National University of Computer and Emerging Sciences, Lahore Campus

Course Name:	Statistical and Mathematical methods for data science	Course Code:	DS 501
Program:	MS Data Science	Semester:	Fall 2018
Duration:	60 Minutes	Total Marks:	25
Paper Date:	October 02, 2018	Weight	17.5
Section:	N/A	Page(s):	4
Exam Type:	Midterm Exam 1 Solutions		

Student : Name:_____ Roll No._____ Section:_____

QUESTION 1 (Marks: 5)

Suppose the age of patients, with diabetes, seen by a doctor has expectation $\mu = 60$ and a standard deviation $\sigma = 5$. What can you say about the probability of the age of a patient being more than 70, if you are to use Chebysehv's inequality. Show all working.

Solution

(do the working yourself)

$P(|X-60| > 10) <= 1/4$

$P(X > 70) <= 1/4$

QUESTION 2 (Marks: 4+1)

a. Compute the covariance matrix for this data. Show all working.

b. What would be the shape of the Gaussian distribution if it is fitted to this data and why? Draw the contours of this distribution.

x_1	x_2
1	0
0	0
-1	2
-1	-2

Solution

Do the working yourself:

BIASED ESTIMATE

$$\text{cov}(X) = \begin{pmatrix} 11/16 & 0 \\ 0 & 2 \end{pmatrix}$$

UNBIASED ESTIMATE

$$\text{cov}(X) = \begin{pmatrix} 11/12 & 0 \\ 0 & 8/3 \end{pmatrix}$$

The contours of the Gaussian distribution are ellipses with major axis along x_2 axis. (draw it yourself)

QUESTION 3 (Marks: 5)

Given the data below:

x_1	x_2	x_3	**Label**
1	1	1	1
0	0	0	1
0	0	0	1
0	1	1	1
0	0	1	2
1	1	1	2
1	0	1	2
1	1	0	2
1	1	1	2
0	0	0	2

Using naive Bayes' assumption determine P(Label=2|**x**=(0,0,0)). Show all working.

Solution

(do the working yourself)

use this expression to compute the denominator

P(**x**=(0,0,0)) = P(**x**=(0,0,0)|label=2)P(label=2) + P(**x**=(0,0,0)|label=1)P(label=1)

P(label=2| **x**=(0,0,0)) = 72/2160 / (72/2160+48/640) = 4/13

QUESTION 4 (Marks: 5)

Suppose the table of question 1 has only two columns, x_1 and x_2 (so ignore the x_3 and label column). Determine P(x_1=0,x_2=1) if independence assumption is **not** applied. Show all working.

Solution

$P(P(x_1=0,x_2=1)$ = total rows having (0,1) / total rows = 1/10

QUESTION 5 (Marks: 5)

Suppose the probability of getting a job after doing data science is 90%. The probability of getting a job if data science is not studied is 60%. There are overall 20% students enrolling in data science. What is the probability/chances of finding a job according to this data? Write down the stated facts, the formula you will use, and your working clearly.

Solution

P(Job) = P(Job|dataScience)*P(dataScience)+P(~Job|dataScience)*P(~dataScience)

= 0.9*0.2+0.6*0.8 = 0.66