CSE 5243 Introduction to Data Mining
Homework 3

Ashwini Joshi
10/03/2017

Q. Chapter 2 : Question 19

a) $x = (1,1,1,1)$ $y = (2,2,2,2)$

$\|x\| = \sqrt{1^2+1^2+1^2+1^2} = 2.$ $\|y\| = \sqrt{2^2+2^2+2^2+2^2} = 4.$

$$\text{cosine} = \frac{x \cdot y}{\|x\| \|y\|} = \frac{1\times2+1\times2+1\times2+1\times2}{2\times4} = \frac{8}{8} = \boxed{1.}$$

$$\text{covariance} = \frac{1}{n-1}\sum_{i=n}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

$\bar{x} = 1$ $\bar{y} = 2.$

$$\therefore \text{covariance}_{xy} = \frac{1}{3}\left[(1-1)(2-2)+(1-1)(2-2)+(1-1)(2-2)+(1-1)(2-2)\right]$$

$$= 0.$$

$$\text{correlation} = \frac{\text{cov}}{\sigma_x \sigma_y}$$

$\because$ covariance is 0, correlation $= \boxed{0}$.

$$\text{Euclidean} = \sqrt{\sum_{k=1}^{n}(x_k-y_k)^2}$$

$$= \sqrt{(1-2)^2+(1-2)^2+(1-2)^2+(1-2)^2}$$
$$= \sqrt{4}$$
$$= \boxed{2}$$

$\therefore$ cosine = 1
correlation = 0
Euclidean = 2.

b) $x = (0,1,0,1)$ $y = (1,0,1,0).$

$$\text{cosine} = \frac{0\times1+1\times0+0\times1+1\times0}{\sqrt{2}\cdot\sqrt{2}} = \boxed{0}$$ $\bar{x} = \frac{1}{2}$ $\bar{y} = \frac{1}{2}$

Q. 2.19 b) $\text{covariance}_{xy} = \frac{1}{3}\left[\left(0-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)+\left(1-\frac{1}{2}\right)\left(0-\frac{1}{2}\right)+\left(0-\frac{1}{2}\right)\left(1-\frac{1}{2}\right)+\left(1-\frac{1}{2}\right)\left(0-\frac{1}{2}\right)\right]$

$= \frac{1}{3}\left[-\frac{1}{4}-\frac{1}{4}-\frac{1}{4}+\frac{1}{4}\right]$

$= \frac{1}{3} \times (-1) = -\frac{1}{3}.$

$\sigma_x \text{ (std. dev)} = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k-\bar{x})^2}$

$= \sqrt{\frac{1}{3}\left[\left(0-\frac{1}{2}\right)^2+\left(1-\frac{1}{2}\right)^2+\left(0-\frac{1}{2}\right)^2+\left(1-\frac{1}{2}\right)^2\right.}$

$= \sqrt{\frac{1}{3} \times \left[\frac{1}{4} \times 4\right]}$

$= \frac{1}{\sqrt{3}}$

$\sigma_y = \sqrt{\frac{1}{3}\left[\left(1-\frac{1}{2}\right)^2+\left(0-\frac{1}{2}\right)^2+\left(1-\frac{1}{2}\right)^2+\left(0-\frac{1}{2}\right)^2\right.}$

$= \frac{1}{\sqrt{3}}$

$\text{correlation} = \frac{-1/3}{\frac{1}{\sqrt{3}} \times \frac{1}{\sqrt{3}}}$

$= \boxed{-1.}$

$\text{Euclidean} = \sqrt{(0-1)^2+(1-0)^2+(0-1)^2+(1-0)^2} = \boxed{2.}$

For jaccard,

$f_{01} = 2 \quad f_{10} = 0 \quad f_{11} = 0.$

$\text{jaccard} = \frac{f_{11}}{f_{01}+f_{10}+f_{11}} = \boxed{0.}$

$\therefore$ cosine = 0 correlation = -1

Euclidean = 2 Jaccard = 0

Q.2.19 c) $x = (0,-1,0,1)$ $y = (1,0,-1,0)$

$$\text{cosine} = \frac{0\times1 + (-1)\times0 + 0\times(-1) + 1\times0}{\sqrt{2}\times\sqrt{2}} = \boxed{0}$$

$\bar{x} = 0$ $\bar{y} = 0$

$$\text{covariance}_{xy} = \frac{1}{3}\left[0\times1 + (-1)\times0 + 0\times(-1) + 1\times0\right] = 0.$$

∵ covariance is 0, correlation = $\boxed{0}$

$$\text{Euclidean} = \sqrt{(0-1)^2 + (-1-0)^2 + (0+1)^2 + (1-0)^2} = \boxed{2}$$

cosine = 0 correlation = 0 Euclidean = 2

d) $x = (1,1,0,1,0,1)$ $y = (1,1,1,0,0,1)$

$\|x\| = \|y\| = 2.$

$$\text{cosine} = \frac{1\times1 + 1\times1 + 0\times1 + 1\times0 + 0\times0 + 1\times1}{2\times2} = \boxed{0.75}$$

$\bar{x} = \frac{2}{3}$ $\bar{y} = \frac{2}{3}$

$$\text{covariance}_{xy} = \frac{1}{5}\left[\left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right) + \left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right) + \left(0-\frac{2}{3}\right)\left(1-\frac{2}{3}\right) + \left(1-\frac{2}{3}\right)\left(0-\frac{2}{3}\right) + \left(0-\frac{2}{3}\right)\left(0-\frac{2}{3}\right) + \left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right)\right]$$

$$= \frac{1}{5}\left[3\times\left(1-\frac{2}{3}\right)\left(1-\frac{2}{3}\right) + \left(-\frac{2}{3}\right)\left(\frac{1}{3}\right) + \left(\frac{1}{3}\right)\left(-\frac{2}{3}\right) + \left(-\frac{2}{3}\right)\left(-\frac{2}{3}\right)\right]$$

$$= \frac{1}{5}\left[\frac{1}{3} - \frac{2}{9} - \frac{2}{9} + \frac{4}{9}\right]$$

$$= \frac{1}{15}$$

$$\sigma_x = \sqrt{\frac{1}{5}\left[\left(1-\frac{2}{3}\right)^2 + \left(1-\frac{2}{3}\right)^2 + \left(0-\frac{2}{3}\right)^2 + \left(1-\frac{2}{3}\right)^2 + \left(0-\frac{2}{3}\right)^2 + \left(1-\frac{2}{3}\right)^2\right]}$$

$$= \sqrt{\frac{1}{5}\times\frac{12}{9}} = \sqrt{\frac{4}{15}} = \frac{2}{\sqrt{15}}$$

Q.2.19 d) $\sigma_y = \sqrt{\frac{1}{5}\left[\left(1-\frac{2}{3}\right)^2+\left(1-\frac{2}{3}\right)^2+\left(1-\frac{2}{3}\right)^2+\left(0-\frac{2}{3}\right)^2+\left(0-\frac{2}{3}\right)^2+\left(1-\frac{2}{3}\right)^2\right]}$

$= \frac{2}{\sqrt{15}}$

$\text{correlation} = \frac{1/15}{\frac{2}{\sqrt{15}} \times \frac{2}{\sqrt{15}}}$

$= \boxed{\frac{1}{4}}$

Jaccard

$f_{11} = 3 \quad f_{01} = 1 \quad f_{10} = 1$

$\text{Jaccard} = \frac{3}{3+1+1} = \frac{3}{5} = \boxed{0.6}$

cosine = 0.75 correlation = 0.25 Jaccard = 0.6

e) $x = (2, -1, 0, 2, 0, -3) \quad y = (-1, 1, -1, 0, 0, -1)$

$\text{cosine} = \frac{2\times(-1)+(-1)(1)+0\times(-1)+2\times0+0\times0+(-3)(-1)}{\sqrt{4+1+4+9} \times \sqrt{1+1+1+1}}$

$= \boxed{0}$

$\bar{x} = 0 \quad \bar{y} = -1/3$

$\text{covariance}_{xy} = \frac{1}{5}\left[(2-0)(-1+\frac{1}{3})+(-1-0)(1+\frac{1}{3})+0+(2-0)(0+\frac{1}{3})+\right.$

$\left.0+(-3-0)\left(-1+\frac{1}{3}\right)\right]$

$= \frac{1}{5}\left[2\times\left(\frac{-2}{3}\right)+(-1)\frac{4}{3}+2\left(\frac{1}{3}\right)+2\right]$

$= \frac{1}{5}[0]$

$= 0$

∵ ~~correl~~ covariance is 0, correlation is 0.

∴ cosine = 0

correlation = 0.

Q.4.2 a) For overall collection:

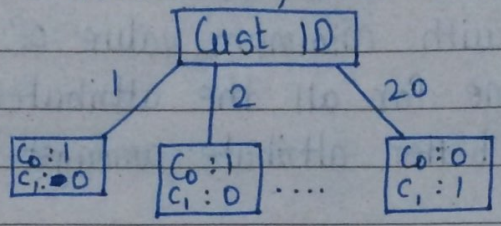
Number of samples = 20 $C_0 = 10$ $C_1 = 10$.

$$Gini(t) = 1 - \sum_{i=0}^{C-1} [P(i/t)]^2$$

$$\therefore Gini = 1 - \left(\frac{10}{20}\right)^2 - \left(\frac{10}{20}\right)^2$$

$$= \boxed{0.5}$$

b) For customer ID, tree will look like



For each leaf node, Gini = 0.

Because $Gini = 1 - (1)^2 - 0$

$$= \boxed{0}$$

$$\therefore Gini \text{ for grouping} = \frac{1}{20} \times 0 + \frac{1}{20} \times 0 + \cdots + \frac{1}{20} \times 0$$

$$= \boxed{0}$$

c) Gender :

	Male N_1	Female N_2
class = C_0	6	4
class = C_1	4	6

$$\therefore Gini(N_1) = 1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 0.48$$

$$Gini(N_2) = 1 - \left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2 = 0.48$$

$$Gini \text{ for grouping} = \frac{10}{20} \times 0.48 + \frac{10}{20} \times 0.48$$

$$= \boxed{0.48}$$

Q.4.2 d) Car type :

	Family N_1	Sports N_2	Luxury N_3
class = C_0	1	8	1
class = C_1	3	0	7

$$Gini(N_1) = 1 - \left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$$

$$Gini(N_2) = 1 - \left(\frac{8}{8}\right)^2 - 0 = 0$$

$$Gini(N_3) = 1 - \left(\frac{1}{8}\right)^2 - \left(\frac{7}{8}\right)^2 = 0.218$$

$$\text{Gini for grouping} = \frac{4}{20} \times 0.375 + \frac{8}{20} \times 0 + \frac{8}{20} \times 0.218$$

$$= \boxed{0.1622}$$

e) Shirt size :

	S N_1	M N_2	L N_3	EL N_4
class = C_0	3	3	2	2
class = C_1	2	4	2	2

$$Gini(N_1) = 1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$$

$$Gini(N_2) = 1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.4897$$

$$Gini(N_3) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$$

$$Gini(N_4) = 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$$

$$\text{Gini for grouping} = \frac{5}{20} \times 0.48 + \frac{7}{20} \times 0.4897 + \frac{4}{20} \times 0.5 + \frac{4}{20} \times 0.5$$

$$= \boxed{0.4913}$$

Q.4.2 f) Gini (Gender) = 0.48

∴ Gain (Gender) = 0.5 − 0.48 — Gain = Gini(parent) − Gini(child)

= 0.02

Gini (car type) = 0.1622

Gain (car type) = 0.5 − 0.1622

= **0.3378**

Gain (Shirt size) = 0.5 − 0.4913

= 0.0087.

Other way to compute which attribute is better is to choose attribute with minimum value of Gini index since gini of parent is same for all the attributes.

∴ Car type is a better attribute amongst three.

g) Customer ID has zero gini index.

But, it cannot be used as an attribute for splitting because it has unique value for each record.

∴ Whenever a new record from test data comes, it won't fit in the tree since its value will be different from the previous records.

∴ It is not a predictive attribute.

∴ Customer ID should not be used for splitting.

Q.4.3 a) Entropy (t) = $-\sum_{i=0}^{c-1} P(i/t) \log_2 P(i/t)$

∴ entropy with respect to positive class is,

$\frac{-4}{9} \log_2 \frac{4}{9}$ = **0.5199**

b) a_1

	T	F
	N_1	N_2
+	3	1
−	1	4

a_2

	T	F
	N_1	N_2
+	2	2
−	3	2

Q.4.3 b) (q1) Entropy $(N_1) = -\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}$

$= 0.8112$

(q1) Entropy $(N_2) = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5}$

$= 0.7219$

Entropy $(q_1) = \frac{4}{9} \times 0.8112 + \frac{5}{9} \times 0.7219$

$= \boxed{0.7615}$

(q_2) Entropy $(N_1) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$

$= 0.9709$

(q_2) Entropy $(N_2) = -\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}$

$= 1.$

Entropy $(q_2) = \frac{5}{9} \times 0.9709 + \frac{4}{9} \times 1$

$= \boxed{0.9838}$

overall entropy for collection $= -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9}$

$= \boxed{0.9910}$

Gain (q_1) (Δinf_o) = Entropy(overall) − Entropy (splitting)

$= 0.9910 - 0.7615$

$= \boxed{0.2295}$

Gain $(q_2) = 0.9910 - 0.9838$

$= \boxed{0.0072}$

Q. 4.3 c) Possible splits for a_3

	+	−	+	−1−	+	+1−	−
	1.0	3.0	4.0	5.0	6.0	7.0	8.0

	0.5		2		3.5		4.5		5.5		6.5		7.5		8.5	
	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
+	0	4	1	3	1	3	2	2	2	2	3	1	4	0	4	0
−	0	5	0	5	1	4	1	4	3	2	3	2	4	1	5	0
Entropy	0.9910		0.8483		0.9885		0.9182		0.9838		0.9727		0.8888		0.9910	
Gain	0		0.1426		0.0024		0.0727		0.0071		0.0182		0.1021		0	

For split 0.5,

$N_1 (\leq 0.5) = 0$ $\quad N_2 (> 5) = -\frac{4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9}$

$= 0.9910$

Entropy $= 0 + 1 \times 0.9910 = 0.9910$.

Gain = Total entropy (before splitting) − entrop (after splitting)

$= 0.9910 - 0.9910$

$= \boxed{0}$

For split 2,

$N_1 (\leq 2) = 0$ $\quad N_2 (> 2) = -\frac{3}{8}\log_2\frac{3}{8} - \frac{5}{8}\log_2\frac{5}{8}$

$= 0.9544$

Entropy $= \frac{1}{9} \times 0 + \frac{8}{9} \times 0.9544$

$= 0.8483$

Gain $= 0.9910 - 0.8483 = \boxed{0.1426}$

For split 3.5,

$N_1 (\leq 3.5) = -\frac{1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1.$

$N_2 (> 3.5) = -\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}$

$= 0.9852$

Q.4.3 c) Entropy $= \frac{2}{9} \times 1 + \frac{7}{9} \times 0.9852$

$= 0.9885.$

Gain = $\boxed{0.0024}$

For split 4.5,

$N_1 (\leq 4.5) = -\frac{2}{3} \log_2 \frac{2}{3} - \frac{1}{3} \log_2 \frac{1}{3}$

$= 0.9182$

$N_2 (>4.5) = -\frac{2}{6} \log_2 \frac{2}{6} - \frac{4}{6} \log_2 \frac{4}{6}$

$= 0.9182.$

Entropy $= \frac{3}{9} \times 0.9182 + \frac{6}{9} \times 0.9182$

$= \boxed{0.9182}$

Gain $= 0.9910 - 0.9182$

$= \boxed{0.0727}$

For split 5.5,

$N_1 (\leq 5.5) = -\frac{2}{5} \log_2 \frac{2}{5} - \frac{3}{5} \log_2 \frac{2}{5}$

$= 0.9709$

$N_2 (>5.5) = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{2} \log_2 \frac{1}{2}$

$= 1.$

Entropy $= \frac{5}{9} \times 0.9709 + \frac{4}{9} \times 1$

$= \boxed{0.9838}$

Gain $= 0.9910 - 0.9838 = \boxed{0.0071}$

Q.4.3 c) For split 6.5,

$N_1(\leq 6.5) = \frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$N_2(>6.5) = \frac{-1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3} = 0.9182$

Entropy $= \frac{6}{9} \times 1 + \frac{3}{9} \times 0.9182 = \boxed{0.9727}$

Gain $= 0.9910 - 0.9727 = \boxed{0.0182}$

For split 7.5,

$N_1(\leq 7.5) = \frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2} = 1$

$N_2(>7.5) = 0.$

Entropy $= \frac{8}{9} \times 1 + 0 = \boxed{0.8888}$

Gain $= 0.9910 - 0.8888 = \boxed{0.1021}$

For split 8.5,

$N_1(\leq 8.5) = \frac{-4}{9}\log_2\frac{4}{9} - \frac{5}{9}\log_2\frac{5}{9} = 0.9910$

$N_2(>8.5) = 0.$

$\therefore$ Entropy $= 1 \times 0.9910 = 0.9910.$

Gain $= 0.$

Since, split at ≤ 2 and >2 is giving maximum gain, it is the best split.

Q.4.3 d) Gain$(q_1) = 0.2295$ Gain$(q_2) = 0.0072$ Gain$(q_3) = 0.1426$

Since gain is maximum for q_1, it is best among q_1, q_2, q_3.

Q.4.3 e)

q_1	T	F
	N_1	N_2
+	3	1
−	1	4

q_2	T	F
	N_1	N_2
+	2	2
−	3	2

Overall classification error for training examples is

Q.4.3 e) given by,

$$1 - \max\left(\frac{4}{9}, \frac{5}{9}\right) = 1 - \frac{5}{9} = \frac{4}{9} = \boxed{0.444}$$

Now,

(a1) class error $(N_1) = 1 - \max\left(\frac{3}{4}, \frac{1}{4}\right)$

$= 1/4$

(a1) class error $(N_2) = 1 - \max\left(\frac{1}{5}, \frac{4}{5}\right)$

$= 1/5$.

class error $(a_1) = \frac{4}{9} \times \frac{1}{4} + \frac{5}{9} \times \frac{1}{5} = \frac{2}{9} = \boxed{0.222}$

Similarly,

(a_2) class error $(N_1) = 1 - \max\left(\frac{2}{5}, \frac{3}{5}\right)$

$= 2/5$.

(a_2) class error $(N_2) = 1 - \max\left(\frac{2}{4}, \frac{2}{4}\right)$

$= 1/2$.

$\therefore$ class error $(a_2) = \frac{5}{9} \times \frac{2}{5} + \frac{4}{9} \times \frac{1}{2}$

$= \boxed{0.444}$

Since attribute a_1 will give gain $= 0.444 - 0.222 = \boxed{0.222}$ unlike attribute a_2 with gain 0, a_1 is the best split among 2.

Q.4.3 F) (a_1) Gini $(N_1) = 1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$

(a_1) Gini $(N_2) = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.32$

Q.4.3 f) ∴ Gini (q_1) grouping $= \frac{4}{9} \times 0.375 + \frac{5}{9} \times 0.32$

$= \boxed{0.3444}$

(q_2) Gini (N_1) $= 1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2 = 0.48$

(q_2) Gini (N_2) $= 1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$

Gini (q_2) grouping $= \frac{5}{9} \times 0.48 + \frac{4}{9} \times 0.5$

$= \boxed{0.488}$

overall gini for training samples $= 1 - \left(\frac{4}{9}\right)^2 - \left(\frac{5}{9}\right)^2$

$= \boxed{0.493}$

Since q_1 has lower ~~gini~~ Gini index, it will give maximum gain.
∴ q_1 is the best split among q_1 and q_2.

Q.4.8 g) A is 0 and B is 0 → class is +.
No. of errors for this case = 0.

A is 0 and B is 1 → class is −.
No. of errors for this case = 1 (Instance 3)

A is 1 and C is 0 → class is +.
No. of errors = 3 (Instance 7, 9, 10)

A is 1 and C is 1 → class is −.
No. of errors = 1 (Instance 8)

Q.4.8 9) Total no. of records = 10.

∴ Training error rate = $\frac{5}{10}$.

In optimistic error approach, generalization error is same as training error.

∴ Generalization error = **0.5**

b) Generalization error using pessimistic approach with a factor of 0.5 = $\frac{5 + 4 \times 0.5}{10}$ since number of leaf nodes is 4.

∴ Generalization error = **0.7**

c) For validation set,

A is 0 and B is 0 → class +.

No. of errors = 0.

A is 0 and B is 1 → class -.

No. of errors = 1. (Instance 12)

A is 1 and C is 0 → class +.

No. of errors = 0.

A is 1 and C is 1 → class -.

No. of errors = 0.

No. of records = 5.

∴ Generalization error using optimistic approach = $\frac{1}{5}$ = **0.2**

Generalization error using pessimistic approach,

$\frac{1 + 4 \times 0.5}{5} = \frac{3}{5}$ = **0.6**

Q.5.5 a) Total no. of positive examples = 29.
Total no. of negative examples = 21.
Total = 29+21 = 50.

For Rule R_1, $f_+ = 12$
$f_- = 3$. Total = 15.

$e_+ = 15 \times \frac{29}{50}$ $\quad e_- = 15 \times \frac{21}{50}$

$= 8.7$ $\quad = 6.3$

Likelihood Ratio Statistic is given by,

$$R(R_1) = 2 \sum_{i=1}^{K} f_i \log(f_i/e_i).$$

$$\therefore R(R_1) = 2\left[12 \log_2 \frac{12}{8.7} + 3 \log_2 \frac{3}{6.3}\right]$$

$$= \boxed{4.7123}$$

For Rule R_2, $f_+ = 7$
$f_- = 3$. Total = 10

$e_+ = 10 \times \frac{29}{50}$ $\quad e_- = 10 \times \frac{21}{50}$

$= 5.8$ $\quad = 4.2$

$$R(R_2) = 2\left[7 \log_2 \frac{7}{5.8} + 3 \log_2 \frac{3}{4.2}\right]$$

$$= \boxed{0.8856}$$

For Rule R_3, $f_+ = 8$
$f_1 = 4$

$e_+ = 12 \times \frac{29}{50}$ $\quad e_- = 12 \times \frac{21}{50}$

$= 6.96$ $\quad = 5.04$

Q.5.5 a) $R(R_3) = 2\left[8\log_2 \frac{8}{6.96} + 4\log_2 \frac{4}{5.04}\right]$

$= 0.547$

Since R_1 has the highest value for likelihood ratio statistic, it is the best rule and R_3 has lowest value, hence it is the worst rule.

b) $n = 15$ for R_1.

$f_+ = 12$.

$K = 2$ since 2 classes are there.

Laplace measure $= \frac{f_+ + 1}{n+K}$.

$\therefore$ Laplace $(R_1) = \frac{12+1}{15+2} = \boxed{0.7647}$

For R_2, $n = 10$. $f_+ = 7$.

$\therefore$ Laplace $(R_2) = \frac{7+1}{10+2} = \frac{8}{12} = \boxed{0.6666}$

For R_3, $n = 12$ $f_+ = 12$.

$\therefore$ Laplace $(R_3) = \frac{8+1}{12+2} = \boxed{0.6428}$

$\therefore$ According to Laplace measure, R_1 is the best rule and R_3 is the worst rule.

c)

m-estimate $= \frac{f_+ + KP_+}{n+K}$

$K = 2$ $P_+ = 0.58$

For R_1, $f_+ = 12$. $n = 15$

Q.5.5 c) $\therefore$ m-estimate $(R_1) = \frac{12 + 2 \times 0.58}{15+2}$

$= \boxed{0.7741}$.

For R_2, $f_+ = 7$. $n = 10$.

$\therefore$ m-estimate $(R_2) = \frac{7 + 2 \times 0.58}{10+2}$

$= \boxed{0.68}$

For R_3, $f_+ = 8$ $n = 12$.

$\therefore$ m-estimate $(R_3) = \frac{8 + 2 \times 0.58}{12+2}$

$= \boxed{0.6542}$

Since R_1 has the highest value, it is the best rule and R_3 has the lowest value, it is the worst rule.

d) If none of the examples from R_1 are discarded after it has been discovered,

For R_3,

no. of positive examples = 8.

no. of negative examples = 4.

$\therefore$ Accuracy $(R_3) = \frac{8}{12} = \boxed{0.6666}$

Since accuracy of $R_1 = \frac{12}{15}$, its value is $\boxed{0.8}$

R_2 does not have any common records with R_1.

$\therefore$ no. of positive examples = 7.

no. of negative examples = 3.

$\therefore$ Accuracy $(R_2) = \frac{7}{10} = \boxed{0.7}$

$\therefore$ Rule R_1 is the best and rule R_3 is the worst.

Q.5.5 e) If positive examples by R_1 are discarded,
for R_3, positive examples = 6.
negative examples = 4.
$\therefore$ Accuracy $(R_3) = \frac{6}{10} = \boxed{0.6}$

Accuracy for R_2 won't change.
$\therefore$ R_1 is the best rule and R_3 is the worst.

f) If both positive and negative examples by R_1 are discarded,
For R_3, positive examples = 6
negative examples = 2.
$\therefore$ Accuracy $(R_3) = \frac{6}{8} = \boxed{0.75}$

Since R_3 has better accuracy than R_2 now,
R_1 is the best rule and R_2 is the worst rule.

Q.5.17 a) For plotting ROC curve for M_1 and M_2, we first arrange the records in ascending order of their posterior probabilities.
For each record, all the records below it are considered as negative and rest are considered positive. From that, count of True Positive (TP), False Positive (FP), True Negative (TN), False Negative (FN) are calculated.
True Positive Rate (TPR) and False Positive Rate (FPR) are computed from that.
ROC curve is plotted with all the obtained points, with TPR on Y-axis and FPR on X axis.

Table below shows the calculations for TPR and FPR with the same,

$$TPR = \frac{TP}{TP+FN} \qquad FPR = \frac{FP}{FP+TN}$$

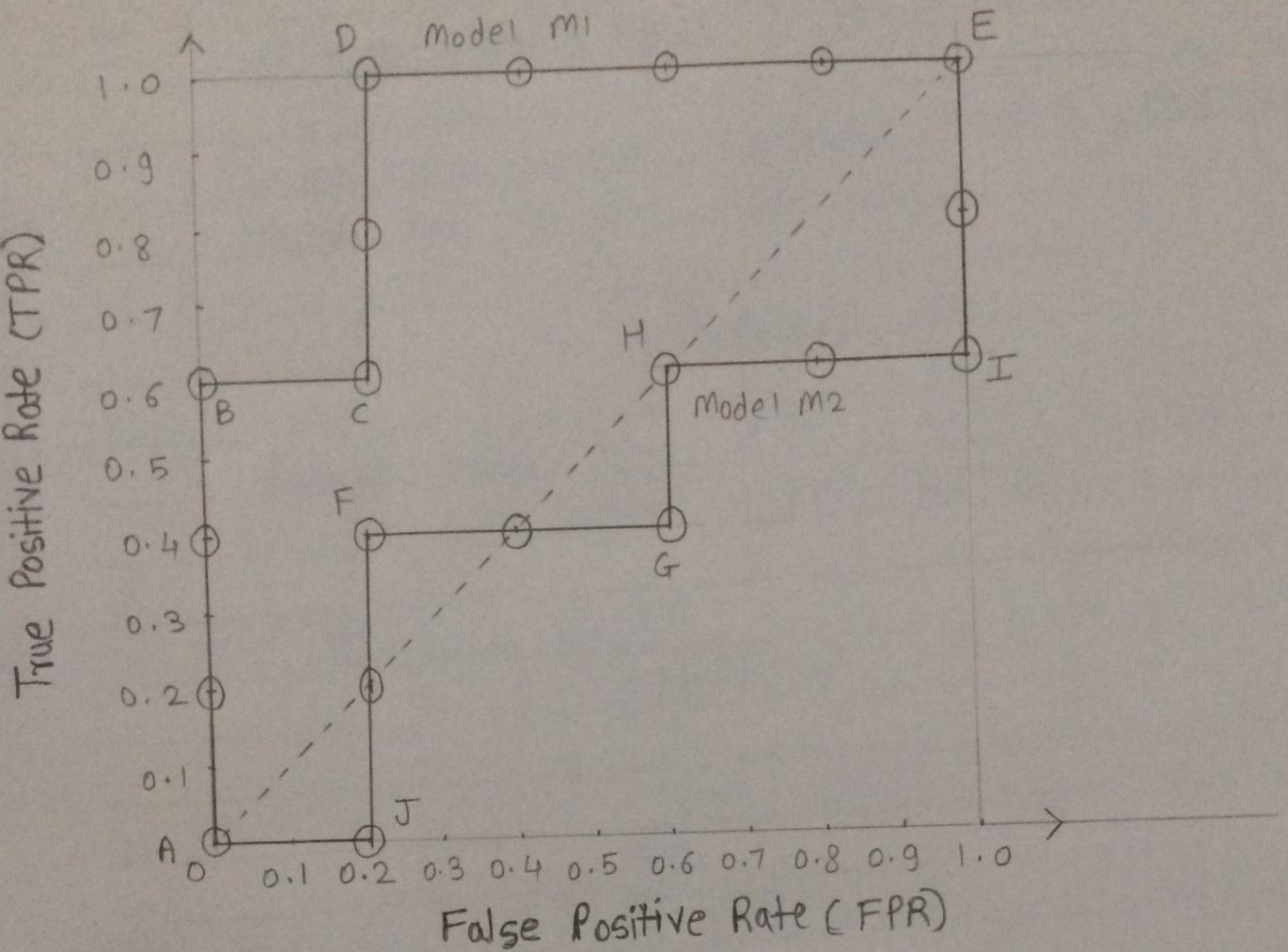
	-	-	-	-	+	+	-	+	+	+	
	0.08	0.15	0.35	0.44	0.45	0.47	0.55	0.67	0.69	0.73	1.0
TP	5	5	5	5	5	4	3	3	2	1	0
FP	5	4	3	2	1	1	1	0	0	0	0
TN	0	1	2	3	4	4	4	5	5	5	5
FN	0	0	0	0	0	1	2	2	3	4	5
TPR	1	1	1	1	1	0.8	0.6	0.6	0.4	0.2	0
FPR	1	0.8	0.6	0.4	0.2	0.2	0.2	0	0	0	0

For Model M_1

	+	+	-	-	+	-	-	+	+	-	
	0.01	0.03	0.04	0.05	0.09	0.31	0.38	0.45	0.61	0.68	1.0
TP	5	4	3	3	3	2	2	2	1	0	0
FP	5	5	5	4	3	3	2	1	1	1	0
TN	0	0	0	1	2	2	3	4	4	4	5
FN	0	1	2	2	2	3	3	3	4	5	5
TPR	1	0.8	0.6	0.6	0.6	0.4	0.4	0.4	0.2	0	0
FPR	1	1	1	0.8	0.6	0.6	0.4	0.2	0.2	0.2	0

For Model M_2

The Following graph shows ROC curves plotted for M1 and M2 on the same graph.



ROC CURVE FOR M1 and M2.

M1 : ABCDE

M2 : AJFGHIE

From the graph, it is clearly visible that almost all the curve for Model M2 lies below the random classifier AE

On the other hand, M1 is above the random classifier for all the points.

Since M1 has a better lift and is close to the upper left corner as compared to M2, M1 is a better mode

Q.5.17 b) For model M1, we arrange the records in ascending order of their posterior probabilities.

-	-	-	-	+	+	-	+	+	+
0.08	0.15	0.35	0.44	0.45	0.47	0.55	0.67	0.69	0.73	1.0

If we take cutoff threshold $t = 0.5$, the confusion matrix for M1 will look like,

		Predicted class	
		+	-
Actual class	+	3 (TP)	2 (FN)
	-	1 (FP)	4 (TN)

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

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

$$= 3/4$$

$$= \boxed{0.75}$$

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

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

$$= 3/5$$

$$= \boxed{0.6}$$

$$\text{F1-measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{recall} + \text{precision}}$$

$$= \frac{2 \times 0.75 \times 0.6}{0.75 + 0.6}$$

$$= \boxed{0.6667}$$

Q.5.17 b) ∴ Precision = 0.75
Recall = 0.6
F1-measure = 0.6667

Q.5.17 c) We will arrange the instances in ascending order of their posterior probabilities.

+	+	−	−	+	−	−	+	+	−
0.01	0.03	0.04	0.05	0.09	0.31	0.38	0.45	0.61	0.68	1

If we take cutoff threshold $t = 0.5$, confusion matrix will look like,

		Predicted Class	
		+	−
Actual class	+	1 (TP)	4 (FN)
	−	1 (FP)	4 (TN)

$$\text{Precision} = \frac{1}{1+1} = \boxed{0.5}$$

$$\text{Recall} = \frac{1}{1+4} = \boxed{0.2}$$

$$\text{F1-measure} = \frac{2 \times 0.5 \times 0.2}{0.2 + 0.5} = \boxed{0.2857}$$

$$M_1 \text{ (F1-measure)} = \boxed{0.6667}$$

∴ F1-measure of model M1 is much higher than that of M2.
Hence, M1 is definitely a better model than M2.
Results are consistent with ROC curve since we got the same result from it.

Q.5.17 d) For cutoff threshold $t = 0.1$, confusion matrix for M1 will look like,

		Predicted Class	
		+	−
Actual class	+	5 (TP)	0 (FN)
	−	4 (FP)	1 (TN)

$$\text{Precision} = \frac{5}{9} = \boxed{0.5556}$$

$$\text{Recall} = \frac{5}{5+0} = 1$$

$$\text{F1-measure} = \frac{2 \times 0.5556 \times 1}{0.5556 + 1}$$

$$= 0.7143$$

For $t = 0.1$, it is classifying almost all records as positive. Though, TP count is increasing, FP count is also increasing.
Though F1-measure for 0.1 is slightly better than 0.5, FPR is quite high as compared to 0.5.
For 0.5,

$$TPR = \frac{3}{3+2} = 0.6 \qquad FPR = \frac{1}{1+4} = 0.2$$

For $t = 0.1$,

$$TPR = \frac{5}{5+0} = 1 \qquad FPR = \frac{4}{4+1} = 0.8$$

$\therefore (TPR, FPR) = (0.6, 0.2)$ for 0.5
$(TPR, FPR) = (1, 0.8)$ for 0.1
Also, $(0.6, 0.2)$ has a better lift than $(1, 0.8)$.
Since both the points lie on the ROC curve for M1, results are consistent with what we expect from ROC curve.
$\therefore$ 0.5 threshold is preferrable over 0.1.

Q. Additional exercise :

From the confusion matrix,

TP = 350

FP = 344

FN = 122

TN = 670.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FN+FP}$$

$$= \frac{350+670}{1486}$$

$$= \boxed{0.6864}$$

Error rate = 1 − Accuracy

= 1 − 0.6864

= $\boxed{0.3135}$

$$\text{True Positive Rate (TPR)} = \frac{TP}{TP+FN}$$

$$= \frac{350}{350+122}$$

$$= \boxed{0.7415}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP+TN}$$

$$= \frac{344}{344+670}$$

$$= \boxed{0.3392}$$

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

$$= \frac{350}{350+344}$$

$$= \boxed{0.5043}$$

$$\text{F-measure} = \frac{2 \times \text{Recall (TPR)} \times \text{Precision}}{\text{Recall (TPR)} + \text{precision}}$$

$$= \frac{2 \times 0.7415 \times 0.5043}{0.7415 + 0.5043}$$

$$= \boxed{0.6003}$$

Accuracy = 0.6864
Error Rate = 0.3135
TPR = 0.7415
FPR = 0.3392
Precision = 0.5043
F-measure = 0.6003