
P_1	1	0.54
$P_2 \cup P_5 \cup P_3 \cup P_4$		1

Finally, combine P_1 with rest as $P_1 \cup (P_2 \cup P_5 \cup P_3 \cup P_4) = 0.54$.

Dendrogram (Min)

4) Association Rule.

Transaction ID as bucket.

(a) Support (s):

Support $\{e\} = \frac{8}{10}$ occurs in $[1, 24, 12, 31, 15, 22, 33, 38]$.

$= 0.8$.

Support $\{b, d\} = \frac{2}{10}$ occurs in $[12, 22]$.

$= 0.2$

Support $\{b, d, e\} = \frac{2}{10}$ occur in $[12, 22]$.

$= 0.2$

(b) Confidence : $c(\{b, d\} \rightarrow \{e\}) = \frac{\sigma(b, d, e)}{\sigma(b, d)} = \frac{2}{2} = 1.$

$c(\{e\} \rightarrow \{b, d\}) = \frac{\sigma(b, d, e)}{\sigma(e)} = \frac{2}{8} = \underline{0.25}$

Customer ID as bucket

transactions with customer ID

$1 \rightarrow \{a, b, c, d, e\}$.

$2 \rightarrow \{a, b, c, d, e\}$.

$3 \rightarrow \{b, c, d, e\}$.

$4 \rightarrow \{a, b, c, d\}$.

$5 \rightarrow \{a, b, d, e\}$.

(c) Support (s).

support $\{e\} = \frac{4}{5} = 0.8$ $[1, 2, 3, 5]$

support $\{b, d\} = \frac{5}{5} = 1$ $[1, 2, 3, 4, 5]$

support $\{b, d, e\} = \frac{4}{5} = 0.8$ $[1, 2, 3, 5]$

(d) Confidence c: $(\{b, d\} \rightarrow \{e\}) = \frac{\sigma(\{b, d, e\})}{\sigma(\{b, d\})} = \frac{4}{5} = \underline{0.8}$.

$\{e\} \rightarrow \{b, d\} \Rightarrow \frac{\sigma(\{e, b, d\})}{\sigma(\{e\})} = \frac{4}{4} = \underline{1}.$