**Section A: -**

1. **What are the basic steps to understand the given set of data for prediction?**
2. Gathering the data (Training Data)
3. Data Preparation (Training data and Evaluation data)
4. Choosing a Model
5. Training
6. Evaluation (of Data)
7. Parameter Tuning
8. Prediction
9. **18 values were observed of a variable. Their mean was found to be 24.11. Another observation was subsequently made and the value observed was 35. What is the new value of the mean? If the sum of the squares of the deviation of these 19 values from their mean is 514.11, what is the standard deviation?**

= = = 24.11

= = = = = 24.68

Sum of squares of the deviation = Variance = σ² = 514.11

Standard Deviation = σ = 22.67

1. **If 𝑥̅ is defined as , then 𝑥̅ is an unbiased estimate of the population mean. True or False?**

True

1. **The standard deviation of two variables, based on 19 observations, is 5.34 and 5.4. The covariance between these two variables, based on the same observations, is -21.00. Estimate the correlation coefficient between these two variables. Are these two variables linearly related?**

Corr(X,Y) or r = = = = -0.72

Threshold = = = 0.45

The variables have moderate negative linear relationship.

1. **Write the expression for the total probability of an event.**

Law of Total Probability:

If B1, B2, B3, ⋯ is a partition of the sample space S, then for any event A we have

P(A) = =

1. **A certain firm has plants A, B, and C producing, respectively, 35%, 15% and 50%, of total output. The probabilities of a non-defective product are, respectively, 0.75, 0.95, and 0.85. A customer receives a defective product. What is the probability that it came from plant C?**

Given that:

P(A) = 0.35

P(B) = 0.15

P(C) = 0.5

Given P(non-defective product from A) = 0.75 🡺 Then, P(defective product from A) = 0.25

Given P(non-defective product from B) = 0.95 🡺 Then, P(defective product from B) = 0.05

Given P(non-defective product from C) = 0.85 🡺 Then, P(defective product from C) = 0.15

Let E be the event that a customer receives a defective product. We have to find out that it came from plant C, that is we have to compute P(C|E).

**From Bayes rule:**

P(E) = P(E|A) x P(A) + P (E|B) x P (B) + P (E|C) x P (C)

= 0.35x0.25 + 0.15x0.05 + 0.5x0.15

= 0.17

P(C|E)\*P(E) = P(CՈE) = P(E|C)\*P(C)

P(C|E) =

=

= **0.44176**

1. Give one example when logarithmic transformation is not a good option to apply on a non-linear relation
2. **In a regression model, the errors follow which distribution?**

Normal distribution or Gaussian distribution

1. **What is Occam’s razor in the context of predictive models?**

Simplest is the best.

More things should not be used than are necessary.

If two algorithms have broadly similar performance for the criteria identified as the most important for a particular project — accuracy and stability, say — we should always prefer the “simpler” one.

1. In a bivariate data, write the equations for determining the value of the regression coefficients which will minimize the sum of the squared errors.

<https://www.cse.wustl.edu/~jain/iucee/ftp/k_14slr.pdf>

**Section B: -**

1. Skin cancer rates have been steadily rising over recent years. It is thought that this may be due to ozone depletion.

The following data are ozone depletion rates in various localities and the rates of skin cancer.

|  |  |
| --- | --- |
| **Ozone depletion (%)** | **Skin cancer rate (%)** |
| 5 | 1 |
| 7 | 1 |
| 13 | 3 |
| 14 | 4 |
| 17 | 6 |
| 20 | 5 |
| 26 | 6 |
| 30 | 8 |
| 34 | 7 |
| 39 | 10 |
| 44 | 9 |

a. Fit a straight-line regression model to the data

b. What is the rate of skin cancer if ozone depletion is 40%?