Ujjwal Abhishek
18CSE132
Group-2

1. Give Brief description of the following (c)

(a) Binning -

Data binning is also called as discrete binning or bucketing. It is a data pre-processing technique used to reduce the effects of minor observation errors. The original data values which fall into a given small interval, a bin, are replaced by a value representative of that interval, often the central value. It is a form of quantization.

There are 3 methods.

(i) Partition into equal-frequency
(ii) Smoothing by bin means
(iii) Smoothing by bin boundaries.

(b) Regression -

Regression is a data mining technique used to predict a range of numeric values (also called continuous values), given a particular dataset. For example, regression might be used to predict the cost of a product or service, given other variables.

Regression is used across multiple industries for business and marketing planning, financial forecasting, environmental modeling and analysis of trends.

(c) Clustering:-

Clustering is the process of making a group of abstract objects into classes of similar objects.

→ a cluster of data objects can be treated as one group.

→ while doing cluster analysis, we first partition the set of data into groups based on data similarity and then assign the labels to the groups.

(d) Smoothing:-

Data smoothing uses an algorithm to remove noise from a data set, allowing important patterns to stand out.

It can be used to predict trends, such as those found in securities prices.

Different data smoothing models include the random method, random walk and the moving average.

(e) Generalization:-

A process that abstracts a large set of task-relevant data in a database from a low conceptual level to higher ones.

Data generalization is a summarization of general features of

objects in a target class and produce what is called characteristic rules.

(f) **Aggregation:-**

Data aggregation is a type of data and information mining process where data is searched, gathered and presented in a report-based, summarized format to achieve specific business objectives or processes and/or conduct human analysis.

Data aggregation may be applied manually or through specialized software.

2. Explain various normalization techniques?

ans:-

(1) min-max normalization: to $[new\_min_A, new\_max_A]$

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

→ ex- Let income ranges \$12,000 to \$98,000 normalized [0.0, 1.0].

Then, \$73,600 is mapped to $\frac{73,600 - 12,000}{98,000 - 12,000} (1.0) + 0$

$= 0.716$

(ii) Z-score normalization:

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

μ - mean
σ - standard deviation

→ ex- Let $\mu = 54,000$, $\sigma = 16,000$

Then, $\frac{73,600 - 54,000}{16,000} = 1.225$

(iii) normalization by decimal scaling

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $Max(|v'|) < 1$

3. Give some examples of data preprocessing techniques?

ans:-

(i) Data cleaning:-
→ fill in missing values
→ Smooth noisy data
→ identify or remove outliers
→ resolve inconsistencies

(ii) Data integration:-
→ integration of multiple databases, data cubes, or files.

(iii) Data reduction:-
→ Dimensionality reduction

→ Numerosity reduction
→ Data compression

(iv) Data Transformation and data discretization
→ normalization
→ Concept hierarchy generation

4. 10, 9, 10, 8, 11, 12, 21, 19, 23, 22 calculate (Mean, Median, Mode, Variance, Standard deviation, Chi-Square Test, Correlation coefficient (1-10), Covariance, min-max normalization by setting min = 0, (U-21), Z-Score normalization, normalization by decimal scaling, Binning, Box Plot, PCA).

Sol^{n}:-

X	$(x_i - \bar{X})^2$	Expected (E)	$(O-E)^2/E$
8	42.25	145/8 = 18.125	5.65
9	30.25	145/9 = 16.11	3.14
10	20.25	145/10 = 14.5	1.40
10	20.25	145/10 = 14.5	1.40
11	12.25	145/11 = 13.18	0.36
12	6.25	145/12 = 12.08	0.0005
19	20.25	145/19 = 7.63	16.94
21	42.25	145/21 = 690	28.81
22	56.25	145/22 = 6.59	36.03
23	72.25	145/23 = 6.30	44.26
$\bar{X} = 14.5$	Variance = 322.25/9 = 35.8		$X^2 = 137.99$

median = 11.5

mode = 10

standard deviation = $\sqrt{\text{variance}}$

= 5.98

chi^2 = 137.99

Normalization

a) Min-max normalization :-

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

given: $v = 21$

$$\therefore v' = \frac{21 - 8}{23 - 8} (1 - 0) + 0$$

$$= \frac{13}{15} = 0.86$$

b) z-score normalisation :-

$$v' = \frac{v - \mu_A}{\sigma_A} \qquad \left[\begin{matrix} \mu_A = 14.5 \\ \sigma_A = 5.98 \end{matrix}\right]$$

$$= \frac{21 - 14.5}{5.98} = \frac{6.5}{5.98} = 1.08$$

c) normalization by decimal scaling:

$j = 02$

$$\therefore v' = \frac{v}{10^j} = \frac{21}{10^2} = 0.21$$

Binning

Step 1: Sort the data

8, 9, 10, 10, 11, 12, 19, 21, 22, 23

(i) Partition into equal-frequency bins.

- Bin 1 8, 9, 10, 10, 11
- Bin 2 12, 19, 21, 22, 23

(ii) Smoothing by bin means.

- Bin 1 10, 10, 10, 10, 10
- Bin 2 19, 19, 19, 19, 19

(iii) Smoothing by bin boundaries

- Bin 1 8, 8, 11, 11, 11
- Bin 2 12, 23, 23, 23, 23

Box-Plot

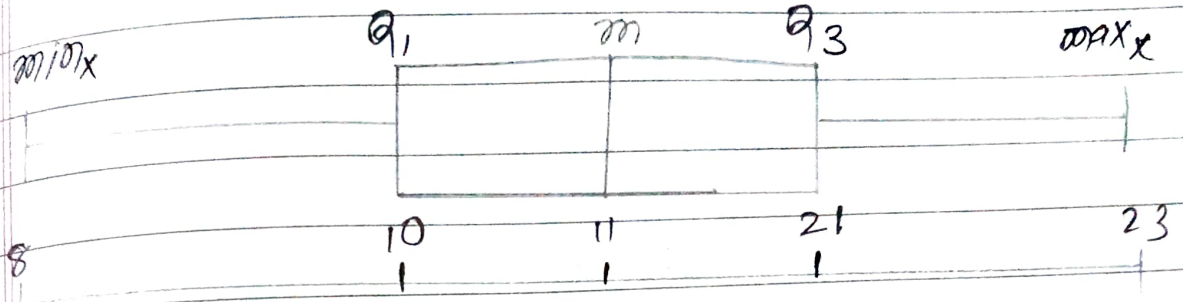
$Q_1 = \frac{10 \times 25}{100} = 2.5 = 3^{rd}$ value $= 10$

$Q_2 = \frac{10 \times 50}{100} = 5^{th}$ value $= 11$

$Q_3 = \frac{10 \times 75}{100} = 7.5 = 8^{th}$ value $= 21$

Range = 23 - 8 = 15

IQR = 21 - 10 = 11



PCA

x	y	$x_i - \bar{x}$	$y_i - \bar{y}$
10	1	-4.5	-4.5
9	2	-5.5	-3.5
10	3	-4.5	-2.5
8	4	-6.5	-1.5
11	5	-3.5	-0.5
12	6	-2.5	0.5
21	7	6.5	1.5
19	8	4.5	2.5
23	9	8.5	3.5
22	10	7.5	4.5
$\bar{x} = 14.5$	$\bar{y} = 15.5$		

Variancey = 82.5/9 = 9.16

$\sigma_y = \sqrt{9.16} = 3.026$

$$r_{x,y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{(n-1)\sigma_A \sigma_B}$$

$$= \frac{16.17}{5.98 \times 3.026} \quad \text{covariance}$$

$$= 0.893$$

covariance $(x, x) = 35.83$

covariance $(x, y) = 16.17$

covariance $(y, x) = 16.17$

covariance $(y, y) = 9.17$

$$C = \begin{bmatrix} 35.83 & 16.17 \\ 16.17 & 9.17 \end{bmatrix}$$

$$C - \lambda I = 0$$

$$\begin{bmatrix} 35.83 & 16.17 \\ 16.17 & 9.17 \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 0$$

$$\Rightarrow (35.83 - \lambda)(9.17 - \lambda) - (16.17)^2 = 0$$

$$\Rightarrow \lambda^2 - 45\lambda - 27.18 = 0$$

$$\Rightarrow \lambda_1 = -0.596$$

$$\lambda_2 = 45.596$$

Taking, $\lambda = -0.596$

$$\begin{bmatrix} 35.83 & 16.17 \\ 16.17 & 9.17 \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} = -0.596 \begin{bmatrix} x_1 \\ y_1 \end{bmatrix}$$

$\Rightarrow 35.83X_1 + 16.17Y_1 = -0.596X_1$

$\Rightarrow 36.42X_1 + 16.17Y_1 = 0 \quad -(1)$

&, $16.17X_1 + 9.17Y_1 = -0.596Y_1$

$\Rightarrow 16.17X_1 + 9.766Y_1 = 0 \quad -(11)$

from eq^n. (1),

$36.42X_1 = -16.17Y_1$

$\Rightarrow X_1 = -0.44Y_1$

$\begin{bmatrix} -0.44 \\ 1 \end{bmatrix} = \sqrt{1+(0.44)^2} = 1.09$

$\begin{bmatrix} -0.44/1.09 \\ 1/1.09 \end{bmatrix} = \begin{bmatrix} -0.403 \\ 0.91 \end{bmatrix}$ Ans.

---

taking $\lambda = 45.596$

$\begin{bmatrix} 35.83 & 16.17 \\ 16.17 & 9.17 \end{bmatrix} \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix} = 45.596 \begin{bmatrix} X_2 \\ Y_2 \end{bmatrix}$

$\Rightarrow 35.83X_2 + 16.17Y_2 = 45.596X_2$

$\Rightarrow 9.76X_2 = 16.17Y_2$

$\Rightarrow X_2 = 1.656Y_2$

$\begin{bmatrix} 1.656 \\ 1 \end{bmatrix} = \sqrt{(1.656)^2+1^2} = 1.93$

$= \begin{bmatrix} 1.656/1.93 \\ 1/1.93 \end{bmatrix} = \begin{bmatrix} 0.85 \\ 0.91 \end{bmatrix}$ Ans