Data transformation can have the following activities

- **Smoothing :** It involves removal of noise from the data.
- **Aggregation :** It involves summarisation and data cube construction.
- **Generalization :** In generalization data is replaced by higher level concepts using concept hierarchy.
- **Normalization :** In normalization, attribute scaling is performed for a specified range.

  Example : To transform V in [min, max] to V' in [0,1], apply

  $$V' = (V\text{-Min}) / (\text{Max-Min})$$

  Scaling by using mean and standard deviation (useful when min and max are unknown or when there are outliers) :

  $$V' = (V\text{-Mean}) / \text{Std. Dev.}$$

- **Attribute/feature construction :** In this process new attributes may be constructed and used for data mining process.

1.6.3(B) Data Discretization

- The range of a continuous attribute is divided into intervals.
- Categorical attributes are accepted by only a few classification algorithms.
- By Discretization the size of the data is reduced and prepared for further analysis.
- Dividing the range of attributes into intervals would reduce the number of values for a given continuous attribute.
- Actual data values may be replaced by interval labels.
- Discretization process may be applied recursively on an attribute.

1.6.3(C) Data Transformation by Normalization

SPPU - May 17

Q. What are the different data normalization methods? Explain them in brief. **(May 17, 6 Marks)**

- Data Transformation by Normalization or standardization is the process of making an entire set of values have a particular property.
- Following methods may be used for normalization :

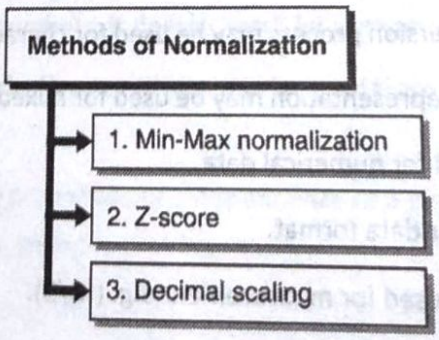
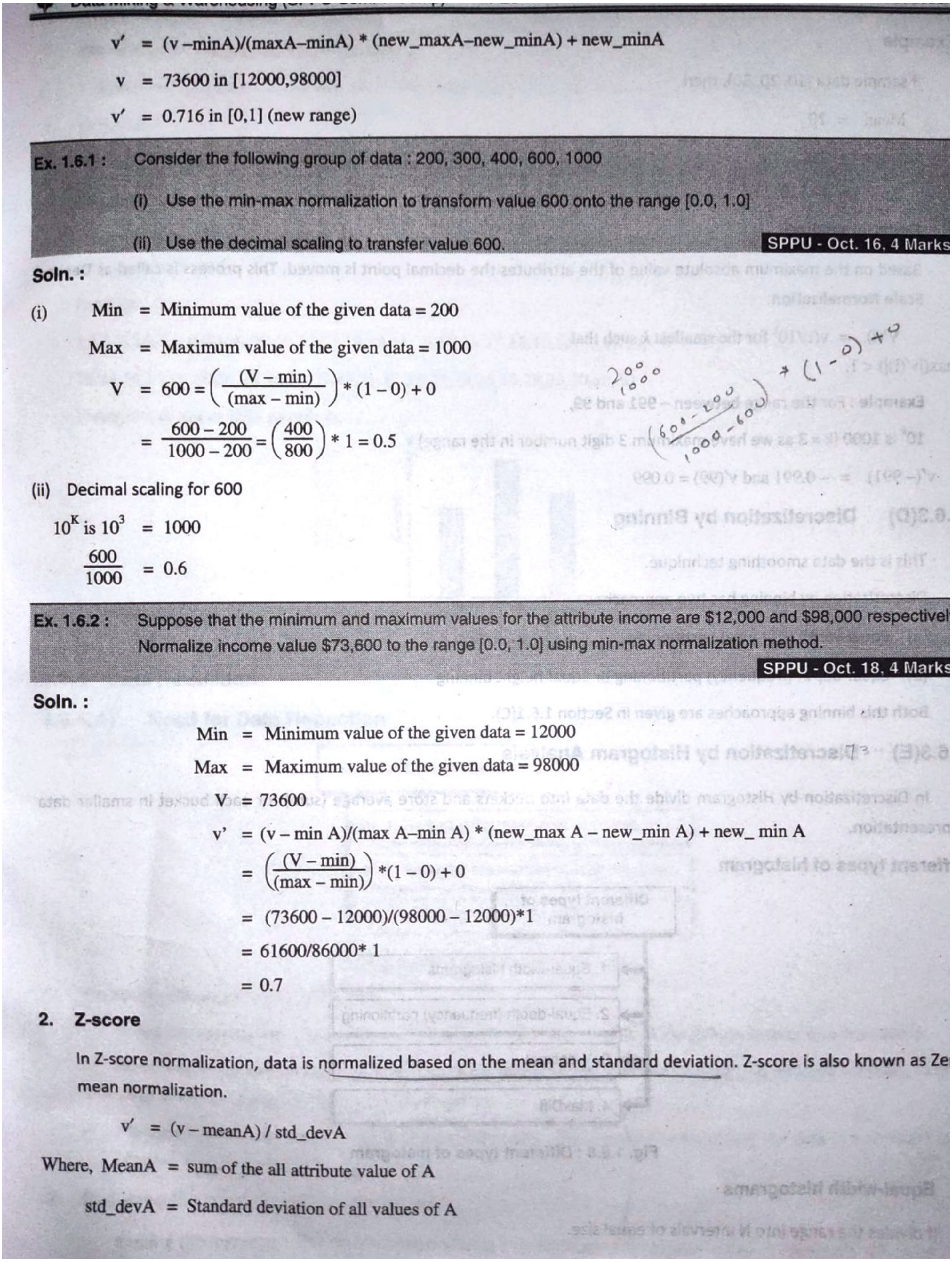


Fig. 1.6.7 : Methods of Normalization

1. Min-Max normalization

- Min-max normalization results in a linear alteration of the original data. The values are within a given range.
- Following formula may be used to perform mapping a v value, of an attribute A from range [minA,maxA] to a new range [new_minA,new_maxA],



$$v' = (v - minA)/(maxA - minA) * (new\_maxA - new\_minA) + new\_minA$$

$$v = 73600 \text{ in } [12000, 98000]$$

$$v' = 0.716 \text{ in } [0,1] \text{ (new range)}$$

Ex. 1.6.1 : Consider the following group of data : 200, 300, 400, 600, 1000

(i) Use the min-max normalization to transform value 600 onto the range [0.0, 1.0]

(ii) Use the decimal scaling to transfer value 600.

SPPU - Oct. 16, 4 Marks

Soln. :

(i) Min = Minimum value of the given data = 200

Max = Maximum value of the given data = 1000

$$V = 600 = \left(\frac{(V - min)}{(max - min)}\right) * (1 - 0) + 0$$

$$= \frac{600 - 200}{1000 - 200} = \left(\frac{400}{800}\right) * 1 = 0.5$$

$\frac{2000}{1000}$

$\left(\frac{600-200}{1000-600}\right) * (1-0) + 0$

(ii) Decimal scaling for 600

10^K is 10^3 = 1000

$$\frac{600}{1000} = 0.6$$

Ex. 1.6.2 : Suppose that the minimum and maximum values for the attribute income are \$12,000 and \$98,000 respectivel Normalize income value \$73,600 to the range [0.0, 1.0] using min-max normalization method.

SPPU - Oct. 18, 4 Marks

Soln. :

Min = Minimum value of the given data = 12000

Max = Maximum value of the given data = 98000

V = 73600

$$v' = (v - min\ A)/(max\ A - min\ A) * (new\_max\ A - new\_min\ A) + new\_\ min\ A$$

$$= \left(\frac{(V - min)}{(max - min)}\right) *(1 - 0) + 0$$

$$= (73600 - 12000)/(98000 - 12000)*1$$

$$= 61600/86000* 1$$

$$= 0.7$$

2. Z-score

In Z-score normalization, data is normalized based on the mean and standard deviation. Z-score is also known as Ze mean normalization.

$$v' = (v - meanA) / std\_devA$$

Where, MeanA = sum of the all attribute value of A

std_devA = Standard deviation of all values of A

Example

If sample data {10, 20, 30}, then

$$\text{Mean} = 20$$

$$\text{std\_dev} = 10$$

$$\text{So } v' = (-1, 0, 1)$$

3. Decimal scaling

Based on the maximum absolute value of the attributes the decimal point is moved. This process is called as **Decimal Scale Normalization.**

$$V'(i) = v(i)/10^k \text{ for the smallest } k \text{ such that}$$

$$\max(|v'(i)|) < 1.$$

Example : For the range between – 991 and 99,

10^k is 1000 (k = 3 as we have maximum 3 digit number in the range)

$$v'(-991) = -0.991 \text{ and } v'(99) = 0.099$$

1.6.3(D) Discretization by Binning

- This is the data smoothing technique.
- Discretization by binning has two approaches :
  - (a) Equal-width (distance) partitioning
  - (b) Equal-depth (frequency) partitioning or Equal-height binning
- Both this binning approaches are given in Section 1.6.1(C).

1.6.3(E) Discretization by Histogram Analysis

In Discretization by Histogram divide the data into buckets and store average (sum) for each bucket in smaller da representation.

Different types of histogram

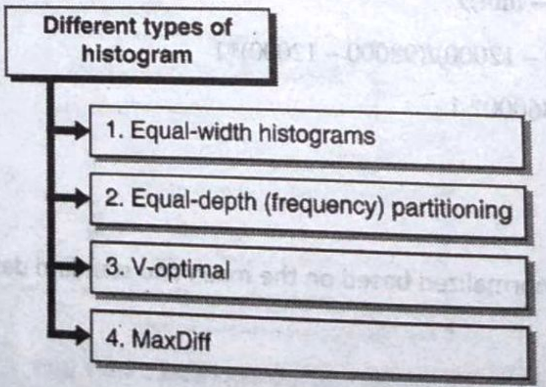

Different types of histogram
1. Equal-width histograms
2. Equal-depth (frequency) partitioning
3. V-optimal
4. MaxDiff

Fig. 1.6.8 : Different types of histogram

1. Equal-width histograms

It divides the range into N intervals of equal size.

Ex. 1.6.2 : For the given attribute marks values :

35, 45, 50, 55, 60, 65, 75

Compute mean, median, mode.

Also compute Five number summary of above data. **Oct. 18, 4 Marks**

Soln. :

(1) Mean

$$\bar{x} = \frac{(x_1 + x_2 + \ldots\ x_n)}{n}$$

$$\bar{x} = \frac{35 + 45 + 50 + 55 + 60 + 65 + 75}{7}$$

$$\bar{x} = \frac{385}{7} = 55$$

(2) Median

Sort the elements in ascending order

35	45	50	55	60	65	75

Middle element is 55

∴ median is 55

(3) Mode

Mode is most frequent value in data set. As each number appears one, so the frequency of all number is same. Therefore all 7 number are mode

(4) Five number summary

- Median → 55
- 1st Quartile → middle value of lower half
- 3rd Quartile → middle value of super half
- Minimum → 35
- Maximum → 75

$$\underbrace{35\quad 45\quad 50}\quad 55\quad \underbrace{60\quad 65\quad 75}$$

∴ First Quartile = Q_1 = 45

Third Quartile = Q_3 = 65

2. Outlier analysis by clustering

- Partition data set into clusters and one can store cluster representation only, i.e. replace all values of the cluster by that one value representing the cluster.
- Outliers can be detected by using clustering techniques, where related values are organized into groups or clusters.

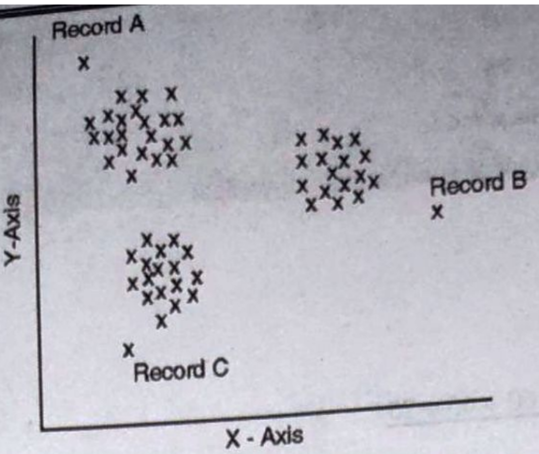


Fig. 1.6.5 : Graphical Example of Clustering

- Perform clustering on attributes values and replace all values in the cluster by a cluster representative.

3. Regression

- Regression is a statistical measure used to determine the strength of the relationship between one dependent variable denoted by Y and a series of independent changing variables.
- Smooth by fitting the data into regression functions.
- Use regression analysis on values of attributes to fill missing values.
- The two basic types of regression are linear regression and multiple regressions.
- The difference between Linear and multiple regressions is that former uses one independent variable to predict the outcome, while the later uses two or more independent variables to predict the outcome.
- The general form of each type of regression is :

Linear Regression : $Y = a + bX + u$

Multiple Regression : $Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + u$

Where, Y = The variable that we are trying to predict

X = The variable that we are using to predict Y

a = The intercept

b = The slope

u = The regression residual.

- In multiple regressions each variable is differentiated with subscripted numbers.
- Regression uses a group of random variables for prediction and finds a mathematical relationship between them This relationship is depicted in the form of a straight line (Linear regression) that approximates all the points the best way.
- Regression may be used to determine for e.g. price of a commodity, interest rates, the price movement of asset influenced by industries or sectors.

Log linear model

- In Log linear regression a best fit between the data and a log linear model is found.
- **Major assumption :** A linear relationship exists between the log of the dependent and independent variables.
- Log linear models are models that postulate a linear relationship between the independent variables and logarithm of the dependent variable.