# Parampreet Singh CPSC 8650

## Assignment - 2

3.3 Given data in ascending order

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) Below are the steps to smooth the above data using smoothing by bins with a bin depth of 3.

Step 1 : Need to sort the data
Step 2 : Divide the data into equal bins of depth 3
Given depth size = 3
Total bins = $\frac{\text{Total values}}{\text{depth size}} = \frac{27}{3} = 9$ bins

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

Step 3 : Cal. arithmetic mean of each bin

Bin 1 : $\frac{13+15+16}{3} = \frac{44}{3} = 14.67$

Bin 2 : $\frac{16+19+20}{3} = \frac{55}{3} = 18.33$

Bin 3 = $\frac{30+21+22}{3} = \frac{63}{3} = 21$

Bin 4 : $\frac{22+25+25}{3} = \frac{72}{3} = 24$

Bin 5 : $\frac{25+25+30}{3} = \frac{80}{3} = 26.67$

Bin 6 : $\frac{33+33+35}{3} = \frac{101}{3} = 33.67$

Bin 7 : $\frac{35+35+35}{3} = \frac{105}{3} = 35$

Bin 8 : $\frac{36+40+45}{3} = \frac{121}{3} = 40.33$

Bin 9 : $\frac{46+52+70}{3} = \frac{168}{3} = 56$

Step 4 : Now, replace each value in the bin with the respective bin mean

Bin 1 : 14.67, 14.67, 14.67
Bin 2 : 18.33, 18.33, 18.33
Bin 3 : 21, 21, 21
Bin 4 : 24, 24, 24
Bin 5 : 26.67, 26.67, 26.67
Bin 6 : 33.67, 33.67, 33.67
Bin 7 : 35, 35, 35
Bin 8 : 40.33, 40.33, 40.33
Bin 9 : 56, 56, 56

b) outliers can be identified/detected by clustering methods, generally those are extreme values from the mean. So, in clustering, they organise the given data into similar values into groups or clusters, values that fall outside of the set of clusters may be considered outliers.

c) Data Smoothing Methods -
There are many methods used for data smoothing other than bin smoothing by means method. We have couple of others

methods -
1) Equal frequency bins
2) Smoothing by bin boundaries.

So the equal frequency bins divide the data into equal depth i.e having equal frequency bins and apply the techniques to smooth the data.
Other than these binning methods there are regression techniques to smooth the data. These can be linear or multiple regression.

3.5 (a) Min max Normalization :-
linear transformation is performed on the original data and the minimum & maximum values from data is fetched & each value is replaced according to the formula and range to $[new\_min_A, new\_max_A]$

$$v' = \frac{v - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A$$

Here, $v$ = original value

(b) Z score Normalization :-
This technique is used to normalize

the data values based on mean & standard deviation.

Here the value range is

$$\left[\frac{min_A - \bar{A}}{\sigma_A}, \frac{max_A - \bar{A}}{\sigma_A}\right]$$

$$v' = \frac{v - \mu_A}{\sigma_A}$$

where $\mu_A$ : mean

$\sigma_A$ : S td

c) z score normalization using the mean absolute deviation :-

This method is similar to the above z ~~score~~ zscore normalization but by replacing the standard deviation with the mean absolute deviation of A represented as $S_A$.

$$S_A = \frac{1}{n}\left(|v_1 - \bar{A}| + |v_2 - \bar{A}| + |v_3 - \bar{A}| + \cdots + |v_n - \bar{A}|\right)$$

$$v' = \frac{v - \bar{A}}{S_A}$$ The value-range is

$$\left[\frac{min_A - \bar{A}}{S_A}, \frac{max_A - \bar{A}}{S_A}\right]$$

d) Decimal scaling :-

This method is used for normalization by decimal scaling which normalizes by

moving the decimal point of values of attribute A. The value range is

$$\left[ \frac{min_A}{10^j} , \frac{max_A}{10^j} \right]$$

$v' = \frac{v}{10^j}$ where $j$ is the smallest integer such that $\max |v'| < 1$

3.6 Given data,

200, 300, 400, 600, 1000

(a) Min-max normalization by setting min = 0 and max = 1

$$v_i' = \frac{v_i - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A$$

So from the above data $max_A = 1000$

$min_A = 200$

$new\_max_A = 1$

$new\_min_A = 0$

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

So, $v_1' = \left(\frac{v_1 - 200}{800}\right)(1-0)+0$

$= \frac{200-200}{800} = 0$

$v_2' = \frac{300-200}{800} = 0.125$

$v_3' = \frac{400-200}{800} = 0.25$

$v_4' = \frac{600-200}{800} = 0.5$

$v_5' = \frac{1000-200}{800} = 1$

So the normalized data are: 0, 0.125, 0.25, 0.5, 1

b) z score normalization :-

$$v' = \frac{v - \mu_A}{\sigma_A}$$

$\mu_A = \frac{1}{5}(200 + 300 + 400 + 600 + 1000)$

$= 500$

$\sigma_A = \sqrt{\frac{1}{5}\left[(200-500)^2 + (300-500)^2 + (600-500)^2 + (400-500)^2 + (1000-500)^2\right]}$

$\sigma_A = 282.8427$

$\sigma_A \approx 282.843$

$$v_1' = \frac{200 - 500}{282.843} = -1.0606$$

$$v_2' = \frac{300 - 500}{282.843} = -0.7071$$

$$v_3' = \frac{400 - 500}{282.843} = -0.3535$$

$$v_4' = \frac{600 - 500}{282.843} = 0.3535$$

$$v_5' = \frac{1000 - 500}{282.843} = 1.7677$$

So, normalized data :

$-1.0606, -0.7071, -0.3535, 0.3535, 1.7677$

c) z score normalization using mean absolute deviation instead of standard deviation

We know, $\mu_A = 500$ & $n = 5$

Mean Absolute Deviation ($s_A$) $= \frac{1}{n}(|v_1 - \bar{\mu}| + |v_2 - \bar{\mu}| + \cdots |v_n - \mu|)$

$$= \frac{1}{5}(|200-500| + |300-500| + |400-500| + |600-500| + |1000-500|)$$

$$= \frac{1}{5}(300 + 200 + 100 + 100 + 500)$$

$$= 240$$

$$v_i' = \frac{v_i - \mu_A}{\sigma_A}$$

$$v_1' = \frac{200-500}{240} = -1.25$$

$$v_2' = \frac{300-500}{240} = -0.8333$$

$$v_3' = \frac{400-500}{240} = -0.4166$$

$$v_4' = \frac{600-500}{240} = 0.4166$$

$$v_5' = \frac{1000-500}{240} = 2.0833$$

Normalized data:

$-1.25, -0.8333, -0.4166, 0.4166, 2.0833$

d) Normalization by decimal scaling:

Need to $j = 4$

$$v_1' = \frac{200}{10000} = 0.02$$

$$v_2' = \frac{300}{10000} = 0.03$$

$$v_3' = \frac{400}{10000} = 0.04$$

$$v_4' = \frac{600}{60000} = 0.06$$

$$v_5' = \frac{6000}{60000} = 0.01$$

3.7) (a) Min-max normalization -
Transform value of 35 for age onto the range [0.0, 1.0]

$$v_i' = \frac{v - min_A}{max_A - min_A} (new\_max_A - new\_min_A) + new\_min_A$$

$min_A = 13$ $max_A = 70$

$v = 35$

$$v_1' = \frac{35 - 13}{70 - 13} (1 - 0) + 0$$

$$= \frac{22}{57} = 0.3859$$

(b) z score Normalization :

$v = 35$ std = 12.94 yrs

$$\bar{A} = \frac{1}{27}(13 + 15 + \cdots + 70) = \frac{809}{27} = 29.96$$

$\sigma_A = 12.94$

$$v_i' = \frac{v_i - A}{\sigma_A} = \frac{35 - 29.62}{12.702} = 0.3966$$

(c) Normalization by decimal scaling

$$v_i' = \frac{v_j}{10^j} = \frac{35}{10^2} = 0.35$$

d) I would prefer normalization by decimal scaling and my preffered method because it is simple to compute the normalized form of data and it can also be used to simply interpret the given data set in its normalized form. For determining the min-max values in the min max normalization I believe we can retrive the outbound errors. Z score normalization measures the distance b/w values, however it does not increase the information value of characteristics. As a result, decimal scaling is my preffered method of scaling.

3.8 Z-score normalization

(a) $v_i' = \frac{v_i - \bar{A}}{\sigma_A}$

Age Data

$\mu_A = \frac{\sum x_i}{n} = \frac{1}{18}(23+23+\cdots+61) = 46.44$

$$\sigma_A = \sqrt{\frac{1}{18}(23-46.44)^2 + \cdots + (61-46.44)^2}$$

$$\sigma_A = 12.86$$

$$v_1' = \frac{23-46.44}{12.86} = -1.82$$

$$v_2' = \frac{23-41.44}{12.86} = -1.82$$

$$v_3' = \frac{27-46.44}{12.86} = -1.51$$

So, z score normalized age data

$-1.82, -1.82, -1.51, -1.51, -0.58, -0.42, 0.04,$
$0.20, 0.28, 0.43, 0.59, 0.59, 0.74, 0.82, 0.90,$
$0.90, 1.05, 1.13$

y. fat data

$$v_i' = \frac{v_i - \bar{B}}{\sigma_B}$$

$$\mu_A = \Sigma B_i = \frac{1}{18}(9.5 + \cdots + 35.7) = 28.783$$

$$\sigma_A = \sqrt{\frac{1}{18}(9.5-28.78)^2 + \cdots (35.7-28.78)^2}$$

$$\sigma_A = 8.99$$

$$v_1' = \frac{9.5-28.78}{8.99} = -2.14$$

$$v_2' = \frac{21.5 - 28.78}{8.99} = -0.25$$

So, z score for normalized data

$-2.14, -0.25, -2.33, -1.22, 0.29, -0.32, -0.15, -0.18, 0.27, 0.65, 1.53, 0.00, 0.51, 0.16, 0.59, 0.46, 1.38, 0.77$

b) Correlation Coefficient

So from the above calculations & given data

$\mu_{age} = 46.44$ $\quad \sigma_{age} = 12.86$

$\mu_{fat} = 28.783$ $\quad \sigma_{fat} = 8.99$

Steps to calculate

Step 1: Cal. mean
Step 2: Cal. std
Step 3: for all values $(x, y)$ data set multiply $(x - \bar{x})(y - \bar{y})$
Step 4: Do step 3 for all values of data & find result
Step 5: Divide step 4 by $\sigma_x \times \sigma_y$
Step 6: divide result by $n-1$

∴ the correlation coefficient

$$r_{x,y} = \frac{\sum_{i=0}^{n} (x_i y_i) - \bar{x}\bar{y}}{(n-1)(\sigma_x \cdot \sigma_y)}$$

S1: $\bar{x} = 46.44$ , $\bar{y} = 28.783$

S2: $S_x = 12.86$ , $S_y = 8.99$

S3: $(x - \bar{x})(y - \bar{y})$

S4: Sum of result = 1700.3334

S5:

$$r_{x,y} = \frac{1700.3}{(18-1)(12.86 \times 8.99)} = 0.865$$

As the $r_{x,y}$ is positive these 2 data set values are positively/directly correlated.

Co-variance:

$$\text{Cov}(A,B) = E(A-\bar{A})(B-\bar{B})$$

$$= \frac{(23 \times 95 + 23 \times 26.5 + \cdots + 61 \times 35.7) - (\bar{x} \cdot$$

$$= 1431.289 - (46.44)(28.78)$$

$$= 94.4607$$

$\text{Cov}(A,B) > 0$ positively correlated

3.9. Sorted sales price values

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

Need to partition into 3 bins

(a) Equal frequency partitioning :-

d = 4

Bin 1 : 5, 10, 11, 13
Bin 2 : 15, 35, 50, 55
Bin 3 : 72, 92, 204, 215

(b) Equal width partitioning :-

max = 215
min = 5

Bin 1 : 5, 10, 11, 13, 15, 35, 50, 55, 72
Bin 2 : 92
Bin 3 : 204, 215

(b) Clustering :-

It is the data that is grouped to similarity values and that is closest to means

Then we divide them into groups below with 3 bins :-

Bin 1 : 5, 10, 11, 13, 15, 35
Bin 2 : 50, 55, 72, 92
Bin 3 : 204, 215