Assignment - 4

Q-1 Use min-max normalization method to normalize the following data by setting minimum = 0 & max = 1
200, 300, 400, 600, 1000

→ Formula for min-max normalization,

$$v' = \frac{v - old(min)}{old(max) - old(min)} (new(max) - new(min)) + new(min)$$

new(min) = 0 old(min) = 200
new(max) = 1 old(max) = 1000

* for V = 300

$$v' = \frac{(300-200)}{(1000-200)} (1-0) + 0$$

$$= 0.125$$

* for V = 200

$$v' = \frac{(200-200)}{(1000-200)} (1-0) + 0$$

$$= 0$$

* for V = 400

$$v' = \frac{(400-200)}{(1000-200)} (1-0) + 0$$

$$= 0.25$$

* for V = 600

$$v' = \frac{(600-200)}{(1000-200)} (1-0) + 0$$

$$= 0.5$$

* for V = 1000

$$v' = \frac{(1000-200)}{(1000-200)} (1-0) + 0$$

$$= 1$$

Data before normalization,
200, 300, 400, 600, 1000
Data after normalization,
0, 0.125, 0.25, 0.5, 1

VISION

Q-2 If mean salary is 54,000₹ and std. deviation is 16,000₹ then find Z-score value of 73,600₹

→ Z-score normalization formula,

$$v' = \frac{v - \text{mean}}{\text{Std. deviation}}$$

$$= \frac{73,600 - 54000}{16,000}$$

$$v' = 1.225$$

Q-3 Suppose a group of sales price records has been sorted as follows 6, 9, 12, 13, 15, 25, 50, 70, 72, 92, 204, 232.
- Partition them into three bins by equal frequency (equi-depth) partition method.
- Perform data smoothing by mean, median and boundaries.

→ * Data smoothing by bin mean

B1 : 10, 10, 10, 10 ((6 + 9 + 12 + 13)/4 = 10)
B2 : 40, 40, 40, 40 ((15 + 25 + 50 + 70)/4 = 40)
B3 : 150, 150, 150, 150 ((72 + 92 + 204 + 232)/4 = 150)

* Data smoothing by bin median

B1 : 10.5, 10.5, 10.5, 10.5 ($\frac{9+12}{2}$ = 10.5)
B2 : 37.5, 37.5, 37.5, 37.5 ($\frac{25+50}{2}$ = 37.5)
B3 : 148, 148, 148, 148 ($\frac{92+204}{2}$ = 148)

* Data smoothing by bin boundaries

B1 : 6, 6, 13, 13
B2 : 15, 15, 70, 70
B3 : 72, 72, 232, 232

Q-4 Suppose that the data for analysis includes the attribute age. The age values for the data tuple are 13, 15, 16, 16, 19, 20, 23, 29, 35, 41, 44, 53, 62, 69, 72. use min-max normalization to transform the value for age onto the range [0.00, 1.00]

V = 45
old_min = 13
old_max = 72
new_min = 0
new_max = 1

$$V' = \frac{V - old\_min}{old\_max - old\_min} (new\_max - new\_min) + new\_min$$

$$= \frac{45 - 13}{72 - 13} (1 - 0) + 0$$

$$= \frac{32}{59}$$

$$V' = 0.54$$

Q-5 minimum salary is 20,000 Rs and maximum salary is 1,70,000 Rs.
- Map the salary 1,00,000 in new range of [60,000, 26,000] Rs using min-max normalization method.

—> V = 100000
new_min = 60000 old_min = 20,000
new_max = 26000 old_max = 170000

$$V' = \frac{V - old\_min}{old\_max - old\_min} (new\_max - new\_min) + new\_min$$

$$= \frac{100000 - 20000}{170000 - 60000} (260000 + 60000) + 60000$$

$$= \$1,06,666 + 60000$$

$$V' = 166666$$

Q-6 In real-world, data tuples with missing values for some attributes are a common occurence Describe various methods for handling this problem.

—> Real World's data tends to be incomplete, noisy and inconsistent and an important task when preprocessing the data is to fill in missing values, smooth-out noise and correct inconsistencies.

- Choosing the right technique is a choice that depends on the problem domain, data domain and our goal for the data mining process.

★ Ignore the data row

- This is usually done when the class label is missing or many attributes are missing from the row.

- However, you'll obviously get poor performance if the percentage of such rows is high.

★ Use a global constant to fill in for missing values.

- Decide on a new global constant value like "unknown", "N/A" or minus infinity that will be used to fill all the missing values.

- This technique is used because sometimes it just doesn't make sense to try and predict the missing value.

★ Use attribute mean

- Replace missing values of an attribute with the mean value for that attribute in database.

★ Use a data mining algorithm to predict the most probable value

- The value can be determined using regression inference based tools using bayesian normalisation, decision trees, clustering algorithm.

Q-7 Suppose that the data for analysis includes the attribute age.

- The age values for data tuples are,
13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

a) what is the mean of the data? what is median?

b) what is the mode of the data? comment on data's modality.

→ mean : It's an average of the numbers

$$\text{mean} = \frac{(13+15+16+16+19+20+20+\ldots+70)}{27}$$

$$= \frac{809}{27}$$

$$= 29.96$$

- median : It's the "middle" value of a sorted list of numbers.

In given data 14th element will be median of data. ∵ 25 is the median.

- mode : mode is simply the number which occurs most or appears most often.

☆ Bimodal :-

A dataset is bimodal if it has two modes.

- Instead of single value, there are two data values that tie for having the highest frequency

☆ Trimodal :-

A dataset is trimodal if it has three modes

- Modes of the given data is in the given series of data, there are two numbers that appears most are 25 & 35.

- So, the given data set is Bimodal, as it is having two modes.

☆ Multimodal :-

A set of numbers with four or more modes is multimodal.