EE871

OFFLINE ENDSEM ASSIGNMENT

SUBMITTED BY :

MEGH MANOJ BHALERAO

16EE234

ACADEMIC SESSION January-May 2020

End Semester Examination (Partial)

Course Code	: EE871	Course Title	: **Machine Learning**
Credits	: 3-1-2 : 5	Instructor	: Dr. Jora M. Gonda
Weight	: 20%	Marks	: $5 \times 7 = 35$

1. Die name (H_M) and $\Pr(H_M \mid \underline{x})$; $H_M \in \mathscr{H} = \{H_4, H_6, H_8, H_{12}, H_{20}\}$ and $\underline{x} = [4, 2, 4, 7, 5]$

Die name (H_M)	$\Pr(H_M \mid \underline{x})$
H8	0.878

2. $\Pr(\text{like}|\underline{x})$, $\underline{x}$ = [round, thick, grey, medi, dark]= 0.25

3. The filled Table is:

	Predicted		
Actual	0	1	Total
0	217	46	263
1	26	226	252
Total	243	272	515

i) Sensitivity	ii) F1-Score
0.8968	0.8625

4. List of indices used for deciding optimal threshold in classification:

i) From ROC

ii) Geometric mean of sensitivity and specificity

iii) Youden's J-statistic

iv) From precision v/s recall curve

v) From F-score or F1-score

5. The confusion matrix is:

$p \geq 0.7$	Predicted		
Actual	RED	BLUE	Total
RED	253	10	263
BLUE	95	157	252
Total	348	167	515

6. The Entropies are:

Outlook	Temperature	Humidity	Wind
0.8364	0.9111	0.78845	0.9242

and the root node is: Humidity

7. The results are:

- The null Hypothesis (H_0) is : Mean Burning rate is 50 cm/s
- The Alternative Hypothesis (H_01) is : Mean Burning rate differs from 50 cm/s
- The test-statistic is (formula) :(sample mean-population mean)/(standard deviation/sqrt(sample size))
- Value of the test-statistic is : 3.25
- Threshold on Test statistic is : greater than 1.96 and less than -1.96
- Conclusion about the Hypothesis is : Reject the null hypothesis and the mean burning rate differs from 50 cm/s. There is strong evidence to suggest that mean burning rate is greater than 50 cm/s

D) Since the probability of the 8-faced die is the highest after the last draw = 0.8776, hence it is the die which was chosen.

	D = 4				
Num faces	Hypothesis	Prior (P(H))	P(D|H)	P(H) * P(D|H)	Posterior (P(H|D))
4	H4	0.2	0.25	0.05	0.355029585798816
6	H6	0.25	0.166666666666667	0.041666666666667	0.295857988165681
8	H8	0.2	0.125	0.025	0.177514792899408
12	H12	0.2	0.083333333333333	0.016666666666667	0.118343195266272
20	H20	0.15	0.05	0.0075	0.053254437869823
				0.140833333333333	

	D = 2				
Num faces	Hypothesis	Prior (P(H))	P(D|H)	P(H) * P(D|H)	Posterior (P(H|D))
4	H4	0.355029586	0.25	0.088757396449704	0.513698630136987
6	H6	0.295857988	0.166666666666667	0.04930966469428	0.285388127853881
8	H8	0.177514793	0.125	0.022189349112426	0.128424657534246
12	H12	0.118343195	0.083333333333333	0.009861932938856	0.057077625570776
20	H20	0.053254438	0.05	0.002662721893491	0.01541095890411
				0.172781065088757	

		D = 4				
28	Num faces	Hypothesis	Prior (P(H))	P(D|H)	P(H) * P(D|H)	Posterior (P(H|D))
29	4	H4	0.51369863	0.25	0.128424657534247	0.650022871175096
30	6	H6	0.285388128	0.166666666666667	0.047564687975647	0.240749211546332
31	8	H8	0.128424658	0.125	0.016053082191781	0.081252858896887
32	12	H12	0.057077626	0.083333333333333	0.004756468797565	0.024074921154633
33	20	H20	0.015410959	0.05	0.000770547945205	0.003900137227051
34					0.197569444444444	

		D = 7				
39	Num faces	Hypothesis	Prior (P(H))	P(D|H)	P(H) * P(D|H)	Posterior (P(H|D))
40	4	H4	0.650022871	0	0	0
41	6	H6	0.240749212	0	0	0
42	8	H8	0.081252859	0.125	0.010156607362111	0.821874441937107
43	12	H12	0.024074921	0.083333333333333	0.002006243429553	0.1623455687777
44	20	H20	0.003900137	0.05	0.000195006861353	0.015779989285193
45					0.012357857653016	

		D = 5				
48	Num faces	Hypothesis	Prior (P(H))	P(D|H)	P(H) * P(D|H)	Posterior (P(H|D))
49	4	H4	0	0	0	0
50	6	H6	0	0.166666666666667	0	0
51	8	H8	0.821874442	0.125	0.102734305242138	0.877680138972523
52	12	H12	0.162345569	0.083333333333333	0.013528797398142	0.115579277560168
53	20	H20	0.015779989	0.05	0.00078899946426	0.006740583467309
54					0.11705210210454	

2) $\underline{x}$ = [round, thick, grey, medium, dark]

Determine the probability $\Rightarrow$ $Pr(like \mid \underline{x})$

$Pr(like \mid \underline{x})$

$$Pr(x_i \mid like) = P(x_1 = round \mid like) = \frac{2}{8} = \frac{1}{4}$$

$$P(x_2 = thick \mid like) = \frac{4}{8} = \frac{1}{2}$$

$$P(x_3 = grey \mid like) = \frac{4}{8} = \frac{1}{2}$$

$$P(X_4 = medi \mid like) = \frac{2}{8} = \frac{1}{4}$$

$$P(x_5 = dark \mid like) = \frac{1}{8}$$

$$P(x \mid like) = \prod_{i=1}^{5} P(x_i \mid like) = \frac{1}{512}$$

$$P(\text{like}|x) = \frac{P(x|\text{like}) \times P(\text{like})}{P(x)}$$

~~P(x) = P(x = like) + P(x =~~

$$P(x) = P(x|\text{like}) \cdot P(\text{like}) + P(x|\text{dislike}) \cdot P(\text{dislike})$$

$$P(x|\text{dislike}) \Rightarrow \prod_{i=1}^{5} P(x_i|\text{dislike})$$

$$P(x_1 = \text{round}|\text{dislike}) = 1/4$$

$$P(x_2 = \text{thick}|\text{dislike}) = \frac{1}{4}$$

$$P(x_3 = \text{grey}|\text{dislike}) = \frac{2}{4} = \frac{1}{2}$$

$$P(x_4 = \text{medi}|\text{dislike}) = \frac{3}{4} = \frac{3}{4}$$

$$P(x_5 = \text{dark}|\text{dislike}) = \frac{2}{4} = \frac{1}{2}$$

$$= \frac{3}{256} // = P(x|\text{dislike})$$

$$P(\text{like}) = \frac{8}{12}, \quad P(\text{dislike}) = \frac{4}{12}$$

$$\therefore P(x) = \frac{1}{512} \times \frac{8}{12} + \frac{3}{256} \times \frac{4}{12}$$

$$P(x) = \underline{\underline{\frac{1}{192}}}$$

$$P(\text{like} \mid x) = \frac{\frac{1}{512} \times \frac{8}{12}}{\frac{1}{192}} = \frac{1}{4}$$

$$H(x) = P(\text{like} \mid x) = 1/4 = \underline{\underline{0.25}}$$

$$P(\text{dislike} \mid x) = 1 - \frac{1}{4} = \underline{\underline{0.75}}$$

3)

		Predicted 0	1	
Actual	0	217	46	263
	1	26	226	252
Total		243	272	

(i) Sensitivity $= \dfrac{TP}{TP+FN} = \dfrac{(1,1)}{(1,1)+(1,0)} = 0.8968$

(ii) F1 score $= \dfrac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

Precision $= \dfrac{tp}{tp+fp} = \dfrac{(1,1)}{(1,1)+(0,1)} = 0.8308$

Recall $= \dfrac{tp}{tp+fn} \Rightarrow$ Sensitivity $=$ ~~0.8308~~ 0.8968

F1-score = 0.8625

4) Indices for optimal threshold in classification.

(i) Geometric mean of sensitivity and specificity.

Higher the value, our classifier is better.

Choose the threshold having highest GM of Sensitivity and specificity.

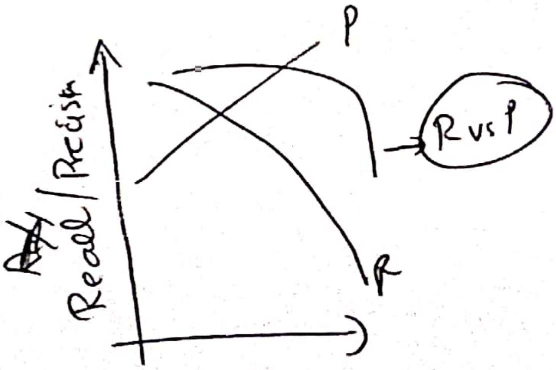
(ii) Youden Index : Youden index is the vertical distance between the 45° line and the point on the ROC curve. Formula for Youden index is ⇒ $YI = \text{sensitivity} + \text{specificity} - 1$

→ Higher values of YI better than lower values

→ choose threshold with highest Youden Index

(iii) Precision and Recall Curve: The precision recall curve shows the trade off between precision and recall for different thresholds. Higher area means higher precision and recall.

⊗ Choose the P and R value closest to (1,1) point



(iv) F-score: Measure of test's accuracy.

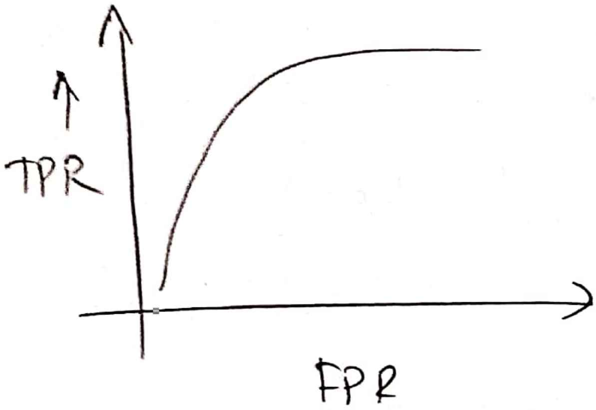
F1-Score → Harmonic mean of precision & recall

Higher F- score → better

* Choose the ~~one~~ threshold with highest F score

(v) <u>ROC Curve:</u>

* Plot of True positive rate against the False Positive rate at various threshold settings



* Choose point on ROC ~~with lowest FPR and highest TPR~~ corresponding to and which is at a closest distance to (0,1) i.e

D = FPR and 1 = TPR

5) P≥0.7

Actual \ Predicted	Red	Blue	Total
Red	253	10	263
Blue	95	157	252
Total	348	167	515

⇒ 2 PDFs' given

(6) Initial Entropy of the system :

$$S_i = -P_+ \log P_+ - P_- \log P_- \Rightarrow -\frac{9}{14}\log_2\left(\frac{9}{14}\right) - \frac{5}{14}\log_2\left(\frac{5}{14}\right)$$

$$= \underline{0.9403}$$

(1) Split based on Outlook:

(5)

Sunny	Overcast	Rainy
Neg : 3	Neg : 1	Neg : 1
Pos : 2	Pos : 3	Pos : 4
$-\frac{3}{5}\log_2\left(\frac{3}{5}\right) - \frac{2}{5}\log_2\left(\frac{2}{5}\right)$	$-\frac{1}{4}\log_2\left(\frac{1}{4}\right)$ $-\frac{3}{4}\log_2\left(\frac{3}{4}\right)$	$-\frac{1}{5}\log_2\left(\frac{1}{5}\right)$ $-\frac{4}{5}\log_2\left(\frac{4}{5}\right)$
$= 0.971$	$= 0.8113$	$= 0.7219$

$$\Rightarrow \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0.8113 + \frac{5}{14} \times 0.7219$$

$$= \underline{0.8364} = S_f$$

$$I.G = \Delta S = 0.9403 - 0.8364 = \underline{\underline{0.1039}}$$

(2) <u>Split based on</u>: Temperature.

Temperature

Hot | Mild | Cool

Hot	Mild	Cool
No : 2	No : 2	No : 1
Yes : 2	Yes : 4	Yes : 3

Entropy = $-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{2}{4}\log_2\left(\frac{2}{4}\right) = 1$

$-\frac{2}{6}\log_2\left(\frac{2}{6}\right) - \frac{4}{6}\log_2\left(\frac{4}{6}\right) = 0.9183$

$-\frac{1}{4}\log\left(\frac{1}{4}\right) - \frac{3}{4}\log\left(\frac{3}{4}\right) = 0.8113$

$$\frac{4}{14}\times 1 + \frac{6}{14}\times 0.9183 + \frac{4}{14}\times 0.8113$$

$= ~~0.9107~~\ 0.91107$

$I.G = S_i - S_f = 0.02902$

Humidity

High	Normal
Yes : 3	Yes : 6
No : 4	No : 1

$-\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log\left(\frac{4}{7}\right) = 0.9852$

$-\frac{6}{7}\log\left(\frac{6}{7}\right) - \frac{1}{7}\log\left(\frac{1}{7}\right) = 0.5917$

Total entropy is =

$$\frac{0.9852 + 0.5917}{2} = 0.78845$$

$IG = S_i - S_f = 0.15185$

Scanned with CamScanner

Wind

Weak | Strong

No : 2 | No : 3

Yes : 5 | Yes : 4

$$-\frac{2}{5}\log\left(\frac{2}{5}\right) - \frac{5}{7}\log\left(\frac{5}{7}\right) \quad \Big| \quad -\frac{3}{7}\log_2\left(\frac{3}{7}\right) - \frac{4}{7}\log_2\left(\frac{4}{7}\right)$$

$$= \underline{0.8631} \quad \Big| \quad = 0.9852$$

$$S_f = \frac{0.8631 + 0.9852}{2} = \underline{0.92415}$$

$$\Delta S = S_i - S_f = \underline{0.01615}$$

→ Hence, highest info gain is in <u>Humidity</u>

Hence, root node is $\boxed{\text{Humidity}}$

(7) $\mu \rightarrow$ mean burning rate : 50 cm/s $= H_0$

$H_1 =: \mu \neq 50$ cm/sec

$\sigma = 2$ cm/sec

$\alpha = 0.05$

The test statistic is (Normalized):

$$Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \Rightarrow \frac{51.3 - 50}{2/\sqrt{25}} = \underline{\underline{3.25}}$$

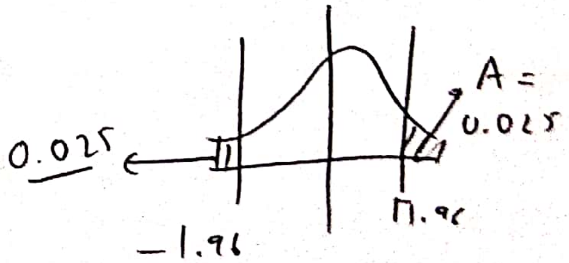
~~Reject~~

Since, $\alpha = 0.05$, $\rightarrow$ and $\frac{0.05}{2} = 0.025$ on either sides of the Normalized gaussian distribution

For $Z_0 > 1.96 \rightarrow \alpha$ (area) $= 0.025$

$Z_0 < -1.96 \rightarrow \alpha$ (area) $= 0.025$

Hence, critical region boundaries are



$Z@0.025 = 1.96$

$Z@-0.025 = -1.96$

(8) Since $Z_0 = 3.25 > 1.96$, we reject H_0 Null Hypothesis $\mu = 50$ at the 0.05 level of significance.

Stated more completely we conclude that the mean burning rate differs from 50 cm/sec, based on a sample of 25 measurements. In fact, there is strong evidence that the mean burning rate exceeds 50 cm/sec.