7/9/21

DWDM

Tutorial - 4

U18co021 - Sahil Bondre

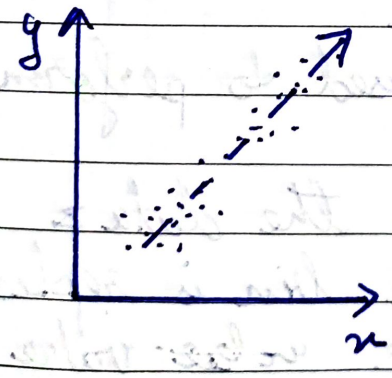
Q1 Scatterplot is a graph that is used to plot data points between two variables. Its a 2 dimensional plot where each variable is plotted on either axis.

It is used to represent correlation between data.

- Types of Correlation

i Positive Correlation

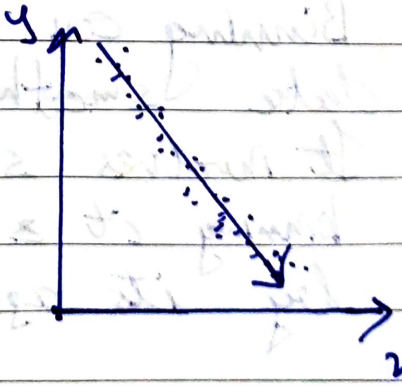
Both variable move in the same direction



x α y

ii) Negative Correlation

Both variables move in opposite direction



x α -y

Q2 We perform data transformation due to the following reasons:

i To remove noise from dataset.

ii Sometimes, discretization can lead to better mining efficiency than using continuous data.

iii New attributes can be constructed that may improve model accuracy

iv Data normalisation maps all the attributes onto a normal scale which is easier to perform analyses on.

Q3 Data smoothing is used to remove noise in the data.
Binning can be used to perform data smoothing.
It involves sorting the data & binning it & each bin is replaced by its aggregate value: value:

i Smoothing by bin means: Replace each value by mean.

ii Smoothing by bin median: Replace each value by median

iii Smoothing by bin boundary: Replace each value by the boundaries of the bin

Q4 $S = \{5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215\}$

i Equal Frequency:

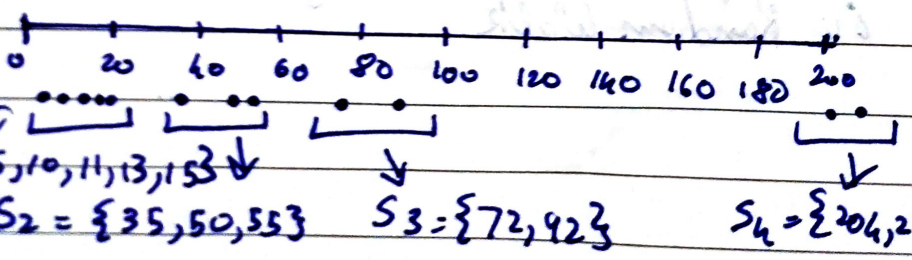
$S_1 = \{5, 10, 11, 13\}$
$S_2 = \{15, 35, 50, 55\}$
$S_3 = \{72, 92, 204, 215\}$

} Size: 4

ii Equal Width

$S_1 (0-49) : \{5, 10, 11, 13, 15, 35\}$
$S_2 (50-99) : \{50, 55, 72, 92\}$
$S_3 \{100-149) : \{\}$
$S_4 (150-199) : \{\}$
$S_5 (200-249) : \{204, 215\}$

iii Clustering



$S_1 = \{5, 10, 11, 13, 15\}$
$S_2 = \{35, 50, 55\}$
$S_3 = \{72, 92\}$
$S_4 = \{204, 215\}$

Q5

a Bin Depth = 3

$S_1 = \{13, 15, 16\} \rightarrow 15$

$S_2 = \{16, 19, 20\} \rightarrow 18$

$S_3 = \{20, 21, 22\} \rightarrow 21$

$S_4 = \{22, 25, 25\} \rightarrow 24$

$S_5 = \{25, 25, 30\} \rightarrow 27$

$S_6 = \{33, 33, 35\} \rightarrow 34$

$S_7 = \{35, 35, 35\} \rightarrow 35$

$S_8 = \{36, 40, 45\} \rightarrow 40$

$S_9 = \{46, 52, 70\} \rightarrow 56$

b This technique is used to smooth the data & remove any noise

b Outliers can be determined by box plot & whisker chart

c Other methods of smoothing data:

i Binning by boundary

ii Binning by median

iii Exponential Smoothing

iv Random Walk

Q6

a $min = 13$, $max = 70$

$$x = 35$$

$$x' = \frac{V - min}{max - min} = \frac{35 - 13}{70 - 13}$$

$$= 0.385$$

b $\sigma = 12.94$

$$\mu = \frac{\Sigma x_i}{n} = \frac{13 + 15 + 16 \ldots + 52 + 20}{27}$$

$$= 29.96$$

$$x_2 = \frac{x - \mu}{\sigma} = \frac{35 - 29.96}{12.94}$$

$$= 0.35$$

c $x_d = \frac{x}{10^j}$ $(j = 2)$

$$= \frac{35}{100} = 0.35$$

d Given data has a lot of values near the central region. Assuming the data is normally distributed we can use Z-score normalisation to preserve the distribution.