INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR

End-Semester Test, Autumn 2022-2023

Date of Examination: **21.11.2022**
Session: **AN**
Subject No.: **CS61061** Subject: **Data Analytics**
Department: Computer Science & Engineering
Full Marks: 100 Time: 03 hours

1. Following questions are of multiple-choice type. More than one options may be correct. Select all the correct answers. No negative marking. **[10×1 = 10]**

i. The concentration of Oxygen, in milligrams per liter air, is
   a. a nominal variable
   b. an ordinal variable
   c. an interval variable
   d. a ratio variable

ii. If the interquartile range is zero, you can conclude that:
   a. the range must also be zero
   b. the mean is also zero
   c. at least 50% of the observation have the same value
   d. all of the observations have the same value
   e. none of the above is correct.

iii. A sample of 100 scores in an examination produced the following statistics:

mean = 95	lower quartile = 70
median =100	upper quartile = 120
mode = 75	standard deviation = 30

Which of the following statement(s) is(are) correct?
   a. Half of the scores are less than 95
   b. The middle 50% of the scores are between 100 and 120
   c. One - quarter of the scores are greater than 120
   d. The most common score is 95

iv. Which of the following is not true about the probability distribution function? All symbols bear their usual meanings.
   a. $0 \le f(x) \le 1$
   b. $\int_{-\infty}^{\infty} f(x)dx = 0$
   c. $P(a \le X \le b) = \int_{a}^{b} f(x)dx$
   d. $\mu = \int_{-\infty}^{\infty} x.f(x)dx$
   e. $\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2.f(x)dx$

v. Which of the following statement(s) is(are) true?
   a. If f(x) is probability mass function, then
      $0 \le f(x) \le 1$
   b. If f(x) is a probability distribution function, then
      $f(x) \ge 0$
   c. If f(x) is probability mass function, then x is any discrete value in the range [a, b], such that $a \le b$
   d. If f(x) is a probability distribution function, then x any continuous value in the range [a, b], such that a<b.

vi. A quiz test was conducted among 10000 students in a placement drive. It was fund μ=90 and σ =20. A random sample of 100 students from the population was chosen and the mean score was found as 86. What is the standard error rate in this case? Write your final answer only.

vii. H_0: μ =250 and
H_1 : $\mu \neq 250$ is equivalent to

a. **H_0: μ = 250**
H_1 : μ< 250

b. **H_0: μ =250**
H_1 : μ > 250

c. H_0: $\mu \geq 250$
H_1 : μ < 250

d. H_0: $\mu \leq 250$
H_1 : μ > 250

e. None of the above are equivalent.

viii. In a hypothesis test the *p*-value is 0.043. This means that we can find statistical significance at

a. both the 0.05 and 0.01 levels.
b. the 0.05 but not at the 0.01 level.
c. **the 0.01 but not at the 0.05 level.**
d. neither the 0.05 or 0.01 level.
e. None of the above.

ix. If the value of any test statistics does not fall in the rejection region, the decision is:

a. Reject the null hypothesis.
b. Reject the alternative hypothesis.
c. **Fail to reject the null hypothesis.**
d. Fail to reject the alternative hypothesis.
e. There is insufficient information to make a decision.

x. If the null hypothesis is really false, which of these statements characterize a situation where the values of the test statistics does not fall in the rejection region?

a. The decision is correct.
b. A Type-I error has been committed.
c. **A Type-II error has been committed.**
d. Insufficient information has been given to make a decision.
e. None of the above is correct.

2. A study was performed to determine whether the type of cancer differed between **Old-aged (O), Middle-aged (M)** and **Children (C)** in a country. A sample of 100 of each type of population diagnosed as having cancer was categorized into one of three types of cancer. The results are shown in Table 1.

Table 1

	Lung	**Stomach**	**Kidney**
O	53	17	30
M	10	67	23
C	30	30	40

(a) State the null hypothesis that we would like to test the following χ^2 -test of the correlation analysis. **[1]**

H_0: Type of population is independent of Type of cancer

(b) Draw the contingency table with all expected frequencies. **[3]**

	Lung	Stomach	Kidney
O	31	38	31
M	31	38	31
C	31	38	31

(c) Compute the χ^2 value. **[3]**

$$\chi^2 = \sum \frac{(O-E)^2}{E} = 70.0$$

(d) Test the null hypothesis for $\alpha = 0.05$. **[3]**

The rejection region for this test is $\chi^2 > 9.488$ (degree of freedom, $df = (3-1)(3-1) = 4$) for $\alpha = 0.05$

So, reject the null hypothesis

Conclude that: The Type of Cancer is related with Age Group of the Population

(e) Find the correlation of coefficient. Hint: Use Cramer's V rule. **[3]**

$$V = \sqrt{\frac{\varphi}{min(r-1,c-1)}}, \quad \varphi = \frac{\chi^2}{n}$$

Here, $\chi^2 = 70$

$n = 300$, size of the population

$$\varphi = 0.23334$$

$$r-1 = c-1 = 2$$

$$V = \sqrt{\frac{0.23334}{2}} = \sqrt{0.11667} = 0.342$$

3. (a) Define **Sum of Squares Total (SST)** and **Sum of Squares Error (SSE)** and hence $\mathbf{R^2}$, the quality of fit to measure the goodness of fit in regression analysis. **[4]**

(b) Consider a simple dataset of size 15 is given below (Table 2).

Table 2

Sl No.	No. of days (x)	Sold cars (y)
1	168	272
2	428	300
3	296	311
4	392	365
5	80	167
6	56	149
7	352	366
8	444	310
9	168	192

10	200	229
11	4	88
12	52	118
13	20	62
14	228	319
15	72	193

For this data two linear regression models ($y = f(x)$) obtained as follows:
A simple linear regression model:

Model 1: $y = 114.4963 + 0.582282\, x$

Simple non-linear regression model with degree 2:

Model 2: $y = 63.85097 + 1.409457\, x - 0.00185\, x^2$

Calculate the R^2 values of Model 1 and Model 2 and decides which model is better than other. **[4+5]**

4. (a) When a logistic regression model is called binary logistic regression? **[1]**

(b) When a logistic regression model is called linear logistic regression with multiple explanatory variables? **[1]**

(c) Consider a binary logistic regression model with single explanatory variable. Define the following.

i. Logistic function
ii. odds
iii. logit **[6]**

(d) Draw curves for the following.

i. Logistic function with single explanatory variable.
ii. logit versus input. You may make assumption, if any. **[4]**

5. Consider a dataset which is shown in Table 3.

Table 3

Sl No.	Weather	Temperature	Humidity	Windy	Play
0	Rainy	Hot	High	No	No
1	Rainy	Hot	High	Yes	No
2	Overcast	Hot	High	No	Yes
3	Sunny	Mild	High	No	Yes
4	Sunny	Cool	Normal	No	Yes
5	Sunny	Cool	Normal	Yes	No
6	Overcast	Cool	Normal	Yes	Yes
7	Rainy	Mild	High	No	No
8	Rainy	Cool	Normal	No	Yes
9	Sunny	Mild	Normal	No	Yes
10	Rainy	Mild	Normal	Yes	Yes
11	Overcast	Mild	High	Yes	Yes
12	Overcast	Hot	Normal	No	Yes
13	Sunny	Mild	High	Yes	No

The column "**Play**" is the class label in this dataset.

(a) Obtain the contingency table containing all prior and posterior probabilities for the above dataset. **[6]**

Attributes		Play: Yes	Play: No
Outlook	Sunny	$\frac{3}{9} = 0.33$	$\frac{2}{5} = 0.4$
	Overcast	$\frac{4}{9} = 0.44$	$\frac{0}{5} = 0.0$
	Rainy	$\frac{2}{9} = 0.22$	$\frac{3}{5} = 0.6$

Temparature	Hot	$\frac{2}{9} = 0.22$	$\frac{2}{5} = 0.4$
	Mild	$\frac{4}{9} = 0.44$	$\frac{2}{5} = 0.4$
	Cool	$\frac{3}{9} = 0.33$	$\frac{1}{5} = 0.2$
Humidity	High	$\frac{3}{9} = 0.33$	$\frac{4}{5} = 0.8$
	Normal	$\frac{6}{9} = 0.67$	$\frac{1}{5} = 0.2$
Windy	No	$\frac{6}{9} = 0.67$	$\frac{2}{5} = 0.4$
	Yes	$\frac{3}{9} = 0.33$	$\frac{3}{5} = 0.6$
Class Probability		$\frac{9}{14} = 0.64$	$\frac{5}{14} = 0.36$

(b) An unseen test data "**today**" is given as follows.

today =	Sunny	Hot	Normal	No

Compute **[4]**

i. P(Yes | today)

ii. P(No | today)

$$P(Yes \mid today) = \frac{P(Outlook = Sunny \mid Yes)P(Temparature = Hot \mid Yes)P(Humidity = Normal \mid Yes)P(Windy =}{P(today)}$$

$$P(No \mid today) = \frac{P(Outlook = Sunny \mid No)P(Temparature = Hot \mid No)P(Humidity = Normal \mid No)P(Windy = N}{P(today)}$$

Since, $P(today)$ is common in both probabilities, we can ignore $P(today)$ and find proportional probabilities as:

$$P(Yes \mid today) \propto \frac{3}{9} \cdot \frac{2}{9} \cdot \frac{6}{9} \cdot \frac{6}{9} \cdot \frac{9}{14} \approx 0.0329$$

$$P(No \mid today) \propto \frac{2}{5} \cdot \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{2}{5} \cdot \frac{5}{14} \approx 0.0128$$

Now, since

$$P(Yes \mid today) + P(No \mid today) = 1$$

These numbers can be converted into a probability by making the sum equal to 1 (normalization):

$$P(Yes \mid today) = \frac{0.0329}{0.0329 + 0.0128} = 0.822$$

$$P(No \mid today) = \frac{0.0128}{0.0329 + 0.0128} = 0.178$$

Since,

$$P(Yes \mid today) > P(No \mid today)$$

So, prediction that golf would be played is 'Yes'

(c) Explain M-estimate approach and compute all the calculations in **Q. 5(b)** with M-estimate. You may make reasonable assumption, if any.

[3]

$$P(X_j = x_j \mid C_i) = \frac{n_{C_i} + m \cdot p}{n + m}$$

Here, $n =$ total number of instances of class C_i

$n_{C_i} =$ number of training examples from class C_i that take the value $X_j = x_j$

$m = 1$, equivalent sample size (constant)

$p = \frac{1}{k}$, assuming uniform prior probability estimate

where k is the number of values the attribute X_j can take

Attributes		Play: Yes	Play: No
Outlook	Sunny	0.33	0.39
	Overcast	0.43	0.06
	Rainy	0.23	0.56
Temparature	Hot	0.23	0.39
	Mild	0.43	0.39
	Cool	0.33	0.22
Humidity	High	0.35	0.75
	Normal	0.65	0.25
Windy	No	0.65	0.42
	Yes	0.35	0.58
Class Probability		0.64	0.36

$P(Yes \mid today)$

$= \frac{P(Outlook = Sunny \mid Yes)P(Temparature = Hot \mid Yes)P(Humidity = Normal \mid Yes)P(Windy =}{P(today)}$

$P(No \mid today)$

$= \frac{P(Outlook = Sunny \mid No)P(Temparature = Hot \mid No)P(Humidity = Normal \mid No)P(Windy = N}{P(today)}$

Since, $P(today)$ is common in both probabilities, we can ignore $P(today)$ and find proportional probabilities as:

$$P(Yes \mid today) \propto 0.33 \cdot 0.23 \cdot 0.65 \cdot 0.65 \cdot \frac{9}{14} \approx 0.0329$$

$$P(No \mid today) \propto 0.39 \cdot 0.39 \cdot 0.25 \cdot 0.42 \cdot \frac{5}{14} \approx 0.0158$$

Now, since

$$P(Yes \mid today) + P(No \mid today) = 1$$

These numbers can be converted into a probability by making the sum equal to 1 (normalization):

$$P(Yes \mid today) = \frac{0.0329}{0.0329 + 0.0158} = 0.79$$

$$P(No \mid today) = \frac{0.0158}{0.0329 + 0.0158} = 0.21$$

Since,

$$P(Yes \mid today) > P(No \mid today)$$

So, prediction that golf would be played is 'Yes'

6. (a) Express the matrix representation of a hyperplane in *n*-dimensional Euclidean space.

$$W^T.X + b = 0$$

$$Where, \quad W = [w_1, w_2, w_3, \ldots., w_n]$$

$$X = [x_1, x_2, x_3, \ldots., x_n]$$

[2]

(b) Write down the problem of finding the maximum margin hyperplane as an optimization problem.

$$Minimize \ \frac{||W||}{2}$$

$$Subject\ to\ \ y_i\,(w.x_i + b) \geq 1\ for\ i = 1,2,\ldots.,N$$

$$Where\ (x_i, y_i), i = 1,2,\ldots N\ are\ the\ i'th\ data$$

[3]

(c) Write the Lagrangian (L) and KKT constraints to solve the optimization problem as stated in **Q. 6(b)** using Lagrangian multiplier method.

$$Lagrangian$$

$$L = \frac{||W||}{2} - \sum_{i=1}^{N} \lambda_i\,(y_i\,(w.x_i + b) - 1)$$

$$KKT\ constraints\ are$$

$$\frac{\delta L}{\delta w} = 0$$

$$\frac{\delta L}{\delta b} = 0$$

$$\lambda_i \geq 0, i = 1,2,\ldots.,N$$

$$\lambda_i\,[y_i\,(w.x_i + b) - 1] = 0$$

$$y_i\,(w.x_i + b) - 1 \geq 1$$

[2+4]

(d) Once all unknowns are solved, what should be the form of the classifier?

$$\delta(x) = \sum_{i=1}^{N} (\lambda_i . y_i \; x . x_i + b)$$

[2]

7. Consider a dataset D which is shown in Table 4.

Table 4

Day	**Outlook**	**Temp.**	**Humidity**	**Wind**	**Decision**
1	Sunny	85	85	Weak	No
2	Sunny	80	90	Strong	No
3	Overcast	83	78	Weak	Yes
4	Rain	70	96	Weak	Yes
5	Rain	68	80	Weak	Yes
6	Rain	65	70	Strong	No
7	Overcast	64	65	Strong	Yes
8	Sunny	72	95	Weak	No
9	Sunny	69	70	Weak	Yes
10	Rain	75	80	Weak	Yes
11	Sunny	75	70	Strong	Yes
12	Overcast	72	90	Strong	Yes
13	Overcast	81	75	Weak	Yes
14	Rain	71	80	Strong	No

(a) Calculate the entropy **E(D)** of the data *D* in Table 4.

Entropy(Decision) = ∑ — p(I) . log p(I) = — p(Yes) . log p(Yes) — p(No) . log2(No) = — (9/14) . log(9/14) — (5/14) . log(5/14) = 0.4098 + 0.531 = 0.940

[3]

(b) For the attribute "**Wind**" obtain the following.

i. Weighted average entropy, $\mathbf{E_{Wind}}$ **(D)**

Sol:

Wind is a nominal attribute. Its possible values are weak and strong.

Entropy(Decision|Wind=Weak) = – p(No) . log2p(No) – p(Yes) . log2p(Yes) = – (2/8) . log2(2/8) – (6/8) . log2(6/8) = 0.811

Entropy(Decision|Wind=Strong) = – (3/6) . log2(3/6) – (3/6) . log2(3/6) = 1

Ewind(D) = (8/14) * 0.811 + (6/14)* 1 = 0.8919

ii. Information gain, α **(Wind, D)**

Sol:

Gain(Decision, Wind) = Entropy(Decision) – ∑ (p(Decision|Wind) . Entropy(Decision|Wind))

Gain(Decision, Wind) = Entropy(Decision) – [p(Decision|Wind=Weak) . Entropy(Decision|Wind=Weak)] + [p(Decision|Wind=Strong) .

[10]

Entropy(Decision|Wind=Strong)] Gain(Decision, Wind) = 0.940 – (8/14).(0.811) – (6/14).(1) = 0.940 – 0.463 – 0.428 = 0.049

iii. Split information, $E^{*}_{\textbf{Wind}}$ **(D)**

Sol: `SplitInfo(A) = −∑ |Dj|/|D| x log|Dj|/|D|` There are 8 decisions for weak wind, and 6 decisions for strong wind. SplitInfo(Decision, Wind) = -(8/14).log2(8/14) – (6/14).log2(6/14) = 0.461 + 0.524 = 0.985

iv. Gain ratio β **(Wind, D)**

Sol: GainRatio(A) = Gain(A) / SplitInfo(A) GainRatio(Decision, Wind) = Gain(Decision, Wind) / SplitInfo(Decision, Wind) = 0.049 / 0.985 = 0.049

8. A prediction system identifies 150 out of 1000 patients to have a disease. When tested with gold standard diagnostic test (it reveals ground truth), 200 patients test positive including 100 of those identified by the prediction system.

(a) Obtain the confusion matrix representing the observations in the prediction system.

100 (++)	100 (+-)
50 (-+)	750 (--)

[5]

(b) Calculate error rate of the prediction system.

Error rate = number of incorrect prediction(FP+FN)/total number of a dataset(P+N) = 150/1000 = 0.15 Mean Error rate = 0.15 x N = 0.15 x 1000 = 150 std error rate = root(e(1-e)/N) = 0.01129

[2]

(c) Calculate the following:

i. Precision
ii. Recall
iii. Sensitivity
iv. Specificity

[6]

i	Precision $=\frac{TP}{TP+FP}$	$\frac{100}{150} = 66\%$

ii	Recall $= \frac{TP}{TP+FN}$	$\frac{100}{200} = 50\%$
iii	Sensitivity $= \frac{TP}{TP+FN}$	$\frac{100}{200} = 50\%$
iv	Specificity $= \frac{TN}{TN+FP}$	$\frac{750}{800} = 93.75\%$

----*----