3.3 Age : 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) Use Smoothing by bin means to smooth these data using bin depth of 3. Comment on the effects of this technique for the given data.

(b) How might you determine outliers in the data.

(c) what other methods are there for data smoothing.

Solution:

(a)

Bin 1 : 13, 15, 16
Bin 2 : 16, 19, 20
Bin 3 : 20, 21, 22
Bin 4 : 22, 25, 25
Bin 5 : 25, 25, 30
Bin 6 : 33, 33, 35
Bin 7 : 35, 35, 35
Bin 8 : 36, 40, 45
Bin 9 : 46, 52, 70

Smoothing by bin means:

Bin 1 : 15, 15, 15
Bin 2 : 18, 18, 18
Bin 3 : 21, 21, 21
Bin 4 : 24, 24, 24
Bin 5 : 27, 27, 27
Bin 6 : 34, 34, 34
Bin 7 : 35, 35, 35
Bin 8 : 40, 40, 40.
Bin 9 : 56, 56, 56.

(b) Outliers can be detected by using clustering. All the similar values are organised into groups or clusters. Values that fall outside the set of clusters are outliers.

(c) Data smoothing can be done by bin boundaries as follows.

Bin 1 : 13, 16, 16
Bin 2 : 16, 20, 20
Bin 3 : 20, 22, 22
Bin 4 : 22, 25, 25
Bin 5 : 25, 25, 30
Bin 6 : 33, 33, 35
Bin 7 : 35, 35, 35
Bin 8 : 36, 36, 45
Bin 9 : 46, 46, 70.

– Methods other than binning are regression technique & clustering.

(3.5) what are the value range for the following normalization methods.

(a) min-max normalization.
(b) z-score normalization.
(c) z-score normalization using the mean absolute deviation instead of standard deviation.
(d) normalization by decimal scaling.

Solution :

(a) min-max normalization.

It performs a linear transformation on the orignal data. Suppose that min_A and max_A are the minimum and maximum values

Page No.
Date

Value ranges of min-max normalization are
[new_min , new_max]

b) Value range of z-score normalization is
[(old_min_mean) / std_deviation , (old_max-mean)/std_dev]
i.e [−∞ , +∞]

c) Value range of z-score normalization using the mean absolute deviation

$$S_A = \frac{1}{n}\left(|v_1 - \bar{A}| + |v_2 - \bar{A}| + \cdots + |v_n - \bar{A}|\right)$$

Value range is $[\min_{S_A} \bar{A}, \max_{S_A} A - \bar{A}]$

d) Value range for normalization by decimal scaling is
[−1.0 , 1.0]

3.7

a). min-max normalization.

min age : 13

max age : 70

min max normalization to transform value 35

$$\frac{35-13}{70-13}\,(1.0-0)+0$$

$$= \frac{22}{57}$$

Ans. = $\boxed{0.385}$.

b) Mean age = 29.96

Standard deviation = 12.94

Z-score normalization to transform value 35 is

$$\frac{35-29.96}{12.94}$$

$$= \frac{5.04}{12.94} = 0.389.$$

Ans = $\boxed{0.389}$.

c) Decimal scaling.

Using the equation where $j=2$, $v=35$

By using decimal scaling

$$v' = \frac{35}{100} \rightarrow \text{as } j=2$$

$\boxed{v' = 0.35.}$

d) I would prefer decimal scaling for normalization as it would maintain data distribution and be intuitive to interpret, we can still continue mining on specific age groups.

Min max normalization does not permit any value to fall outside the current minimum and maximum values without encountering an 'out of bond error'.

Z-score normalization would not increase any information value of the attribute in terms of intuitiveness to users or the results.

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

5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215.

Partition them into 3 bins by each of the following methods.

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

(b) equal-width partitioning.

(c) clustering.

(a) equal-frequency partitioning.

bin 1 : 5, 10, 11, 13

bin 2 : 15, 35, 50, 55

bin 3 : 72, 92, 204, 215.

(b) equal-width partitioning.

The width of each interval will be $(215-5)/3 = 70$.

bin 1 : 5, 10, 11, 13, 15, 35, 50, 55, 72

bin 2 : 92

bin 3 : 204, 215.

(c) clustering.

we will form clusters among big gaps among data

bin 1 : 5, 10, 11, 13, 15

bin 2 : 35, 50, 55, 72, 92

bin 3 : 204, 215.