Roll Number_____

Thapar Institute of Engineering and Technology, Patiala

Department of Computer Science and Engineering

EST

B.E. (Third Year), Semester: VI **Date:** 28-05-2022 **Time: 2 hr**

Couse Name (Code): Predictive Analytics using Statistics (UCS654) **Max Marks:** 35

Course Instructor(s): Dr. Prashant Singh Rana, Dr. Suresh Chandra Raikwar

Note: Attempt all questions. Answer must be in brief and make suitable assumption (if any). No marks for incorrect/steps calculations and formula writing.

Q1(a) Consider a distribution with a parameter θ. Assume $\widehat{\theta_1}$ is an unbiased estimate of θ. Prove that $\widehat{\theta_2}$ [defined as $\widehat{\theta_2} = \widehat{\theta_1} + \delta$, where $\delta \sim N(0,1)$] will be an unbiased estimate of θ. [4]

Q1(b) Assume $X_1, X_2, X_3 \ldots\ldots\ldots X_n$ be independent and identically distributed (i.i.d.) sample from a distribution with probability defined as:
$P_{X_i}(x;\theta) = (1-\theta)^{x-1} * \theta$, where θ is the parameter of the distribution.
Find the estimator of the parameter θ using Maximum Likelihood Estimation. [3]

Q2(a) Assume $X_1, X_2, X_3 \ldots\ldots\ldots X_n$ be i.i.d. N(0,1) random variables. If $S_n = \sum_{i=1}^{n} X_i^2$ and variance$(S_n) = 2n$, then show that $\lim_{n\to\infty} P(S_n \leq a) = 0$, for any a>0, by using central limit theorem. [3]

Q2(b) The guaranteed average life of a certain type of electric light bulbs is 1000 hours with a standard deviation of 125 hours. It is proposed to sample the output so as to assure that 90% of the bulbs do not fall short of the guaranteed average by more than 2.5%. What should be the minimum size of the sample? [Note: area under normal curve from z=0 to z=1.28 is 0.4000] [4]

Q3 Find the association rules for the given below transaction dataset with following conditions. [3.5 * 2]

(a) Support = 50%, Confidence = 75%

(b) Support = 75%, Confidence = 50%

Transaction ID	Items Purchased
1	A1, A2, A3, A4, A6
2	A1, A2, A4, A7
3	A1, A5, A6, A8
4	A1, A4, A5, A7
5	A2, A4, A5
6	A5, A6, A7, A8

Q4(a) Calculate the Proximity Measures (Symmetric distance, Asymmetric distance, Coherence for given below dataset. Where X2 is a symmetric attribute: [3]

Name	X1	X2	X3	X4	X5	X6	X7
A	1	1	1	0	0	0	1
B	0	1	1	1	1	0	1
C	1	0	0	0	0	1	0

Q4(b) (i) Calculate the document-term frequency matrix for given below documents (remove the stop words). [2+2]

Doc1	Predictive Analytics used Big Data, Data Mining and Data Science Software.
Doc2	Top LinkedIn Groups for Analytics Big Data Mining and Data Science
Doc3	Introduction to Data Science a free online course on Coursera already started

(ii) Calculate the cosine similarity between above three given documents.

Q5 Considered five prediction models M1 to M5 of classification and regression. In Table-1 and Table-2, M1 to M5 are predicted values from five different models. Apply the ensemble approach to calculate the accuracy for M1, M2, M3, M4, M5 and ensemble model. Fill the accuracy table given below (right side). [2*3.5]

Note: Use voting for classification and average for regression model.

Actual	M1	M2	M3	M4	M5
0	0	0	1	1	1
0	0	0	0	0	1
0	1	0	0	1	0
1	1	1	1	0	0
0	0	0	1	0	0
1	0	1	1	1	0
1	1	1	0	1	1
1	1	0	0	1	0
1	0	1	0	1	0
1	0	1	1	0	1

Table 1: Classification Model

Actual	M1	M2	M3	M4	M5
5.13	6.84	6.29	6.76	10.08	5.38
8.83	9.18	9.44	6.22	9.6	5.07
9.46	7.61	7.86	9.05	5.46	7.3
9.35	7.83	5.89	8.64	5.65	10.89
10.15	7.04	7.4	9.43	8.06	10.12
7.36	8.15	10.09	5.7	6.66	10.66
8.39	5.46	8.52	5.77	6.72	7.7
6.48	8.69	7.77	5.83	10.41	9.29
10.87	8.64	6.09	7.77	5.19	5.52
9.64	10.96	8.13	6.27	9.99	8.96

Table 2: Regression Model

Accuracy for Classification models					
M1	M2	M3	M4	M5	Ensemble Model

Accuracy for Regression models (with ±1 error)					
M1	M2	M3	M4	M5	Ensemble Model

Roll Number_____

Thapar Institute of Engineering and Technology, Patiala

Department of Computer Science and Engineering

Quiz-II

B.E. (Third Year), Semester: VI **Date:** 28-05-2022 **Time:** 10 Min

Couse Name (Code): Predictive Analytics using Statistics (UCS654) **Max Marks:** 10

Course Instructor(s): Dr. Prashant Singh Rana, Dr. Suresh Chandra Raikwar

Note: Attempt all questions. Submit this exam before taking theory exam. No negative marking. No Correction allowed (Zero marks will be awarded for over written or modified answer(s)).

	Write your correct answer here ↓	
Q1	A random sample of size 15 has 50 as mean, the sum of squared deviations from mean is 130. Consider the following statements, related to the test to be preferred? S1: We should prefer t-test since population standard deviation is not known. S2: We should prefer t-test since sample size is small. Which of the following statement(s) is(are) true? A. S1 is true and S2 is false B. Both S1 and S2 are true C. S1 is false and S2 is true D. Both S1 and S2 are false	
Q2	The test statistic for a two-sided significance test for a population mean is z = 2.12. The corresponding P-value is: A) pnorm(2.12) B) pnorm(-2.12) C) 2*pnorm(2.12) D) 2*pnorm(-2.12)	
Q3	It is believed that 15% of all students taking a particular course receive "A" grade. In a sample of 200 students, it is found that 35 made an "A". Then, the test statistic (with 4 significant digits) for testing that the true proportion is 15% is: A) Z = 0.9901 B) Z = 39.22 C) Z = 0.9311 D) Z = 39.11	
Q4	If the heights of women are normally distributed with a mean of 64 inches, which of the following is the highest? The probability of randomly choosing: A) One woman and finding that her height is between 63 and 65 inches B) 15 women and finding that their mean height is between 63 and 65 inches C) 100 women and finding that their mean height is between 63 and 65 inches D) All of these events have the same probability	
Q5	If C is the region of rejection and T is the test statistic, then which of the following is/are FALSE regarding Type II errors? A. Type II error is Pr $(T \in C \mid H_0 \; is \; true)$ B. Type II error is Pr $(T \notin C \mid H_0 \; is \; true)$ C. Type II error is Pr $(T \in C \mid H_0 \; is \; false)$ D. Type II error is Pr$(T \notin C \mid H_0 \; is \; false)$	
Q6	A test of H0: $\mu = 0$ versus Ha: $\mu > 0$ is conducted on the same population independently by two different researchers. They both use the same sample size and the same value of $\alpha = 0.05$. Which of the following will be the same for both researchers? A. The p-value of the test B. The power of the test if the true $\mu = 6$ C. The value of the test statistic D. The decision about whether or not to reject the null hypothesis	
Q7	If the third moment about the mean is zero, then the distribution is: A. Positively skewed B. Negatively skewed C. Symmetrical D. Asymmetrical	

Q8	Suppose X is normally distributed with mean 5 and standard deviation 0.4. Using the standard transformation $Z = (X - \mu)/\sigma$, we have noted that $P(X \leq a) = P(Z \leq 1.3)$. What is the value of a, correct to two decimal places? A. 5.52 B. 5.50 C. 4.50 D. 4.52	
Q9	Consider an n-bit binary number X, representing standard deviation of some observations. S1: If each observation is multiplied by 4, then the Least Significant Bit of X will be set to zero. S2: If each observation is multiplied by 4, then the Most Significant Bit of X will be set to one. Which of the following statement(s) is(are) true? A. S1 is true and S2 is false B. S1 is false and S2 is true C. Both S1 and S2 are true D. Both S1 and S2 are false	
Q10	If a random sample of size 36 is drawn from a population with a mean 63 and variance 81, then P(Xbar > 66.75) is: [Given that $P(Z < 2.5) = 0.9938$, $P(Z < 1.96) = 0.9750$] A) 0.9938 B) 0.0062 C) 0.9750 D) 0.0250	