BITS Pilani
Pilani Campus

Introduction to Statistical Methods

BITS Pilani
Pilani Campus

Course No: DSECL ZC418 / AIML CZC418
Course Title: ISM
WEBINAR 2 : 05/06/2025

Topics – Webinar

- Bayes' Theorem
- Naïve Bayes Classifier
- Random Variables

BAYES' THEOREM:

Suppose that E_1, E_2 , ---------E_n are mutually exclusive and exhaustive events of a sample space " S "

such that P (Ei) > 0 for i = 1,2,3, ---------n and A is any arbitrary event of " S "

such that P (A) > 0 and $A \subseteq \bigcup_{i=1}^{n} E_i$ then the conditional probability of E_i given A

for i = 1,2,3………n is given by $P(E_i/A) = \dfrac{P(E_i)P(A/E_i)}{\sum_{i}^{n} P(E_i)P(A/E_i)}$

Bayes' theorem is also known as the formula for the **probability of "causes"**. **As we know, the E_i's are a partition of the sample space S, and at any given time only one of the events E_i occurs.** Thus we conclude that the Bayes' theorem formula gives the probability of a particular E_i, given the event A has occurred.

Terms related to Bayes Theorem

- **Hypotheses:** Events happening in the sample space **E_1, E_2,… E_n is called the hypotheses**

- **Priori Probability:** Priori Probability is the initial probability of an event occurring before any new data is taken into account. $P(E_i)$ is the priori probability of hypothesis E_i.

- **Posterior Probability:** Posterior Probability is the updated probability of an event after considering new information. Probability $P(E_i|A)$ is considered as the posterior probability of hypothesis E_i.

Applications of Bayes Theorem

Bayesian inference is very important and has found application in various activities, including medicine, science, philosophy, engineering, sports, law, etc., and Bayesian inference is directly derived from Bayes' theorem.

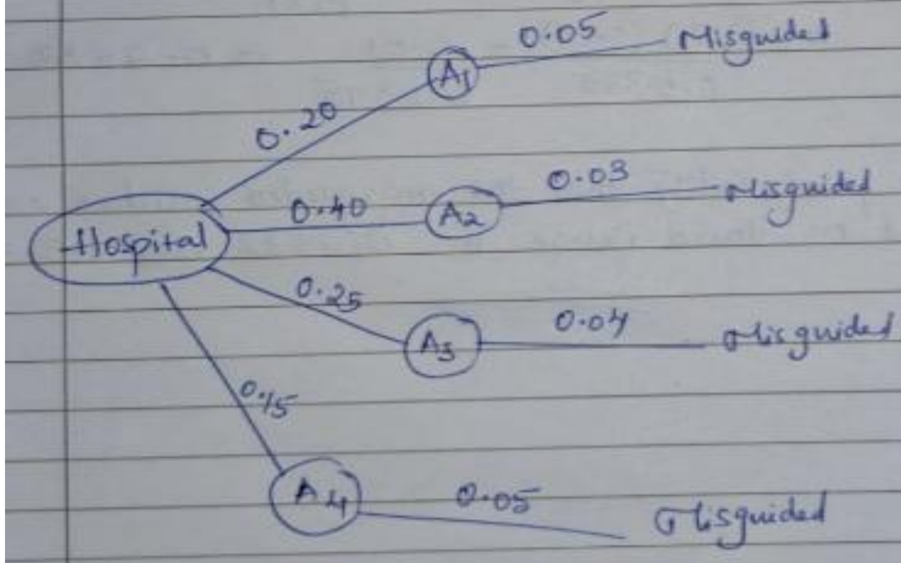
Example: Bayes' theorem defines the accuracy of the medical test by taking into account how likely a person is to have a disease and what is the overall accuracy of the test.

Problem 1

A hospital has 4 nurses handling respectively 20%, 40%, 25% and 15% of the patients of all the doctors coming to the hospital. The probability that they misguide the patients are respectively 0.05, 0.03, 0.04 and 0.05. Find the probability that the misguided incident can be blamed on the third nurse.

Problem 1 - Solution

Let A_1: Event nurse 1 handles the patient
A_2: Event nurse 2 handles the patient
A_3: Event nurse 3 handles the patient
A_4: Event nurse 4 handles the patient
X : Event that the patient is misguided



Given that,

$P(A_1) = 0.20$ $\quad P(X/A_1) = 0.05$

$P(A_2) = 0.40$ $\quad P(X/A_2) = 0.03$

$P(A_3) = 0.25$ $\quad P(X/A_3) = 0.04$

$P(A_4) = 0.15$ $\quad P(X/A_4) = 0.05$

By total probability,

P(a patient is misguided)

$$= P(X) = P(A_1 \cap X) + P(A_2 \cap X) + P(A_3 \cap X) + P(A_4 \cap X)$$

$$= P(A_1)P(X/A_1) + P(A_2)P(X/A_2) + P(A_3)P(X/A_3) + P(A_4)P(X/A_4)$$

$$= (0.20)(0.05) + (0.40)(0.03) + (0.25)(0.04) + (0.15)(0.05)$$

$$= 0.01 + 0.012 + 0.01 + 0.0075$$

$$= 0.0395$$

$$\therefore P(X) = 0.0395$$

To find the probability that the misguided incident can be blamed on third Nurse:

By applying Bayes' theorem, we have

$$P(A_3/X) = \frac{P(A_3 \cap X)}{P(X)} = \frac{P(A_3)P(X/A_3)}{P(X)}$$

$$= \frac{(0.25)(0.04)}{0.0395} = \frac{0.01}{0.0395} \simeq 0.2532$$

$\therefore$ The probability that the misguided incident can be blamed on third nurse is 0.2532 (or) 25.32%

Problem 2

Three machines A, B and C produce respectively 50%, 30% and 20% of the total number of items of a factory. The percentage of defective outputs of these machines are 3%, 5% and 2%. An item is selected at random and is found to be defective. (i) Find the probability that the item was produced by machine C? (ii) What is the probability that the item was produced by machine C or B?(iii) What is the probability that an item selected at random is found to be non-defective and also the probability that the item was produced by machine A given that it is non-defective.

Problem 2 - Solution

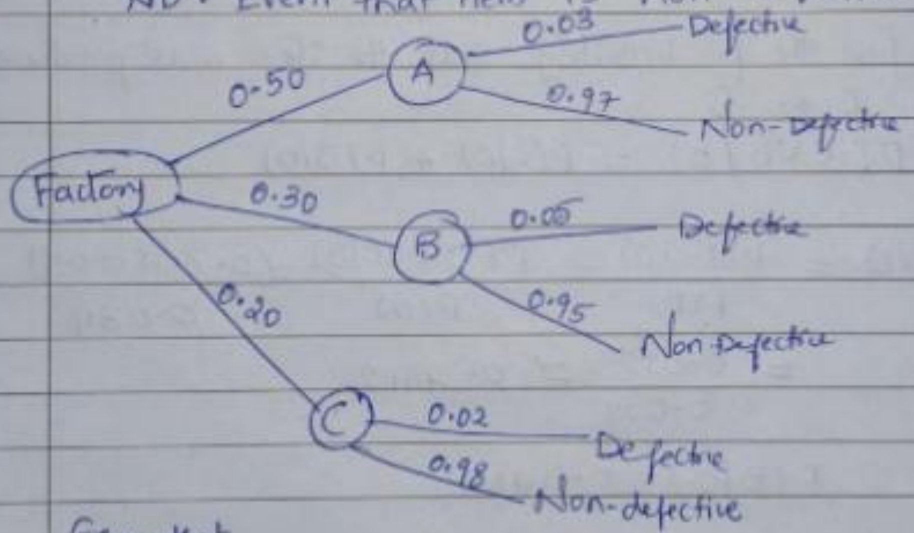
Let D: Event that item is defective

A: Event item produced by machine A

B: Event item produced by machine B

C: Event item produced by machine C

ND: Event that item is non-defective



Given that,

$P(A)=0.50$	$P(D/A)=0.03$	$P(ND/A)=0.97$
$P(B)=0.30$	$P(D/B)=0.05$	$P(ND/B)=0.95$
$P(C)=0.20$	$P(D/C)=0.02$	$P(ND/C)=0.98$

By total probability,

$P(\text{item found defective}) = P(D)$

$= P(A\cap D)+P(B\cap D)+P(C\cap D)$

$= P(A)P(D/A)+P(B)P(D/B)+P(C)P(D/C)$

$= (0.50)(0.03)+(0.30)(0.05)+(0.20)(0.02)$

$= 0.015+0.015+0.004$

$= 0.034$

$\therefore P(D) = 0.034$

(ii) To find the probability that the item was produced by machine C:

$$P(C/D) = \frac{P(C\cap D)}{P(D)} = \frac{P(C)P(D/C)}{P(D)}$$

$$= \frac{(0.20)(0.02)}{0.034} = \frac{0.004}{0.034} = 0.1176$$

$\therefore P(C/D) = 0.1176$

$\therefore$ Probability that the given defective item was produced by machine C $= 0.1176$ (or) 11.76%

Problem 2 - Solution

ii) To find the probability that the item was produced by C or B

$P(C \cup B / D) = P(C/D) + P(B/D)$

$P(B/D) = \frac{P(B \cap D)}{P(D)} = \frac{P(B)P(D/B)}{P(D)} = \frac{(0.30)(0.05)}{0.034}$

$= \frac{0.015}{0.034} \simeq 0.4412$

$\therefore P(B/D) = 0.4412$

$\therefore P(C \cup B/D) = P(C/D) + P(B/D) = 0.1176 + 0.4412 = 0.5588$

$\therefore$ Probability that item was produced by machine C or B, given that it is defective is app. 0.5588 (or) 55.88%

iii) By total probability,

P(Item found Non-Defective) = P(ND)

$= P(A \cap ND) + P(B \cap ND) + P(C \cap ND)$

$= P(A)P(ND/A) + P(B)P(ND/B) + P(C)P(ND/C)$

$= (0.50)(0.97) + (0.30)(0.95) + (0.20)(0.98)$

$= 0.485 + 0.285 + 0.196$

$= 0.966$

$\therefore P(ND) = 0.966$

To find the probability that the item was produced by machine A:

$P(A/ND) = \frac{P(A \cap ND)}{P(ND)} = \frac{P(A)P(ND/A)}{P(ND)}$

$= \frac{(0.50)(0.97)}{0.966} = \frac{0.485}{0.966} = 0.502$

$\therefore P(A/ND) = 0.502$

$\therefore$ Probability that the item was produced by machine A given that it is non-defective is 0.502 (or) 50.2%

Problem 3

EC2R
JULY 2024

innovate achieve lead

Hardik Pandya, Rishabh Pant and Surya Kumar Yadav are in the race of leading Indian cricket team in the next world cup with probabilities 0.2, 0.5 and 0.3 respectively. The probabilities of getting an increase in the match fee by Hardik, Rishabh and Surya are 0.3, 0.6 and 0.5 respectively if they become the Captain.

If there is an increase in match fee then find the probability

I. that it is because of Hardik
II. that it is because of Rishabh
III. that it is because of Surya Kumar

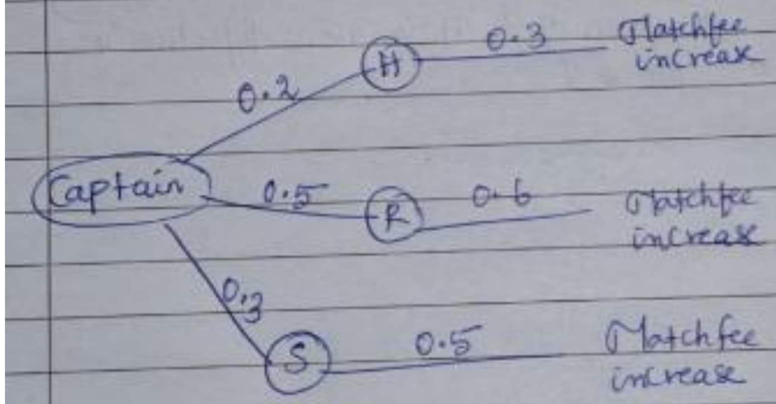
Problem 3 - Solution

Let, Event H: Hardik Pandya becomes the captain

Event R: Rishabh Pant becomes the captain

Event S: Surya Kumar Yadav becomes the captain

Event F: There is an increase in the match fee



Given that, $P(H) = 0.2$, $P(F/H) = 0.3$

$P(R) = 0.5$, $P(F/R) = 0.6$

$P(S) = 0.3$, $P(F/S) = 0.5$

By Law of total probability, we have

$P(F) = P(H)P(F/H) + P(R)P(F/R) + P(S)P(F/S)$

$= (0.2)(0.3) + (0.5)(0.6) + (0.3)(0.5)$

$= 0.06 + 0.3 + 0.15 = 0.51 \Rightarrow P(F) = 0.51$

(I) To find the probability that it is because of Hardik:

$P(H/F) = \frac{P(H)P(F/H)}{P(F)} = \frac{(0.2)(0.3)}{0.51} = \frac{0.06}{0.51} = \frac{6}{51}$

$= \frac{2}{17} = 0.1176$

$\therefore P(H/F) = 0.1176$

$\therefore$ The probability that increase in matchfee is because of Hardik $= 0.1176$ (or) 11.76%

(II) To find the probability that it is because of Rishabh:

$P(R/F) = \frac{P(R)P(F/R)}{P(F)} = \frac{(0.5)(0.6)}{0.51} = \frac{0.30}{0.51} = \frac{30}{51}$

$= \frac{10}{17} = 0.5882$

$\therefore P(R/F) = 0.5882$

$\therefore$ The probability that increase in matchfee is because of Rishabh $= 0.5882$ (or) 58.82%

(III) To find the probability that it is because of Surya Kumar:

$P(S/F) = \frac{P(S)P(F/S)}{P(F)} = \frac{(0.3)(0.5)}{0.51} = \frac{0.15}{0.51} = \frac{15}{51}$

$= \frac{5}{17} = 0.2941$

$\therefore P(S/F) = 0.2941$

$\therefore$ The probability that increase in matchfee is because of Suryakumar yadav is $= 0.2941$ (or) 29.41%

lead

Problem 4

EC2R
JAN 2024

In a manufacturing company Machine 1 produces 40% of the items and Machine 2 and Machine 3 produces 25% and 35% of the items respectively. But from the past records it is found that 15%, 20% and 25% of the items they produce are defective.

Then find

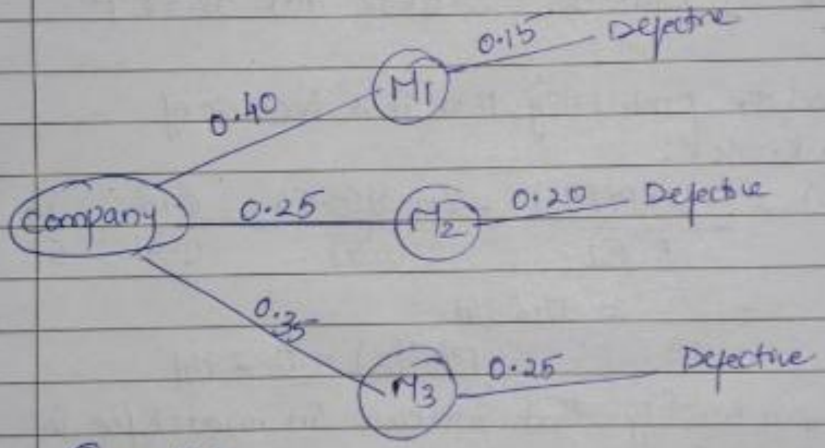
i) Total percentage of defective items produced

ii) If a defective item is selected randomly, then find the probability that it is produced by Machine 1

iii) If a defective item is selected randomly, then find the probability that it is not produced by Machine 2

Problem 4 - Solution

Let Event M_1: Items produced by Machine M_1
Event M_2: Items produced by Machine M_2
Event M_3: Items produced by Machine M_3
Event D : Item is defective



Given that,

$P(M_1) = 0.40$ $\qquad P(D/M_1) = 0.15$

$P(M_2) = 0.25$ $\qquad P(D/M_2) = 0.20$

$P(M_3) = 0.35$ $\qquad P(D/M_3) = 0.25$

(i) By the law of total probability, we have

$$P(D) = P(M_1)P(D/M_1) + P(M_2)P(D/M_2) + P(M_3)P(D/M_3)$$
$$= (0.40)(0.15) + (0.25)(0.20) + (0.35)(0.25)$$
$$= 0.06 + 0.05 + 0.0875 = 0.1975$$
$$\therefore P(D) = 0.1975$$

(ii) To find the probability that it is produced by machine M_1:

$$P(M_1/D) = \frac{P(M_1)P(D/M_1)}{P(D)} = \frac{(0.40)(0.15)}{0.1975}$$
$$= \frac{0.06}{0.1975} = 0.3038 \qquad \therefore P(M_1/D) = 0.3038$$

$\therefore$ Probability that it is produced by Machine M_1 given that it is defective $= 0.3038$ (or) 30.38%

(iii) To find the probability that it is not produced by Machine M_2

$$P(\bar{M_2}/D) = 1 - P(M_2/D)$$
$$= 1 - \left(\frac{P(M_2)\,P(D/M_2)}{P(D)}\right)$$
$$= 1 - \left(\frac{(0.25)(0.20)}{0.1975}\right) = 1 - \left(\frac{0.05}{0.1975}\right)$$
$$= 1 - 0.2532 = 0.7468$$
$$\therefore P(\bar{M_2}/D) = 0.7468$$

$\therefore$ The probability that it is not produced by Machine M_2, given that it is defective $= 0.7468$ (or) 74.68%

Problem 5

EC2M
JAN 2024

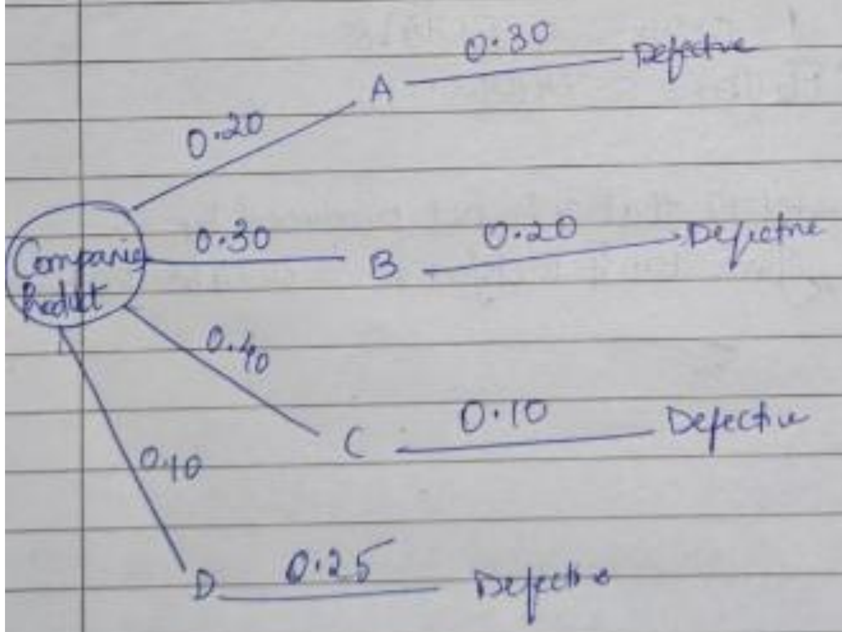
innovate achieve lead

A product is manufactured by four companies A, B,C and D. They produce 20%, 30%, 40% and 10% of the products released in the market. But it is observed that 30%, 20%, 10% and 25% are defective items manufactured by them respectively. Then find

I. Total percentage of defective items manufactured
II. The percentage that the selected defective item is produced by company A
III. The percentage that the selected defective item is not produced by company B
IV. The percentage that the selected defective item is not produced by company D

Problem 5 - Solution

Let Event A: Product manufactured by company A
Event B : Product manufactured by Company B
Event C : Product manufactured by company C
Event D: Product manufactured by Company D
Event X : Defective Item



Given that, $P(A) = 0.20$ $P(X/A) = 0.30$
$P(B) = 0.30$ $P(X/B) = 0.20$
$P(C) = 0.40$ $P(X/C) = 0.10$
$P(D) = 0.10$ $P(X/D) = 0.25$

i) By the law of total probability, we have

$$P(X) = P(A)P(X/A) + P(B)P(X/B) + P(C)P(X/C) + P(D)P(X/D)$$

$$= (0.20)(0.30) + (0.30)(0.20) + (0.40)(0.10) + (0.10)(0.25)$$

$$= 0.06 + 0.06 + 0.04 + 0.025 = 0.185$$

$\therefore P(X) = 0.185$

$\therefore$ Total percentage of the defective items manufactures $= 18.5\%$

ii) To find the percentage that the selected defective item is produced by company A

$$P(A/X) = \frac{P(A)P(X/A)}{P(X)} = \frac{(0.20)(0.30)}{0.185} = \frac{0.06}{0.185} = 0.3243$$

$\therefore P(A/X) = 0.3243$

$\therefore$ The percentage that the selected defective item is produced by company $A = 32.43\%$

Problem 5 - Solution

(III) To find the percentage that the selected defective item is not produced by company B.

$$P(B^c/x) = 1 - P(B/x) = 1 - \left(\frac{P(B)\,P(x/B)}{P(x)}\right)$$

$$= 1 - \left(\frac{(0.30)(0.20)}{0.185}\right) = 1 - \left(\frac{0.06}{0.185}\right)$$

$$= 1 - 0.3243 = 0.6757$$

$\therefore\ P(B^c/x) = 0.6757$

$\therefore$ The percentage that the selected defective item is not produced by B = 67.57%.

(IV) To find the percentage that the selected defective item is not produced by company D.

$$P(D^c/x) = 1 - P(D/x) = 1 - \left(\frac{P(D)\,P(x/D)}{P(x)}\right)$$

$$= 1 - \left(\frac{(0.10)(0.25)}{0.185}\right) = 1 - \left(\frac{0.025}{0.185}\right)$$

$$= 1 - 0.1351 = 0.8649$$

$\therefore\ P(D^c/x) = 0.8649$

$\therefore$ The percentage that the selected defective item is not produced by company D = 86.49%

Naïve Bayes Classifier

- Naïve Bayes classifiers are a collection of algorithms based on Bayes' Theorem. It is not a single algorithm but a family of algorithms where all of them share a common principle, i.e., every pair of features being classified as independent of each other.
- Naïve Bayes classifier is one of the most simple and effective classification algorithms, that aids in the rapid development of machine learning models with rapid prediction capabilities.
- It is highly used in text classification tasks where data contains high dimension (as each word represent one feature in the data). It is fast and making prediction easy with high dimension data.
- It is a probabilistic classifier, that predicts the probability of an instance belongs to a class with a given set of feature value. It uses Bayes theorem in the algorithm for training and prediction.

Why is it called Naïve Bayes?

- The “Naïve” part of the name indicates the simplifying assumption made by the Naïve Bayes Classifier. The classifier assumes that the features used to describe an observation are conditionally independent, given the class label.

Assumptions of Naïve Bayes

The fundamental Naive Bayes assumption is that each feature makes an:

- **Feature independence:** The features of the data are conditionally independent of each other, given the class label.
- **Continuous features are normally distributed:** If a feature is continuous, then it is assumed to be normally distributed within each class.
- **Discrete features have multinomial distributions:** If a feature is discrete, then it is assumed to have a multinomial distribution within each class.
- **Features are equally important:** All features are assumed to contribute equally to the prediction of the class label.
- **No missing data:** The data should not contain any missing values.

Types of Naïve Bayes Model

- **Gaussian Naïve Bayes Classifier :** The Gaussian Naive Bayes classifier assumes that the attributes of a dataset have a normal distribution. Here, if the attributes have continuous values, the classification model assumes that the values are sampled from a Gaussian distribution.
- **Multinomial Naïve Bayes:** When the input data is multinomially distributed, we use the multinomial naive Bayes classifier. This algorithm is primarily used for document classification problems like sentiment analysis.
- **Bernoulli Naïve Bayes:** The Bernoulli Naive Bayes classification works in a similar manner to the multinomial classification. The difference is that the attributes of the dataset contain boolean values representing the presence or absence of a particular attribute in a data point.

Advantages of Naïve Bayes Classifier

- Easy to implement and computationally efficient.
- Effective in cases with a large number of features.
- Performs well even with limited training data.
- It performs well in the presence of categorical features.
- For numerical features data is assumed to come from normal distributions

Disadvantages of Naive Bayes Classifier

- Assumes that features are independent, which may not always hold in real-world data.
- Can be influenced by irrelevant attributes.
- May assign zero probability to unseen events, leading to poor generalization.

Implementation of Naïve Bayes classifier

To implement a Naive Bayes classifier, we perform three steps.

- First, we calculate the probability of each class label in the training dataset.
- Next, we calculate the conditional probability of each attribute of the training data for each class label given in the training data.
- Finally, we use the Bayes theorem and the calculated probabilities to predict class labels for new data points. For this, we will calculate the probability of the new data point belonging to each class. The class with which we get the maximum probability is assigned to the new data point.

Problem 6

Consider the following data set:

Find out whether the object with attribute **Confident = Yes, Sick = No** will Fail or Pass using Bayesian Classification.

Confident	Studied	Sick	Result
Yes	No	No	Fail
Yes	No	Yes	Pass
No	Yes	Yes	Fail
No	Yes	No	Pass
Yes	Yes	Yes	Pass

Data set for classification

Problem 6 - Solution

Confident	Studied	Sick	Result
Yes	NO	NO	Fail
Yes	NO	Yes	Pass
NO	Yes	Yes	Fail
NO	Yes	NO	Pass
Yes	Yes	Yes	Pass

$P(Pass) = \frac{3}{5}$

$P(Fail) = \frac{2}{5}$

(Data set for classification)

Let $X = \{Confident = Yes, Sick = No\}$

Confident \ Result →	Pass	Fail
Yes	2/3	1/2
No	1/3	1/2

Sick \ Result →	Pass	Fail
Yes	2/3	1/2
NO	1/3	1/2

$P(Result = Pass / X) = P(Result = Pass / Confident = Yes, Sick = No)$

$= P(Result = Pass) * P(Confident = Yes / Result = Pass) * P(Sick = No / Result = Pass)$

$= \frac{3}{5} \times \frac{2}{3} \times \frac{1}{3} = \frac{2}{15} = 0.1333$

$\therefore P(Result = Pass / X) = 0.1333$

$P(Result = Fail / X) = P(Result = Fail / Confident = Yes, Sick = No)$

$= P(Result = Fail) \times P(Confident = Yes / Result = Fail) * P(Sick = No / Result = Fail)$

$= \frac{2}{5} \times \frac{1}{2} \times \frac{1}{2} = \frac{1}{10} = 0.1$

$\therefore P(Result = Fail / X) = 0.1$

$\therefore P(Result = Pass / X) > P(Result = Fail / X)$

$\therefore$ The new object with attribut X,

$X = \{Confident = Yes, Sick = No)$ belongs to pass clas.

Problem 7

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

$$New\ Instance = (Red, SUV, Domestic) \rightarrow (Yes\ or\ No)$$

Problem 7 - Solution

Let the New instance be
$X = \{Red, SUV, Domestic\}$

$P(yes) = \frac{5}{10} = \frac{1}{2}$

$P(No) = \frac{5}{10} = \frac{1}{2}$

	Stolen	
Type	Yes	No
Sports	4/5	2/5
SUV	1/5	3/5

	Stolen	
Colour	Yes	No
Yellow	2/5	3/5
Red	3/5	2/5

	Stolen	
Origin	Yes	No
Domestic	2/5	3/5
Imported	3/5	2/5

$P(Yes/x) = P(Yes/Red, SUV, Domestic)$

$= P(Yes) \times P(Red/Yes) \times P(SUV/Yes) \times P(Domestic/Yes)$

$= \frac{1}{2} \times \frac{3}{5} \times \frac{1}{5} \times \frac{2}{5} = \frac{3}{125} = 0.024$

$\therefore P(Yes/Red, SUV, Domestic) = 0.024$

$P(No/x) = P(No/Red, SUV, Domestic)$

$= P(No) \times P(Red/No) \times P(SUV/No) \times P(Domestic/No)$

$= \frac{1}{2} \times \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} = \frac{9}{125} = 0.072$

$\therefore P(No/Red, SUV, Domestic) = 0.024$

$P(No/x) > P(Yes/x)$

$\therefore$ The new instance belongs to 'No' class

Random Variables

Motivation:

In most statistical problems we are concerned with one number or a few numbers that are associated with the outcomes of experiments.

When inspecting a manufactured product we may be interested only in the number of defectives; in the analysis of a road test we may be interested only in the average speed and the average fuel consumption; and in the study of the performance of a miniature rechargeable battery we may be interested only in its power and life length. All these numbers are associated with situations involving an element of chance—in other words, they are values of random variables.

Random Variables

A random variable is a real valued function defined on the sample space

- The random variable is the observation.
- The random variable is a function of the observation.

A **random variable** consists of an experiment with a probability measure P[·] defined on a sample space S and a function that assigns a real number to each outcome in the sample space of the experiment.

Advantages:

- When the elements of the sample space are non-numeric, they can be quantified by assigning a real number to every event of the sample space. A real number X associated with the outcome of a random experiment.

- Infinite sample space can be converted to finite

Random Variables - Classification

Random variables are usually classified according to the number of values they can assume. Random variable can be classified as

1) discrete random variables, which can take on only a finite number
2) an uncountable infinity of values; called continuous random variables

Examples:

- **Number of accidents on a highway – Discrete**
- **Weights of the students of the class – Continuous**

Discrete Random Variables – Problem 8

EC2M July 2022

A random variable X can assume five values:0,1,2,3,4. A portion of the probability distribution is shown here:

X	0	1	2	3	4
P(x)	0.1	0.25	0.25	?	0.2

i) Find p(3)

ii) Calculate the mean

iii) What is the probability that X is greater than 3?

iv) What is the probability that X is 3 or less?

v) What is the probability that X is in between 1 and 3?

Problem 8 - Solution

Given probability $dist^n$ is

X	0	1	2	3	4
$P(X=x)$	0.1	0.25	0.25	?	0.2

(i) We know that, $\sum P(x) = 1$

$\Rightarrow 0.1 + 0.25 + 0.25 + P(3) + 0.2 = 1$

$\Rightarrow 0.8 + P(3) = 1$

$\Rightarrow P(3) = 1 - 0.8$

$\Rightarrow P(3) = 0.2$

(ii) Mean $= E(X) = \sum x\,p(x)$

$= 0(0.1) + 1(0.25) + 2(0.25) + 3(0.2) + 4(0.2)$

$= 0 + 0.25 + 0.5 + 0.6 + 0.8$

$= 2.15$

$\therefore$ Mean $= 2.15$

(iii) $P(X > 3) = P(X=4) = 0.2$

$\therefore P(X > 3) = 0.2$

(iv) $P(X \leq 3) = P(X=0) + P(X=1) + P(X=2) + P(X=3)$

$= 0.1 + 0.25 + 0.25 + 0.2$

$= 0.8$

$\therefore P(X \leq 3) = 0.8$

(v) $P(1 < X < 3) = P(X=2) = 0.25$

$\therefore P(1 < X < 3) = 0.25$

Discrete Random Variables – Problem 9

APR EC2 22

The probabilities that number of times in a day certain computer may malfunction is as follows:

No. of malfunctions in a day	0	1	2	3	4	5	6
Probability	0.17	k	0.27	0.16	0.07	0.03	0.01

Find

(a) the value of k

(b) mean and variance of the malfunction of certain computer

Problem 9 - Solution

8. Given probability distribution is

No. of malfunctions in a day	0	1	2	3	4	5	6
probability	0.17	K	0.27	0.16	0.07	0.03	0.01

(a) We know that

$$\sum p(x) = 1$$

$\Rightarrow 0.17 + K + 0.27 + 0.16 + 0.07 + 0.03 + 0.01 = 1$

$\Rightarrow 0.71 + K = 1$

$\Rightarrow K = 1 - 0.71$

$\Rightarrow K = 0.29$

(b) Mean $= E(x) = \sum x\,p(x)$

$= 0(0.17) + 1(0.29) + 2(0.27) + 3(0.16) + 4(0.07)$
$+ 5(0.03) + 6(0.01)$

$= 0 + 0.29 + 0.54 + 0.48 + 0.28 + 0.15 + 0.06$

$= 1.8$

$\therefore$ Mean $= 1.8$

$$Var(x) = E(x^2) - (E(x))^2$$

$E(x^2) = \sum x^2 p(x)$

$= (0)^2(0.17) + (1)^2(0.29) + (2)^2(0.27) + (3)^2(0.16)$
$+ (4)^2(0.07) + (5)^2(0.03) + (6)^2(0.01)$

$= 0 + 0.29 + 4(0.27) + 9(0.16) + 16(0.07)$
$+ 25(0.03) + 36(0.01)$

$= 0.29 + 1.08 + 1.44 + 1.12 + 0.75 + 0.36$

$= 5.04$

$E(x^2) = 5.04$

$Var(x) = E(x^2) - (E(x))^2$

$= 5.04 - (1.8)^2$

$= 5.04 - 3.24$

$= 1.8$

$\therefore Var(x) = 1.8$

Random Variable – Probability Distribution - Problem 10

A student takes two courses. In each course, the student will earn a B with probability 0.6 or a C with probability 0.4, independent of the other course. To calculate a grade point average (GPA), a B is worth 3 points and a C is worth 2 points. The student's GPA is the sum of the GPA for each course divided by 2. Make a table of the sample space of the experiment and the corresponding values of the student's GPA, G.

The sample space, probabilities and corresponding grades for the experiment are

Outcome	$P[\cdot]$	G
BB	0.36	3.0
BC	0.24	2.5
CB	0.24	2.5
CC	0.16	2

Continuous Random Variables – Problem 11

EC2 M S1-24

A data scientist is working on a machine learning model that predicts house prices based on various features such as size, location, and number of bedrooms. The model has some error in its predictions, represented by the random variable X. The error X follows the following probability density function (PDF):

$$f(x) = \begin{cases} \dfrac{k}{6} & -3 \leq x \leq 3 \\ 0 & otherwise \end{cases}$$

Calculate $i)$ the value of k , $ii)$ the expected value of error $iii)$ Compute the variance of the error.

Problem 11 - Solution

i) For a valid probability distribution, the total probability over the range must equal 1.

$$\therefore \int_{-3}^{3} f(x)dx = 1 \rightarrow \int_{-3}^{3} \frac{k}{6} dx = 1 \rightarrow \frac{k}{6}\big(3-(-3)\big) = 1 \rightarrow k = 1.$$

$$ii)\ \ E(X) = \int_{-3}^{3} x.\frac{1}{6}\, dx = \left[\frac{x^2}{12}\right]_{-3}^{3} = 0$$

$$iii)\ Var(x) = E(X^2) - \big(E(X)\big)^2.$$

$$E(X^2) = \int_{-3}^{3} \frac{x^2}{6}\, dx = \left[\frac{x^3}{18}\right]_{-3}^{3} = \frac{1}{6}\left(\frac{27}{3} - \left(-\frac{27}{3}\right)\right) = \frac{18}{6} = 3\,.$$

$$Var(X) = 3 - (0)^2 = 3.$$

Continuous Random Variables – Problem 12

EC2R
JULY 2024

Let f(x) be a continuous random variable defined in [0,2] as
$f(x) = k(x + 1)/4$, then find
(a) k value
(b) mean of x
(c) Expectation of x^2
(d) Variance
(e) $P(0.5 < x < 1.5)$

Problem 12 - Solution

a) $\int_0^2 \frac{k(x+1)}{4} dx = 1 \Rightarrow \frac{k}{4}\left[\frac{(x+1)^2}{2}\right]_0^2 = 1$ **K =1**

b) $E(x) = \int_0^2 \frac{x(x+1)}{4} dx = \frac{1}{4}\left[\frac{x^3}{3} + \frac{x^2}{2}\right]_0^2 = \frac{7}{6} = \quad 1.1666$

c) $E(x^2) = \int_0^2 \frac{x^2(x+1)}{4} dx \quad = \frac{1}{4}\left[\frac{x^4}{4} + \frac{x^3}{3}\right]_0^2 = \frac{5}{3} = 1.6666$

d) $Var(x) = E(x^2) - [E(x)]^2 = \frac{11}{36} = 0.3055$

e) $P(0.5 < x < 1.5) = \int_{0.5}^{1.5} \frac{x+1}{4} dx$

$$= \left[\frac{(x+1)^2}{8}\right]_{0.5}^{1.5} = 0.5$$

Continuous Random Variables – Problem 13

EC2R
JAN 2024

Let $f(x) = k(x + 1)/2$ where $0 < x < 2$,
$= 0$, otherwise

Then find

i) k value

ii) $P(1/2 < x < 2)$

iii) $P(X > 1/4)$

Problem 13 - Solution

i) To find Value of k:

$$\int_0^2 \frac{k(x+1)}{2}dx = 1$$

$$\frac{k}{2}\left[\frac{x^2}{2}+x\right]_0^2 = 1$$

$$\frac{k}{2}[(2+2)-(0+0)] = 1$$

Therefore $k = \frac{1}{2}$

ii) P(1/2 < x < 2)

$$P(\frac{1}{2} < x < 2) = \int_{1/2}^{2} \frac{1}{2}\cdot\frac{(x+1)}{2}dx$$

$$P\left(\frac{1}{2} < x < 2\right) = \frac{1}{4}\left[\frac{x^2}{2}+x\right]_{\frac{1}{2}}^{2}$$

$$= \frac{27}{32} == 0.84375$$

iii) P(X > ¼)

$$P(X > 1/4) = \frac{1}{4}[\frac{x^2}{2}+x]_{1/4}^{2}$$

$$= \frac{119}{128} = 0.9297$$

Joint Probability Distribution Function

- Let $X = \{x_1, x_2, \ldots, x_m\}$ and $Y = \{y_1, y_2, \ldots, y_n\}$ be two discrete random variables. Then $P(x, y) = J_{ij}$ is called joint probability function of X and Y if it satisfies the conditions:

$$(i)\ J_{ij} \geq 0 \quad (ii) \sum_{i=1}^{m} \sum_{j=1}^{n} J_{ij} = 1$$

- Set of values of this joint probability function J_{ij} is called joint probability distribution of X and Y.

	y_1	y_2	$\ldots$	y_n	Sum
x_1	J_{11}	J_{12}	$\ldots$	J_{1n}	$f(x_1)$
x_2	J_{21}	J_{22}	$\ldots$	J_{2n}	$f(x_2)$
$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$	$\ldots$
x_m	J_{m1}	J_{m2}	$\ldots$	J_{mn}	$f(x_m)$
Sum	$g(y_1)$	$g(y_2)$	$\ldots$	$g(y_n)$	$Total = 1$

Joint Probability Distribution Function

- If X and Y are discrete random variables, the joint probability distribution of X and Y is a description of the set of points (x,y) in the range of (X,Y) along with the probability of each point.

- The joint probability distribution of two discrete random variables is sometimes referred to as the **bivariate probability distribution** or **bivariate distribution.**

- Thus, we can describe the joint probability distribution of two discrete random variables is through a **joint probability mass function**

$$f(x,y) = P(X=x, Y=y)$$

Joint Density Function

When X and Y are continuous random variables, the **joint density function** $f(x, y)$ is a surface lying above the xy plane, and $P[(X, Y) \in A]$, where A is any region in the xy plane, is equal to the volume of the right cylinder bounded by the base A and the surface.

The function $f(x, y)$ is a **joint density function** of the continuous random variables X and Y if

1. $f(x, y) \geq 0$, for all (x, y),
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)\, dx\, dy = 1$,
3. $P[(X, Y) \in A] = \int \int_A f(x, y)\, dx\, dy$, for any region A in the xy plane.

Marginal Distributions

The **marginal distributions** of X alone and of Y alone are

$$g(x) = \sum_y f(x,y) \quad \text{and} \quad h(y) = \sum_x f(x,y)$$

for the discrete case, and

$$g(x) = \int_{-\infty}^{\infty} f(x,y)\, dy \quad \text{and} \quad h(y) = \int_{-\infty}^{\infty} f(x,y)\, dx$$

for the continuous case.

Independent Random Variables

- Let X and Y are two random variables with joint probability function $f(x, y)$ are said to be independent if following condition satisfied:
- $f(x, y) = g(x).h(y) for\ all\ x\ and\ y$, where $g(x)$ is marginal probability function of X and $h(y)$ is marginal probability function of Y.

Example: Suppose $f(x, y) = e^{-(x+y)}, x \geq 0, y \geq 0$. For this probability function, marginal probability function for X is $g(x) = e^{-x}, x \geq 0$ and marginal probability function for Y is $h(y) = e^{-y}, y \geq 0$.

Clearly $f(X, Y) = g(X).h(Y)$ So X and Y are independent variables.

Conditional Probability Distribution Function

- **Discrete Case:**

$$P[X = x_i / Y = y_j] = \frac{P[X = x_i / Y = y_j]}{P[Y = y_j]} = \frac{p_{ij}}{p_{.j}}$$

The set $\{x_i, p_{ij} / p_{.j}\}$, $i = 1, 2, 3, \ldots$ is called the conditional probability distribution of X given $Y = y_j$.

The conditional probability function of Y given $X = x_i$ is given by

$$P[Y = y_i / X = x_j] = \frac{P[Y = y_i / X = x_j]}{P[X = x_j]} = \frac{p_{ij}}{p_{i.}}$$

The set $\{y_i, p_{ij} / p_{i.}\}$, $j = 1, 2, 3, \ldots$ is called the conditional probability distribution of Y given $X = x_i$.

Continuous Case

Conditional Probability Density Function

$$f(y \mid x) = \frac{f(x, y)}{f_X(x)} \quad \textbf{or} \quad f(x \mid y) = \frac{f(x, y)}{f_Y(y)}$$

- the continuous version of Bayes' theorem

$$f(y \mid x) = \frac{f(x \mid y) f_Y(y)}{f_X(x)}$$

- another expression of the marginal pdf

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_{-\infty}^{\infty} f(x \mid y) f_Y(y) dy$$

$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_{-\infty}^{\infty} f(y \mid x) f_X(x) dx$$

Joint Discrete – Problem 14

EC2R
JAN 2024

Consider the following probability distribution of X and Y.

Y	X			
	0	**1**	**2**	**3**
0	0.05	0.25	0.05	0
1	0.05	0.15	0.05	0.05
2	0.10	0.15	0.10	0

i).Find marginal distribution of X

ii).Find marginal distribution of Y

iii).Find P($X < 2 / Y < 2$)

iv). Find P($X < 2 / Y = 1$)

Problem 14 – Solution

Solution:

i).Find marginal distribution of X

X	0	1	2	3	Total
P(x)	0.2	0.55	0.2	0.05	1

ii).Find marginal distribution of Y

Y	0	1	2	Total
P(y)	0.35	0.3	0.35	1

iii).Find P(X < 2 / Y < 2)

$$P(X < 2 \mid Y < 2) = \frac{P(X < 2 \cap Y < 2)}{P(Y < 2)}$$

$$= \frac{P(X = 0, Y = 0) + P(X = 1, Y = 0) + P(X = 0, Y = 1) + P(X = 1, Y = 1)}{P(Y = 0) + P(Y = 1)}$$

$$= \frac{0.05 + 0.05 + 0.25 + 0.15}{0.35 + 0.3} = \frac{10}{13} = 0.7692$$

Problem 14 – Solution

iv). Find P(X < 2 / Y =1)

$$P(X<2 \mid Y=1)=\frac{P(X<2 \cap Y=1)}{P(Y=1)}$$

$$=\frac{P(X=0,Y=1)+P(X=1,Y=1)}{P(Y=1)}$$

$$=\frac{0.05+0.15}{0.3}=\frac{2}{3}=0.67$$

Joint Discrete – Problem 15

EC2R
JAN 2025

The discrete random variables X and Y have joint probability mass function

$$f(x, y) = \begin{cases} cxy & x = 1,2,3, \quad y = 1,2 \\ 0 & otherwise \end{cases}.$$

i) Find the value of constant c.

ii) Find the marginal probability function for X and Y.

Problem 15 – Solution

i)

$$\therefore \sum_x \sum_y cxy = 1$$

$$\therefore c + 2c + 2c + 4c + 3c + 6c = 1$$

$$\therefore 18c = 1$$

$$\therefore c = \frac{1}{18}.$$

ii) Let $g(x)$ be marginal probability function of x.

$$\therefore g(x) = \sum_y f(x,y) = \sum_y cxy = cx(1+2) = \frac{3x}{18} = \frac{x}{6}, x = 1,2,3.$$

Let $h(y)$ be marginal probability function for y:

$$h(y) = \sum_x f(x,y) = \sum_x cxy = cy(1+2+3) = 6cy = \frac{6y}{18} = \frac{y}{3}, y = 1,2.$$

Joint Discrete – Problem 16

EC2R JULY 2024

Let X and Y are two independent random variables with the probability distributions given as

X	-1	0	1	2
P(X)	0.25	0.15	0.35	0.25

Y	0	1	2	3
P(Y)	0.10	0.20	0.30	0.40

Then find the following

a) If possible find the joint distribution of X and Y. If not, justify it.
b) If possible ,find $P(x < 1,Y < 2)$
c) If possible, find $P(X < 1 / Y < 2)$

Problem 16 – Solution

Part (a): Joint Distribution of X and Y

$P(X,Y)$	$Y=0$	$Y=1$	$Y=2$	$Y=3$
$X=-1$	$0.25\times 0.10 = 0.025$	$0.25\times 0.20 = 0.05$	$0.25\times 0.30 = 0.075$	$0.25\times 0.40 = 0.10$
$X=0$	$0.15\times 0.10 = 0.015$	$0.15\times 0.20 = 0.03$	$0.15\times 0.30 = 0.045$	$0.15\times 0.40 = 0.06$
$X=1$	$0.35\times 0.10 = 0.035$	$0.35\times 0.20 = 0.07$	$0.35\times 0.30 = 0.105$	$0.35\times 0.40 = 0.14$
$X=2$	$0.25\times 0.10 = 0.025$	$0.25\times 0.20 = 0.05$	$0.25\times 0.30 = 0.075$	$0.25\times 0.40 = 0.10$

Part (b): P(X<1, Y<2)

$$
\begin{aligned}
P(X<1,Y<2) &= P(X=-1,Y=0)+P(X=-1,Y=1)+P(X=0,Y=0)+P(X=0,Y=1) \\
&= 0.025+0.05+0.015+0.03 \\
&= 0.12
\end{aligned}
$$

Problem 16 – Solution

C) P(X<1|Y<2)

$P(X < 1 \mid Y < 2) = \frac{P(X<1 \text{ and } Y<2)}{P(Y<2)}$

We already calculated $P(X < 1 \text{ and } Y < 2) = 0.12.$

Next, we need to find $P(Y < 2)$:

$P(Y < 2) = P(Y = 0) + P(Y = 1) = 0.10 + 0.20 = 0.30$

Now, we can calculate the conditional probability:

$$\begin{aligned} P(X < 1 \mid Y < 2) &= \frac{P(X < 1 \text{ and } Y < 2)}{P(Y < 2)} \\ &= \frac{0.12}{0.30} \\ &= 0.4 \end{aligned}$$

Joint Continuous – Problem 17

EC2 M JULY 2024

Consider the probability density function $f(X,Y) = k\,(X + Y)/8$ where $0 < X < 2$ and $0 < Y < 2$. Then find

a) k value
b) Marginal probability density function of X
c) Marginal probability density function of Y
d) Are X and Y are independent? Justify it

Problem 17 – Solution

a) $\int_0^2 \int_0^2 \frac{k(x+y)}{8} dxdy = 1,$

Solving this with proper steps yield **K =1**

b) $f(x) = \int_0^2 \frac{x+y}{8} dy = \frac{x+1}{4}$

c) $f(y) = \int_0^2 \frac{x+y}{8} dx = \frac{y+1}{4}$

d) $f(xy) \neq f(x).f(y)$. Hence not independent

Joint Continuous – Problem 18

EC2 M FEB 2024

Consider the following probability distribution.

$f(x, y) = \frac{k}{2}(e^{-(x+y)}, 0 < x < \infty, 0 < y < \infty$

i) Find k value

ii) Check whether x and y are independent or not.

iii) Find P(0 < x < 2 , 0< y < 5)

iv) P(0 < x < 2 / 0 < y < 5)

Problem 18 – Solution

$$\int_0^{\infty}\int_0^{\infty}\frac{k}{2}e^{-(x+y)}dxdy = 1$$

$$\int_0^{\infty}(\frac{k}{2}e^{-y}\int_0^{\infty}e^{-x}dx)\,dy = 1$$

$$\int_0^{\infty}(\frac{k}{2}e^{-y}(0-(-1)))\,dy = 1$$

$$\int_0^{\infty}\frac{k}{2}e^{-y}dy = 1$$

$$\frac{k}{2}[-e^{-y}]_0^{\infty} = 1$$

$$\frac{k}{2}(0-(-1)) = 1$$

$$\frac{k}{2} = 1$$

$$k = 2$$

i) Check whether x and y are independent or not.

As Joint pdf is f(x, y) = $e^{-(x+y)}$ = $f(x).f(y)$

And marginal PDF are

f(x) = $\int_0^{\infty}e^{-(x+y)}dy$

f(y) = $\int_0^{\infty}e^{-(x+y)}dx$

Which gives

$$f(x) = -e^{-x}$$
$$f(y) = -e^{-y}$$

Thus f(x, y) = $f(x).f(y)$

Hence x and y are independent variables.

ii) Find P(0 < x < 2 , 0< y < 5)

$P(0 < x < 2, 0 < y < 5) = \int_0^2 \int_0^5 e^{-(x+y)}dydx$

=$(e^{-2}-1).(e^{-5}-1)$=0.859

iii) P(0 < x < 2 / 0 < y < 5)

$P(0 < x < 2 /\ 0 < y < 5) = \frac{P(0<x<2, 0<y<5)}{P(0<y<5)} = \frac{0.859}{0.993}$= 0.865

innovate achieve lead

THANK YOU