Quiz 1, ME-781, October 16, 2020

Max Marks: 50

Name: Roll No:

2x5

1. Which data scale do the following variables belong to:
    a. Name of a person
    b. Cell number of a person
    c. Marks obtained in an exam
    d. Ranking of a student in his/her class
    e. Rate of change of fever of a hospitalized patient (in deg C per minute)

2x3

2. Let set A={1,2} and set B={-1,0,1}. Then write the following sets:
    a. C = A x B
    b. D = B x A
    c. D ∩ C

3

3. Probability of heart disease (Y) is seen to be related to age (X1), gender(X2) and body mass index (X3) of a person. How would you write this relation, if Y has a quadratic relation with X (with no term of the relation having both X1 and X3 variables)

3x2

4. In a random health checkup a person is tested positive for a rare disease. Less than 1 person in million in the world have this disease. However, the medical test which was conducted is very accurate with 99% accuracy (i.e. Sensitivity = Specificity = 99%).
    a. What is the probability that the person, who has been tested positive by this medical test, has this disease.
    b. However, if the same disease is prevalent in 10% of the population then what would be the probability that the person, who has been tested positive by this medical test, has the disease.

2x3

5. Show (very briefly) whether the following are dissimilarity measures (or not) for two point (x1, y1) and (x2,y2) in a 2D space:
    a.) $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$
    b.) $\sqrt{|(x_1 - x_2).(y_1 - y_2)|}$
    c.) $\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} + \sqrt{|(x_1 - x_2).(y_1 - y_2)|}$

2

6. Let y_i' be the value predicted by a linear regression model at point x_i. If y_i is the true value and y_{mean} be the mean of values (over the entire data set of n values), then **find** $\boldsymbol{\sum_1^n (y_i - y')^2}$ if $\sum_1^n (y_{mean} - y')^2 = 45$ and$\sum_1^n (y_i - y_{mea})^2 = 49$.

2

7. Consider a dichotomous event (for eg tossing a coin etc), let its success probability be *p*= 0.3 and let x_i be 1 for success and 0 for failure. Then the random variable x_i would follow binomial distribution which will have mean equal to *p* and variance equal to *p(1-p)*. Now consider another random variable $y = \sum_1^n x_i$, then the standard deviation of y for n=40 is:

a. 2.89

b. 0.21

c. 8.4

d. 1.33

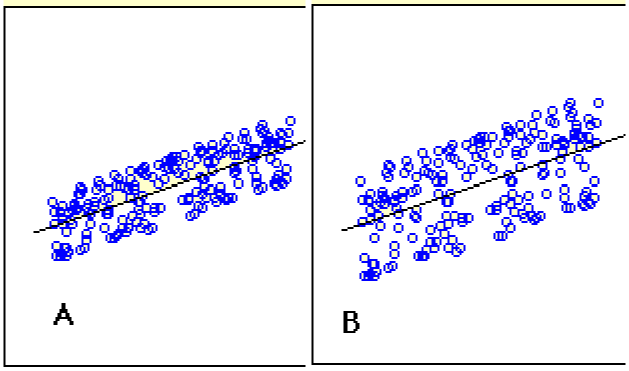
3

8. An intern in an online shopping company develops a linear regression model to predict consumer buying behavior. She takes over 50 variables into consideration while developing the model.

Based on the p-value from the t-statistic, she rejects the null hypothesis (the null Hypothesis is that the buying behavior is not dependent on any of these 50 variables). Is she correct in her analysis? Provide a brief explanation.

2x2

9. Below graphs show two fitted linear regression lines (A & B) on a randomly generated data. Note: 1) Scale is same in both graphs for both axis. 2) X axis is independent variable and Y-axis is dependent variable.



   a. What is the sum of residues of the fitting in both cases A and B
   b. Which of the two will have higher sum of square of residues

2

10. Which of the following is **always true** for Mahalanobis norm? Let *p* be the dimension of datapoints.
   a. Atleast *p* datapoints are required to ensure the existence.
   b. Mutual independence of the datapoints is a sufficient condition for the existence.
   c. Mutual independence of the datapoints is a necessary condition for the existence.
   d. More than *p* datapoints are required to ensure the existence.

2

11. Which of the following are true? Cosine similarity is always preferred over Euclidean distance similarity, if
   a. the scale of the magnitude of each datapoint is irrelevant.
   b. the datapoints are not collinear.
   c. the datapoints are collinear.
   d. the datapoints are sparsely scattered in the space. (scattering need not be random)

   Note: In all the above cases, assume that we can define both the similarity measures.

2x2

12. The relationship between number of beers consumed (x) and blood alcohol content (y) was studied in chimpanzee using least squares regression. The following regression equation was obtained from this study: ***Y = 0.0037 + 0.0180 x***
   a. Which of the following is implied by the above equation?
      i. each beer consumed increases blood alcohol by 0.37%
      ii. on average it takes 1.8 beers to increase blood alcohol content by 1%
      iii. each beer consumed increases blood alcohol by an average of amount of 1.8%
      iv. each beer consumed increases blood alcohol by exactly 0.018
   b. The equation seems to suggest that even if no beer is consumed, there is alcohol present in the blood. What could be the reason for this? (suggest a reason based on data analytics)