The LNM Institute of Information Technology

Department of Computer Science & Engineering

CSE 327 Introduction to Data Science 16UC5126

Midterm Exam

NOTE: *No doubt clarifications in the exam hall. If assumptions are to be made, make your own assumptions, state it and use it. If the assumptions are relevant and it makes sense it will be considered.* **Answer in the same order as it appears in the question paper. If you change the order there will be penalty of 1 mark for each inverted pairs! Bring your own calculators. Calculators should not be shared in the examination hall.**

All the best!

Sep 27, 2018

Total Marks: 50

1. A *blind beer taste test* was organised during IPL T20 tournament held in India. Each of 32 loyal Budweiser drinkers were given two beer containers that are not labelled and asked to give their choice of which one they liked the most. One of the containers was having Budweiser and the other one was having Kingfisher. Of the 32 participants, 13 said they liked the Kingfisher better. Kingfisher said this was an impressive showing. But maybe the subjects were just clueless to find the difference between the two beers. Test whether the results observed here is are or like tossing a coin or not, by providing:
   (a) the null and alternative hypotheses (3)
   (b) the P-value (4)
   (c) the conclusion of the test (3)

   Note: Calculation might look like cumbersome but if you are smart enough you will be able to minimise it!

2. A simple random sample of 50 test cricket scores of Virat Kohli are taken from a circket database. The average cricket score in each innings by Kohli in the sample set is 60 with an SD of 12. Answer the following:
   (a) From the sample set also find in how many tests (with respect to the sample set) Virat would have scored more than 120 runs? (2)
   (b) Construct an approximate 90%-confidence and 99%-confidence intervals for the average test cricket score of Virat Kohli. (4)
   (c) In the previous question, why is the case that one interval is bigger and another interval shorter? What is the semantics behind it. Explain it briefly. (2)

3. The following stem and leaf plot gives the scores on a statistics exam taken by computer science and engineering students.

   9 | 0, 1, 4
   8 | 3, 5, 5, 7, 8
   7 | 2, 4, 4, 5, 7, 7, 8
   6 | 0, 2, 3, 4, 6, 6
   5 | 2, 5, 5, 6, 8
   4 | 3, 6

   Find the percentage of students covered in the area of one SD and two SD away from the mean and and compare with the percentage of an exact normal distribution. Find also the inter-quartile range of this distribution. (6)

4. An internet service provider (ISP) provides internet connections to 100,000 customers. Historically, its customers have sent and received an average of 200 email messages per month. The ISP suspects that email use is increasing, and wants to plan for increased demand. To determine whether its customers currently send and receive an average of more than 200 messages per month, the ISP will take a simple random sample of 25 customers and calculate the sample mean and sample standard deviation of the number of email messages those customers sent and received in the previous month.

(a) State the null and alternate hypothesis that is required to do this test (3)

(b) Which is more appropriate a two-tail test or one-tail test? Briefly explain why (3)

(c) If the sample mean is observed to be 285.9 emails per month, with a sample standard deviation of 127 emails per month find the P-value and conclude if you go with the null hypothesis or reject it (to calculate the t-score use the following portion of t-table). (4)

df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725

5. The **nominal variable** is comprised of values that can be named but not ranked or quantified. The **ordinal variable** is comprised of values that can ranked but not quantified. The **interval-ratio variable** is comprised of values that can be quantified (described by numbers). A database person wants to build a DB for LNMIIT students and wants to define the variables for Age, Number of siblings, Highest degree completed, Overall happiness, Gender and Name from the three variables defined above. Give your best choice to him. Moreover he needs advice on if Mean, Mode and Median can be applied to all these variables or not. Which will make more sense please advice him. (6)

6. Answer the following questions:

(a) *Noise in data is never interesting/desirable but outliers can be.* Explain briefly with example. (2)

(b) Find out Jaccard and Simple matching coefficients between the two vectors $(1, 1, 0, 1, 0, 1)$ and $(1, 1, 1, 0, 0, 1)$. (2)

(c) Consider the data set given below and find out the best splitting condition for the root node of a decision tree using Gini index as the measure of impurity. (6)

Record	A	B	C	Class
1	0	0	0	Yes
2	0	0	1	No
3	0	1	1	No
4	0	1	1	No
5	0	0	1	Yes
6	1	0	1	Yes
7	1	0	1	No
8	1	0	1	No
9	1	1	1	Yes
10	1	0	1	Yes