Data Mining

Quiz # 2

Solution

Roll No: _____ **Section:** _____ **Date: 23-02-2024**

Question 1

In real-world data, tuples with missing values for some attributes are a common occurrence. Describe various methods for handling this problem.

- Ignore Tuple
- Use Statistical Methods. (Mean, Media, Mode)
- Introduce Special values
- Predict missing values using Regression Analysis

Question 2

Suppose a group of 12 sales price records has been sorted as follows:

2,10,11,13,15,35,50,55,72,92,204,210.

Partition them into three bins by each of the following methods:

(a) equal-frequency (equal-depth) partitioning

(b) equal-width partitioning

And then smooth noise by bin means for both partitioning schemes.

Total bins = 3

Data points = 12

(a) Equal-Frequency Binning - $N/K = 12/3 = 4$

Bin 1 = 2, 10, 11, 13

Bin 2 = 15, 35, 50, 55

Bin 3 = 72, 92, 204, 210

(b) Equal-width Partitioning

$$w = B - A / k = \frac{210 - 2}{3} = 69.3$$

$A + w = 2 + 69.3 = 71.3 \Rightarrow 71$

$A + 2w = 2 + 2(69.3) = 140.6 \Rightarrow 141$

$A + 3w = 2 + 3(69.3) = 209.9 \Rightarrow 210$

Bin 1 = 2, 10, 11, 13, 15, 35, 50, 55

Bin 2 = 72, 92

Bin 3 = 204, 210

Smoothing by Bin Means

Equal Frequency

Bin 1 = 9, 9, 9, 9

Bin 2 = 38.75, 38.75, 38.75, 38.75

Bin 3 = 144.5, 144.5, 144.5, 144.5

Equal-width

Bin 1 = 23.8, 23.8, 23.8, 23.8, 23.8, 23.8, 23.8, 23.8

Bin 2 = 82, 82

Bin 3 = 207, 207