INDIAN INSTITUTE OF TECHNOLOGY, KHARAGPUR

End-Autumn Semester 2018-19

Date of Examination: **26.11.2018** Session: **AN**
Subject No. : **CS40003** Subject: **Data Analytics**
Department: Computer Science & Engineering

Full Marks: 100 Time: 03 hours

Answer to ALL questions

1. Answer briefly to the following questions. **[10×2 = 20]**

(a) An observation is recorded with a sample of size 500, which is shown in the following table.

Table Q1(a)

Abortion Law	Male	Female	Transgender
For	82	70	62
Against	92	62	67
Neutral	25	19	21

What sort of correlation analysis is suitable with the above-mentioned data What other data/ table you should need to do the same?

(b) Let the random variable *x* represents the number of defective parts for a machine when three parts are sampled from a production line and tested. The following is the probability distribution of *x*.

Table Q1(b)

x	0	1	2	3
f(x)	0.51	0.38	0.10	0.01

Find the **coefficient of variance** from the above data.

(c) Can χ^2 analysis be applied to ordinal or numerical data? Justify your answer.

(d) Consider the following sample data.

Table Q1(d)

Age	43	21	25	42	57	59
Glucose level	99	65	79	75	87	81

Can you guess the glucose level of a person whose age is 30? What is your idea about guessing the age if glucose level is 120?

(e) Calculate the **coefficient of determination** for the data in **Table Q1(d)**.

(f) In regression analysis, we calculate $\boldsymbol{R^2}$. What this $\boldsymbol{R^2}$ is called? What is the range of values of $\boldsymbol{R^2}$.

(g) A company packages salted peanuts in 200 gm packets using machine. A sample of 16 packets is taken from the production line at random time intervals and their contents weighted. The mean weight of peanuts in the 16 packets is found as 199.5.

For a statistical inference based on the above mentioned statement, what sample distribution

you should choose? Explain your answer. Clearly mention assumption, if any.

(h) For the hypothesis testing with reference to the statement in **Q1(g)**, state the null hypothesis H_0 and alternate hypothesis H_1.

(i) In a hypothesis testing, the following hypotheses are assumed.

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

How your analysis will be if you consider two other cases of H_1 as follows.

Case 1 $H_1 : \mu < \mu_0$
Case 2 $H_1 : \mu > \mu_0$

Assume all other specifications remain same.

(j) If the value of α (the confidence level) decreases, the rejection region decreases and there is a chance that H_0 is acceptable. How you can decide the limiting value of α so that the null hypothesis H_0 will be rejected always.

2. **Table Q2** contains a training data set. Answer to the following questions with reference to the data in **Table Q2**. The last column indicates the class label.

Table Q2

Income	**Age**	**Education**	**Marital Status**	**Usage**
Low	Old	University	Married	Low
Medium	Young	College	Single	Medium
Low	Aged	University	Married	Low
High	Young	University	Single	High
Low	Old	University	Married	Low
High	Young	College	Single	Medium
Medium	Aged	College	Married	Medium
Medium	Old	High School	Single	Low
High	Aged	University	Single	High
Low	Old	High School	Married	Low
Medium	Young	College	Married	Medium
Medium	Aged	High School	Single	Low
High	Aged	University	Single	High
Low	Old	High School	Married	Low
Medium	Young	College	Married	Medium

(a) Calculate the entropy of the data in **Table Q2**. What is the highest possible value of entropy of the data? In what situation, such a value is possible**?** **[2+1+1]**

(b) Split the table at the first level with reference to the attribute "Age" using each of the following metrics for splitting criteria.

i. Information gain
ii. Gain ratio
iii. Gini index

[3×3]

(c) Compare the three decision tree induction algorithms, namely ID3, C4.5 and CART with respect to the following.

i. Quality of tree
ii. Training time

[3×1]

3. Consider the data given in the **Table Q3** as the training set. Answer to the following questions with reference to the above-mentioned data.

Table Q3

Service	Education	Marrital-status	Occupation	Sex	Skill
Government	Bachelor	Unmarried	Clerical	Male	Efficient
Business	Bachelor	Married	Managerial	Male	Inefficient
Private	HS-grade	Divorced	Managerial	Male	Efficient
Government	HS-grade	Married	Clerical	Male	Normal
Business	Bachelor	Married	Managerial	Female	Inefficient
Private	Master	Married	Managerial	Female	Normal
Business	HS-grade	Widowed	Clerical	Female	Efficient
Business	Bachelor	Married	Managerial	Male	Inefficient
Private	Master	Unmarried	Managerial	Female	Efficient
Government	Bachelor	Married	Clerical	Male	Normal

(a) Calculate all the prior and posterior probabilities, which are suitable to predict a class according to the Bayesian model. You are advised to put your results in the form of a contingency table, which can be obtained after scanning the data in **Table Q3**. **[6]**

(b) Consider a test data which is given below.

Private	HS-grade	Married	Manager	Male	?

i. What is the prediction value that the test data belongs to class '**Efficient**'.

ii. For which class, the Bayes' classifier predicts the highest score?

[1+3]

(c) When the Bayes' classifier is called Naïve Bayes' classifier? **[2]**

(d) Mention at least two limitations which the Bayes' classifier suffers from. Suggest an idea to address the limitations. **[2+2]**

4. Consider the data in **Table Q4**. Answer the following questions with reference to the data in **Table Q4**.

Table Q4

Length	Width
5.763	3.312
5.554	3.333
5.291	3.337
5.324	3.379
5.658	3.562
5.386	3.312
6.191	3.561
5.998	3.484
6.154	3.930

(a) Cluster the data with **k** = 3. Show your result with first three iterations. You should produce results in the tabular forms. Clearly mention the similarity measure you have followed in your working. **[3×2]**

(b) How to measure the cluster quality with, L_1 norm, L_2 norm and Cosine Similarity as the similarity measures? Give the cluster quality at the end of third iteration, in your answer to the question at Q4(a). **[3+3]**

(c) Mention at least three situations when the **k-Mean** clustering algorithm fails to give good result. You should mention each situation clearly and explain why k-Means algorithm fails? **[4]**

5. Answer to the following questions with respect to the Support Vector Machine for classifying data.

SVM

(a) Why the main task in SVM is solving an optimization problem? State the objective function and constraints that the SVM has to solve. Clearly state all the symbols you have mentioned. **[2+1+2]**

(b) Express the primal form of the Lagrangian. Gives the KKT constraints of the problem. **[1+2]**

(c) Obtain the dual form of the Langrangian from its primal form. How the dual form of Lagrangian handles non-linearity in data? **[3+2]**

(d) How SVM classify data with multiple classes. Explain one strategy with an illustration. **[2+1]**

6. Consider the following data in **Table Q6(a),** which are defined with three types of attributes having their usual types.

Table Q6(a)

Person	Gender	Age	Salary
I	Female	Young	Low
You	Male	Old	Medium
He	Male	Aged	Medium
She	Female	Aged	High

(a) Obtain the similarity matrix of the table of data. **[3]**

(b) Consider the confusion matrix (see **Table Q6(b)**) showing the result on the performance of a classifier.

Table Q6(b)

	Class A	Class B
Class A	80	25
Class B	15	70

Calculate the following.

(i) Precision (ii) Recall (iii) Sensitivity

You should clearly mention the formula for each metric for measuring the above. **[3×2]**

(c) Plot the ROC curve and clearly show the location of (i) ideal, (ii) worst (iii) ultra-liberal (iv) ultra-conservative and (v) random classifiers in it.

State two applications od ROC curves. Explain your answer with examples. **[3+2+2]**

---- * ----