20/2/23
Monday

Module-2

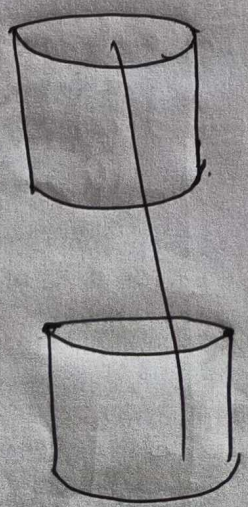
Data prepnocemij

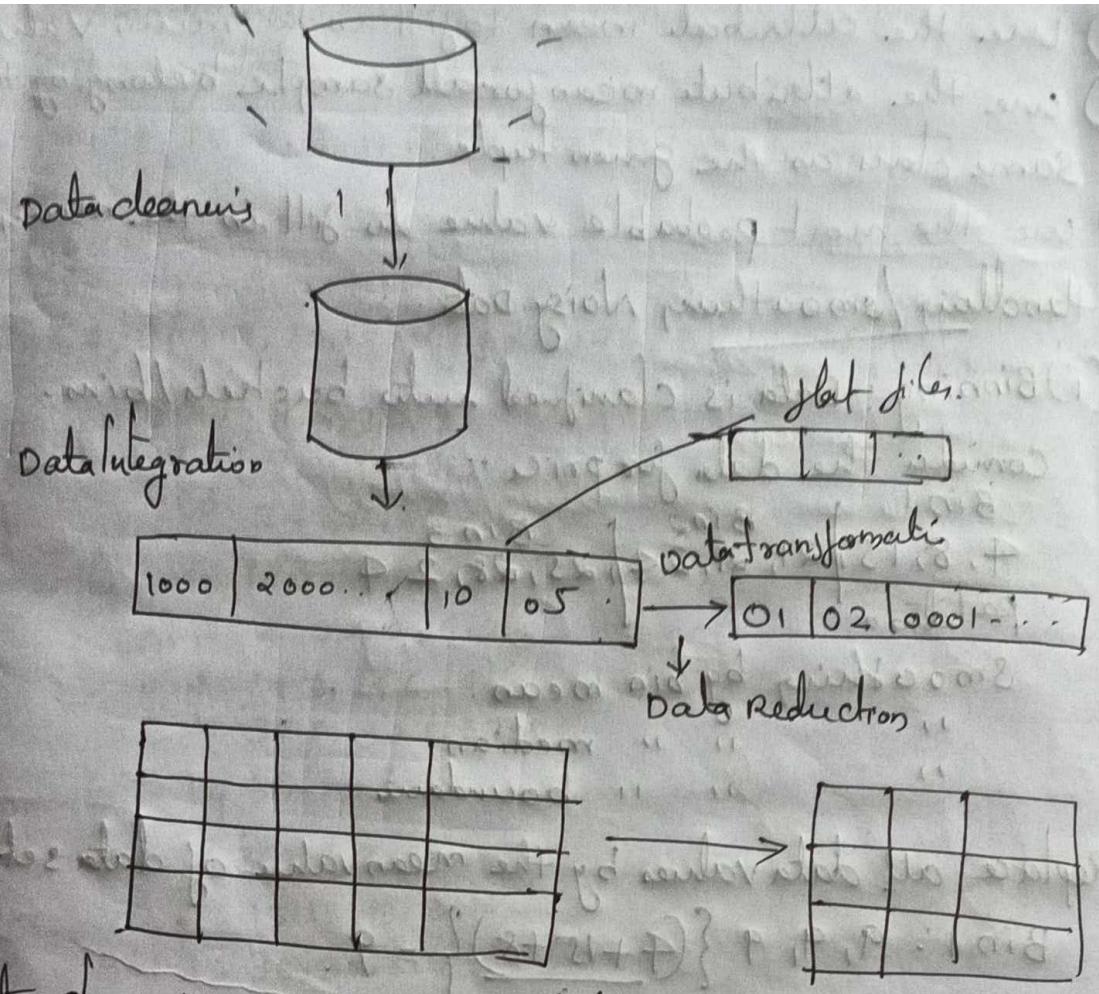
3 types of data
Incomplete
Noisy
Inconsistent

why data prepnocemij
- Real world data
  - Incomplete - malfunctioning of data.
  - Noisy - unwanted data error.
  - Inconsistent
    - Naming convemon : same data and different column names
    - Normalization

+ Major steps in data prepnocemij

(i) Data cleaning
(ii) Data integration
(iii) Data tranformation
(iv) Data reduction





Data cleaning : Remove noisy data.
filling moving.

Data transformation : To convert into common format.
(normalization)

Data reduction : To reduce it size after data transformation.

23/2/23.
Thursday

Data cleaning : filling the missing value, smoothing noisy data identifying or remove outlier, resolving inconsistencies

I Handling missing values

1) Ignore the tuple.
Method is not very effective

2) Filling in the missing values manually.
This approach is time consuming. May not be feasible.

3) use a global constant to fill in the missing value.

4) Use the attribute mean to fill in the missing value
5) Use the attribute mean for all samples belonging to the same class as the given tuple.
6) Use the most probable value to fill in the missing value

II Handling / smoothing Noisy Data:

(i) Binning: Data is classified into buckets / bins.

Consider the data for price.

Bin 1 | Bin 2 | Bin 3

4, 8, 15 | 21, 21, 24 | 25, 28, 34

Partition

Smoothing by bin mean.

" " median.

" " boundaries

Replace all data values by the mean value of data set each bin.

Bin 1: 9, 9, 9 $\left\{\frac{(4+15+8)}{3}\right\} = 9$.

Bin 2: 21, 21, 24 = 22.

22, 22, 22

Bin 3: 25, 28, 34.

29, 29, 29.

Smoothing by median.

Bin 1: 8

Bin 2: 21

Bin 3: 28.

Bin 1: 8, 8, 8.

" 2: 21, 21, 21.

" 3.: 28, 28, 28.

Smoothing by bin boundaries.

4, 4, 15

21, 21, 24

25, 25, 34.

Apply smoothing techniques by min mean, median, boundaries on the following data.

[3, 7, 24, 50, 3, 9, 11, 12, 34, 26, 17, 23, 1, 8, 30]

bin size: 5.

sort

[2, 3, 7, 8] [9, 11, 13, 17, 23] [24, 26, 30, 34, 50]

Bin 1: 1, 2, 3, 7, 8

" 2: 9, 11, 13, 17, 23.

" 3: 24, 26, 30, 34, 50.

smoothing by bin means.

Bin 1: 4.2, 4.2, 4.2, 4.2.

" 2: 14.6, 14.6, 14.6

" 3: 32.8, 32.8, 32.8

smoothing by bin median.

Bin 1: 3, 3, 3, 3, 3.

" 2: 13, 13, 13, 13, 13

Bin 3: 30, 30, 30, 30, 30.

smoothing by bin boundaries

Bin 1: 1, 1, 1, 8, 8

Bin 2: 9, 9, 9, 23, 23

Bin 3: 24, 24, 24, 50, 50.

Regression (prediction)

It is also used for noise recovery.

Types of regression

Simple linear.

Multiple

Polynomial

Logistic

25/2/23 Saturday

Data cleansing as a process.

Step 1: Discrepancy detection

2: Data transformation.

Discrepancy

Human error in data entry

Avoid discrepancy

1) Using metadata.

Commercial tools.

Data scrubbing tools.

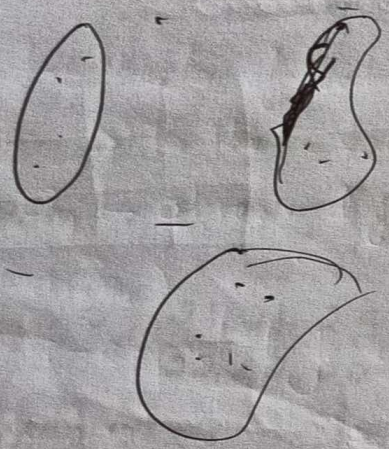
Data auditing tools.

Data transformation

Data Migration tool

ETL (Extraction / Transformation / loading).

Clustering



II Data Integration

The data integration is the process of combines data from multiple sources into a coherent data store.

Issues in data Integration

1. Entity identification problems
2. Data Redundancy problems.
3. Detection and resolution of data value conflicts.

1. Entity Identification problem

Eg: Data analyst or computer cannot ensure that customer_id in one database and cust_number in another refer to the same attribute. ie metadata

helps to avoid errors in schema entity identification problem

Data Redundancy (repetetion of data)

using normalization to avoid redundancy. Derived attribute are also examples of redundancies.

Solution to data Redundancy

1) Correlation Analysis

For numerical attributes we can evaluate the correlation b/w 2 attributes A & B. known as pearson's product moment coefficient (named after its inventor, karl pearson)

$$r_{A,B} = \frac{\sum_{i=1}^{N} (a_i - \bar{A})(b_i - \bar{B})}{N \sigma_A \sigma_B}$$

$$= \frac{\sum_{i=1}^{N} (a_i b_i) - N \bar{A} \bar{B}}{N \sigma_A \sigma_B}$$

N - No: of tuples.

a_i & b_i - respective values of A & B in tuple i

$\bar{A}$ & $\bar{B}$ - respective mean values of A & B.

σ_A & σ_B - respective standard deviation.

A, B - attributes.

$r_{A,B}$ - correlation coefficient.

$-1 \leq r_{A,B} \leq 1$.

if $r_{A,B} > 0$, A & B are +vely correlated. which means the value of A increase as the values of B increase.

Highly correlated

if $r_{A,B} = 0$, A & B are independent & there is no correlation b/w them.

If $r_{A,B} < 0$

A & B are negatively correlated, where the values of one attr. increase, the values of the other attribute decrease.

If $r_{A,B} \neq 0$ A & B dependent to each other.

Correlation analysis for categorial attributes

A correlation relationship b/w 2 attributes A, B can be discovered by a x^2 (chi-square) test. Suppose A has c distinct values, namely $a_1, a_2, \ldots c$. B has r distinct values, namely $b_1, b_2, \ldots, b_c$

$$x^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

The data tuples described by A & B can be shown as a contigency table, with the c values of A

o_{ij} – observed frequency (actual count) of the joint event (A_i, B_j)

e_{ij} – expected frequency of (A_i, B_j)

The sum is computed over all of the

$$e_{ij} = \text{count}(A = a_i) \times \text{count}(B = b_j)$$

Problem of categorial analysis

i) Suppose that a group of 1,500 people was surveyed. The gender of each ~~person~~ person was noted. Each person was polled as to whether their preferred type of reading material was fiction or non fiction. Thus, we have 2 attributes, gender and preferred reading. The observed frequency (or count) of each possible joint events is summarized in the contigency, where the numbers in parantheses are the expected frequencies.

		Male	Female	
b_1	Fiction	250 $(1,1)$	200 $(2,1)$	450
b_2	Nonfiction	50 $(1,2)$	1000 $(2,2)$	1050
		300	1200	1500

Hypothesis

Gender and preferred reading are independent each other.

B_{ij} = Total cost

$$e_{ij} = \left(\frac{a_{ij} \times b_{ij}}{N}\right)$$

$$e_{11} = \frac{a_1 \times b_1}{N} = \frac{300 \times 450}{1500} = 90$$

$$e_{21} = \frac{1200 \times 450}{1500} = 360$$

$$e_{12} = \frac{a_1 \times b_2}{N}$$

$$= \frac{300 \times 1050}{1500} = 210$$

$$e_{22} = \frac{a_2 \times b_2}{N} = \frac{1200 \times 1050}{1500} = 840$$

		a_1 Male	a_2 Female	Total
b_1	Fiction	250 $(1,1)$ (90)	200 $(2,1)$ (360)	450
b_2	Nonfiction	50 $(1,2)$ (210)	1000 $(2,2)$ (840)	1050

$$\chi^2 = \sum_{i=1} \sum_{j} \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

$$= \frac{(250-90)^2}{90} + \frac{(200-360)^2}{360} + \frac{(50-210)^2}{210} + \frac{(1000-840)^2}{840}$$

$$= 284.44 + 71.11 + 121.90 + 30.48$$

$$= 507.93$$

degrees of freedom $= (r-1)(c-1)$

$= (2-1)(2-1) = 1$.

If the critical value from the chi-square distribution table is checked against the chi-square. If calculate chi-square value < critical value, hypothesis can be accepted.

otherwise hypothesis rejected.

Critical value corresponding to $d[(r-1)(c-1)]$ with maximum significants levels of 0.001 is 10.828.

507.93 > 10.828 hypothesis can be rejected. ∴ A & B are strongly correlated ie, preferred reading and gender are correlated

Detection and resolution of data value conflicts.

To avoid this attribute functional dependencies and referential constraints in the source system match those in the target system.

28/8/25 Tuesday

Data transformation

The data are transformed into forms appropriate for mining.

Smoothing which works for remove noise from the data. Such techniques include binning, regression, clustering

Aggregation: where summary or aggregation are applied to the data.

eg: the daily sales data may be aggregated so as to compute monthly and annual total amounts. This step is typically used in constructing a data cube for analysis of the data at multiple

Generalization of the data

Generalization of the data, where low-level data are replaced by higher level concepts through the use of concept hierarchies.

eg: categorical attributes, like street, can be generalized to higher level concepts like city or country.

Smoothing and e.t

Attribute construction (feature construction), where new attributes are constructed and added from the given set of attributes to help the mining process.

eg: we may to wish to add the attribute area based on the attribute height & width ft.

* Normalization

where the attribute data are scaled so as to fall with in a small specified range.

Methods for data normalization

Min max normalization.

Z-score "

Min-max normalization.

Performs a linear transformation on the original data.

let min_A and max_A are the minimum and max values of an attribute A.

Maps a value, v_i of A to v' in the range $[new\_min_A$ $new\_max_A]$.

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

It will encounter an "out of bonds" error if a future i/p case for normalization fall outside of the original data range for A.

Q: Suppose that the minimum and max value for the attribute income are $12,000 & $98,000 respectively. We would like to map income to the range [0.0, 1.0]. By min-max normalization, a value of $73,600 for income is transformed to

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

$min_A = 12000$

$max_A = 98000$

$new\ min_A = 0.0$

$new\ max_A = 1.0$

Given $v = 73600$

$$v' = \left[\frac{73600 - 12000}{98000 - 12000}(1.0 - 0.0)\right] + 0.0$$

$$= \frac{61600}{86000} \times 1 + 0 = 0.716$$

~~Z-score or Z-mean normalization.~~

disadvantages

For eg: consider the data 99200.

$$V' = \frac{99200 - 12000}{98000 - 12000}$$

Z-score or Z-mean normalization

Values for an attribute, A, are normalized based on the mean and standard deviation of A.

Maps a value V, of A to V' by computing

$$V' = \frac{V - \bar{A}}{\sigma_A}$$

where $\bar{A}$ and σ_A are the mean & sd. respectively of attribute A

adv

useful when the actual min & max of attribute A are unknown, or when there are outliers that dominate the min-max normalization.

Suppose that the mean & sd. of the values for the attribute income are $54,000 & $16,000 respectively with Z-score normalization a value of 73,600 for income is transformed.

$$V = 73600$$

$$\bar{A} = 54000 \qquad \sigma_A = 16,000$$

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

$$= \frac{73600 - 54000}{16000}$$

$$= \frac{19600}{16000} = 1.225$$

X Normalization by decimal scaling.

Normalizes by moving the decimal point of values of attribute A.

The no: of decimal points moved depends on the max absolute value of A

A value of v; of A is normalized to v' by computing.

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

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

Q. Suppose that the recorded values of A range from -986 to 917. The max absolute value of A is 986. To normalize by decimal scaling, we therefore

divide each value by €1,000 (i.e., $j = 3$), so that -986 normalizes to -0.986 and 917 normalizes to 0.917.

$v' =$

the following

use the 2 method to normalize data

200, 300, 400, 600, 1000

(A) Min-max normalization by setting min=0 max=1.

(B) Z-score normalization.

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

$new\_max_A = 1$

$new\_min_A = 0$.

$V = 200$

$min_A = 200$, $max_A = 1000$.

$$v' = \frac{200 - 200}{1000 - 200} [1 - 0] + 0 = 0$$

$V = 300$

$$v' = \frac{300 - 200}{1000 - 200} [1 - 0] + 0 = \frac{100}{800} = +0.125$$

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

$$= \frac{400 - 200}{800} = \frac{200}{800} = +0.25$$

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

$$= \frac{600 - 200}{800} = \frac{400}{800} = +0.5$$

$$v' = \frac{1000 - 200}{1000 - 200} = \frac{200 - 1000}{800} = \frac{-800}{800}$$

$$= 1$$

normalized data = 0, 0.125, 0.25, 0.5, 1

B) Z-score normalization.

$$v' = \frac{v - \bar{A}}{\sigma_A}$$

$$\bar{A} = \frac{200 + 300 + 400 + 600 + 1000}{5}$$

$$= \frac{2500}{5} = 500$$

σ^2 Variance $= \frac{\Sigma(x-\bar{x}^2)}{N}$

$SD = \sqrt{\frac{\Sigma(x-\bar{x}^2)}{N}}$

σ^2 variance $= \frac{1}{5}\left[(200-500)^2 + (300-500)^2 + (400-500)^2 + \right.$

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

$= \frac{1}{5}\left[90000 + 40000 + 10000 + 10000 + 250000\right]$

$= 80000$

$SD = \sqrt{80000} = 282.84$

$V^1 = \frac{200-500}{282.84} = -1.06$

$V^1 = \frac{300-500}{282.84} = -0.7$

$V^1 = \frac{400-500}{282.84} = -0.35$

$$v' = \frac{600 - 500}{282.84} = 0.35$$

$$v' = \frac{1000 - 500}{282.84} = 1.76$$

Data Reduction

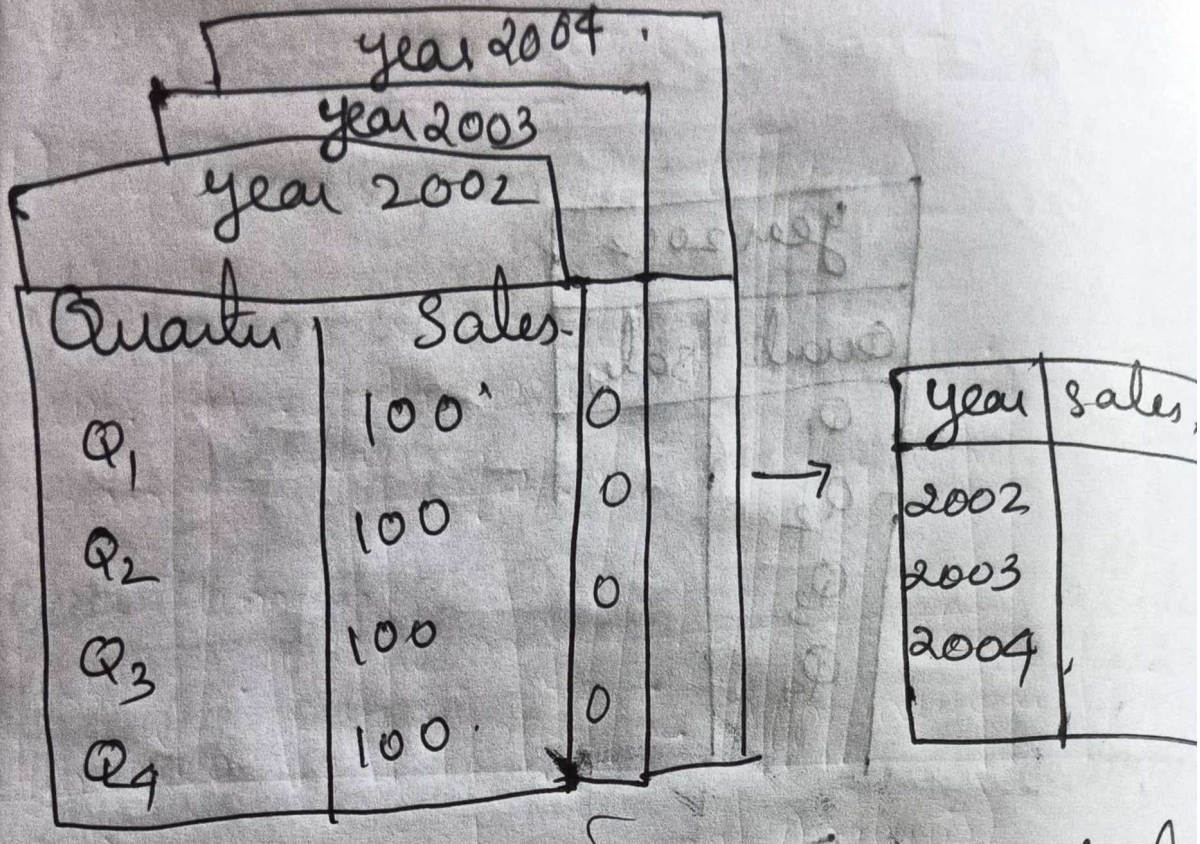
Data reduction is the process of retaining relevant data and eliminating irrelevant attributes from the data set.

Techniques used for data Reduction

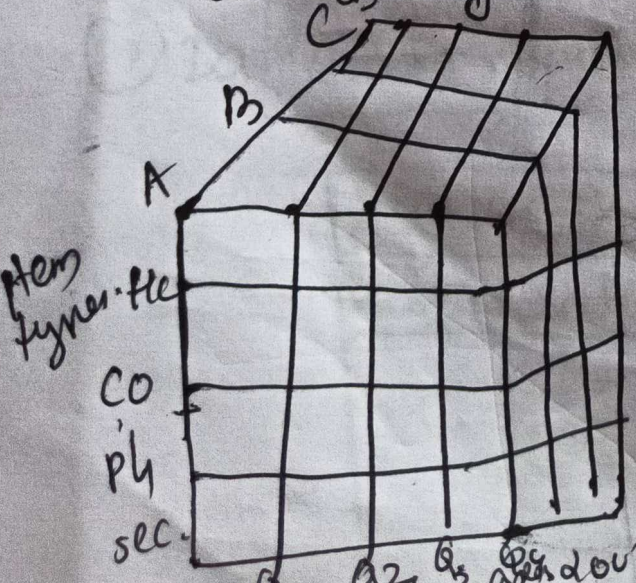
- Data Cube Aggregation
- Attribute Subset Selection
- Dimensionality reduction
- Numerosity reduction
- Discretization & Concept hierarchy generation

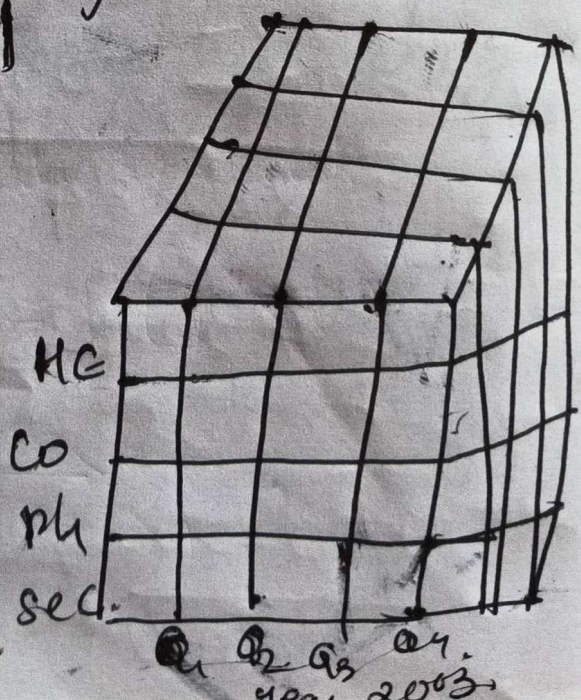
(1) Data Cube Aggregation

year 2002	
Quarter	Sales


year 2004
year 2003
year 2002
Quarter
Sales
Q1
Q2
Q3
Q4
100
100
100
100
0
0
0
0
year
sales
2002
2003
2004


C
B
A
branch
home entertain.
comput
phone
security
500
750
150
50
2002
2003
2004
year
Apex
* Base cuboid
* Apex cuboid
highest level
abstraction level
Represented
highest abstra
level


C
B
A
item
type
HC
CO
PH
SEC
Q1
Q2
Q3
Q4
2002


HC
CO
PH
SEC
Q1
Q2
Q3
Q4
year 2003

13/23 Monday

Dimensionality Reduction

- lossless
- lossy (Approx i/p don't rec original)

1) DWT (Discrete wavelet transform)

Original data used in signal processing. Replaced by discrete wavelet coefficients.

lossless: If the original data can be reconstructed from the ~~reduced~~ o/p - then it is termed as lossless reduction

lossy: If the ~~original data can~~ we can only reconstructed only the approximate i/p.

2) PCA (principal component Analysis)

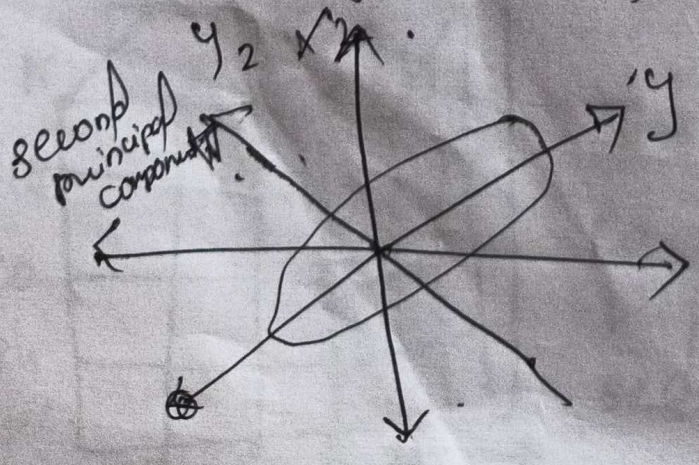
If we have n dimensional data set - we have to calculate k (n dimensional) principal components where k<n.

They are orthogonal vector to original data.

Principal components are unit vectors calculated from original data i/p.

$$z^{\wedge} = \frac{\vec{z}}{|z|}$$

It can be diagramatically represented as.



Numerosity Redⁿ

to reduce the no. of counts.

(i) parametric method.

- Regression.
- Log linear models ⟶ probability concep

strictly based on parameter.
uses some parametric models.

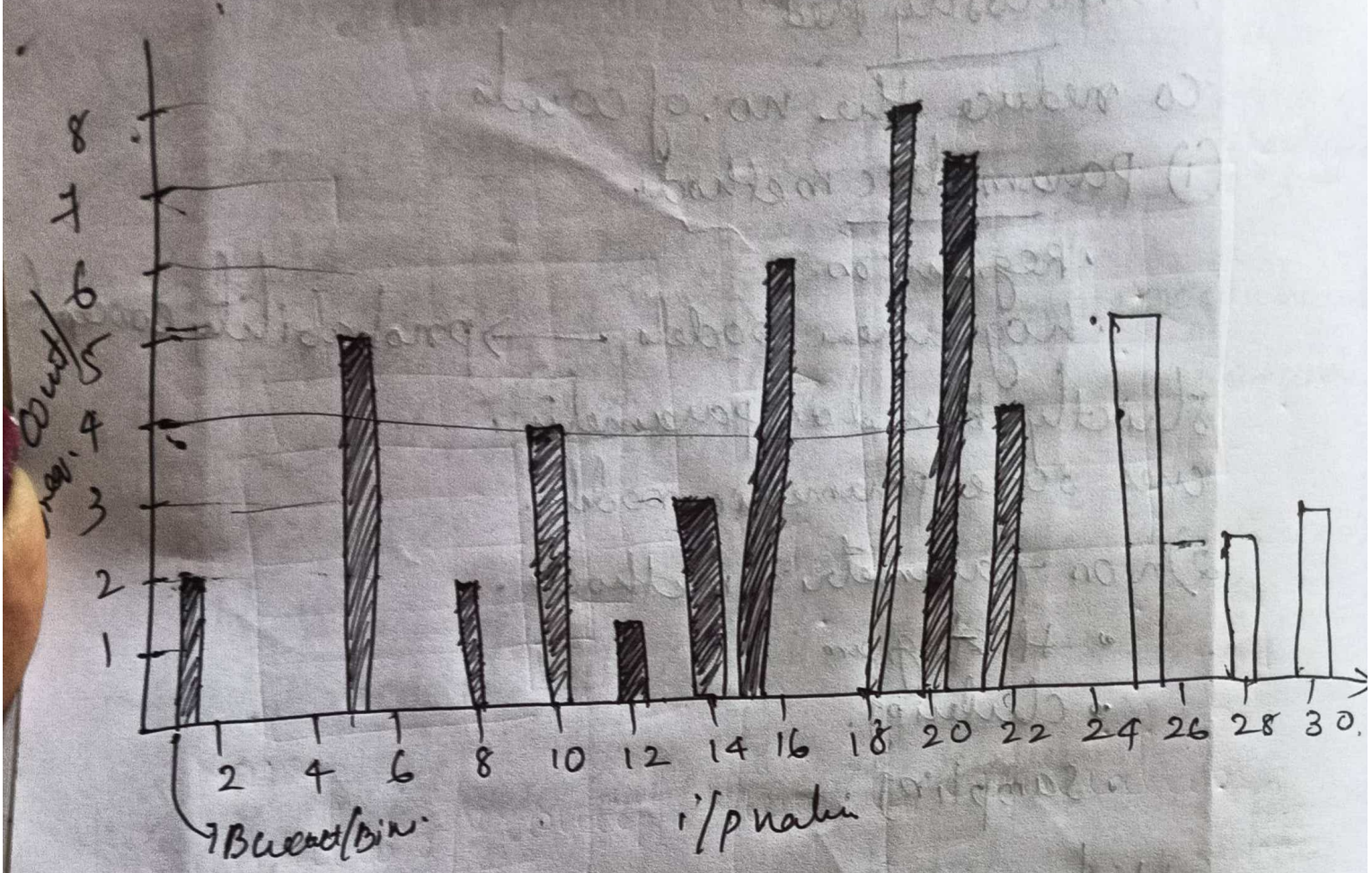
2) Non parametric methods.

- Histogram
- clustering.
- sampling.

• Histogram

convert the data to a histogram

sorted data | 1, 1, 5, 5, 5, 5, 5, 8, 8, 10, 10, 10, 10, 12, 14, 14, 14
15, 15, 15, 15, 15, 15, 18, 18, 18, 18, 18, 18, 18, 18, 20,
unsorted data first will be sorted | 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25
28, 28, 30, 30, 30.



Procedure of partitioning Histograms.

i) (1) Equal width.
↳ width of each bucket is uniform.

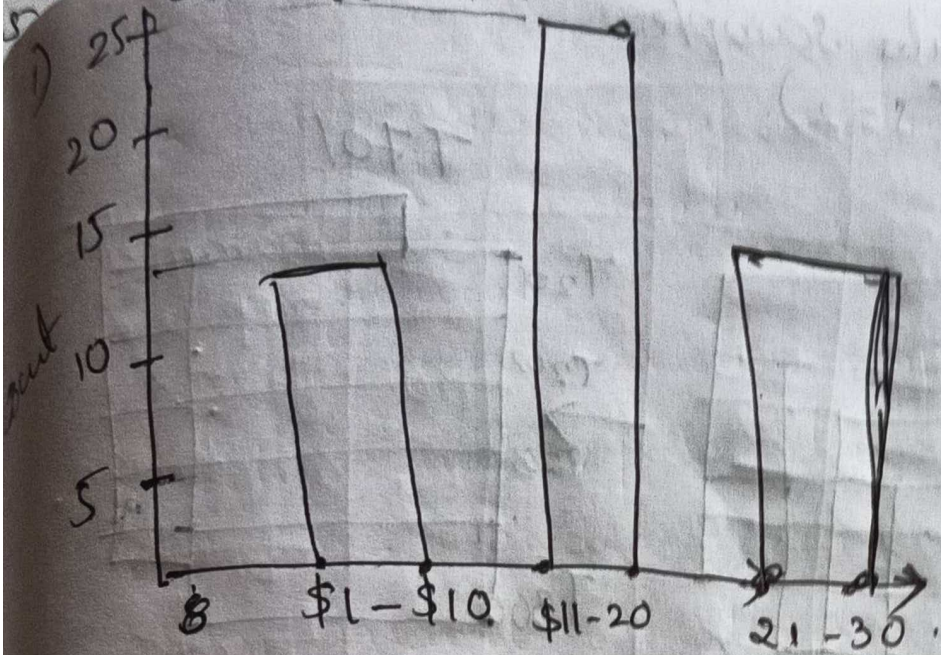
2) Equal frequency.
Frequency of each bucket is a constant
Each bucket contains same no. of elements

3) V-Optimal → Bucket with least variance

4) Max Diff - Diff b/w each pair of adj values.

Aggregation

equal width.

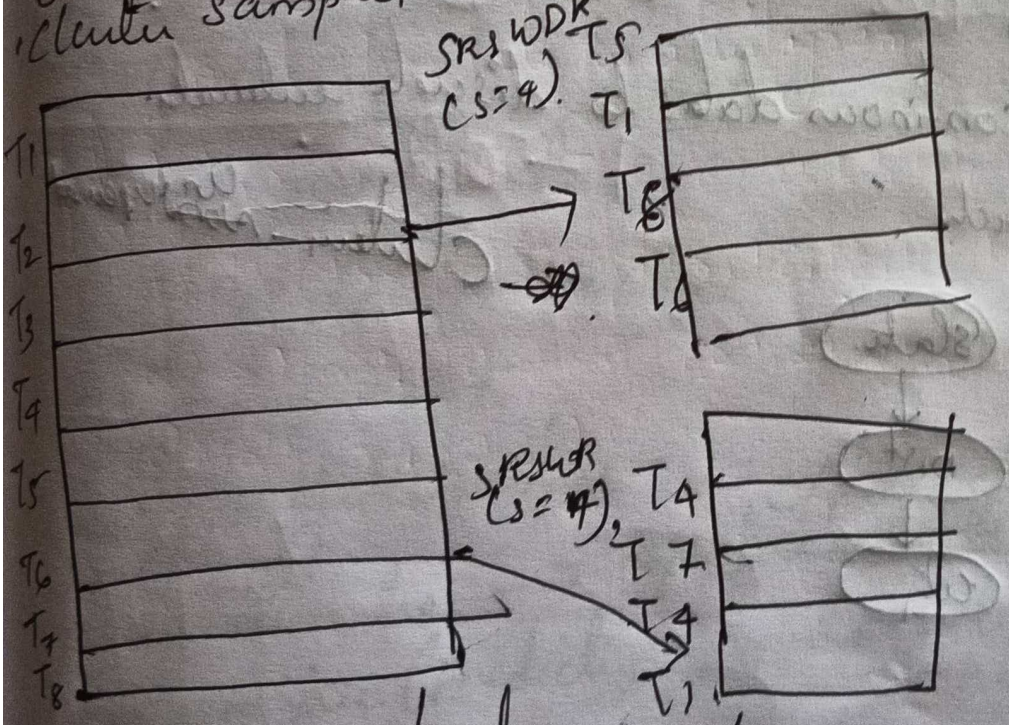


equal width Partitioning.

13/23 Sunday

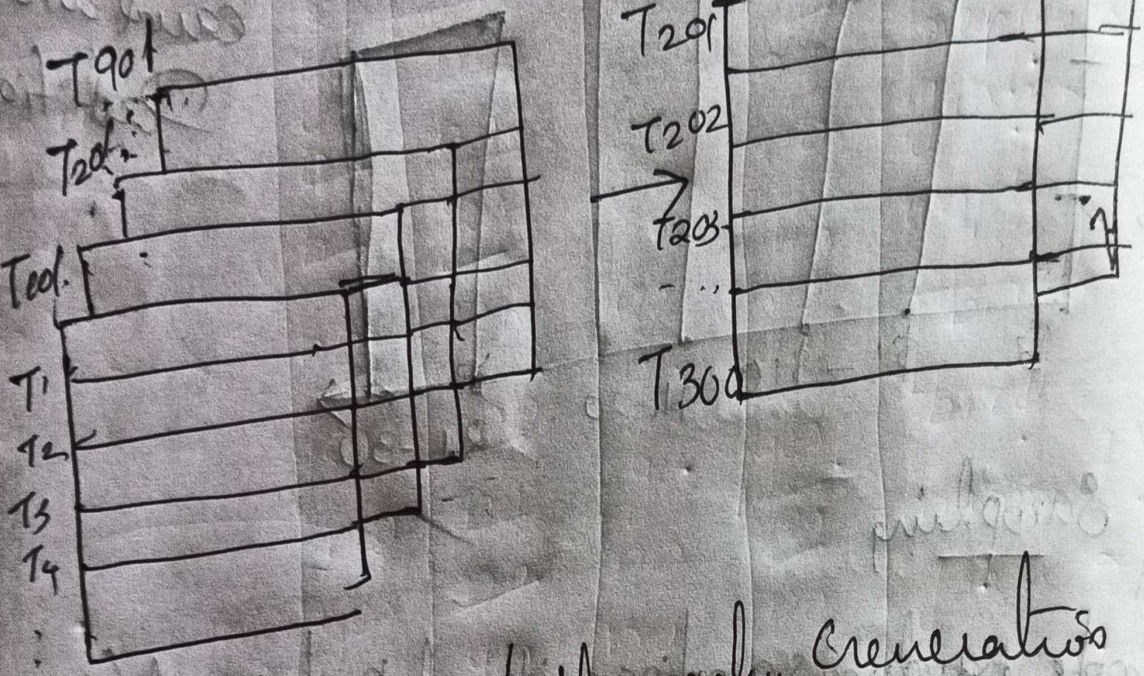
Sampling

- Simple Random sampling without Replacement (SRSWOR)
- Simple Random sampling with Replacing (SRSWR)
- Cluster sample:



Cluster sample

Cluster sample (s=2).



Discretization & concept Hierarchy Generation -

Discretization

Converting continous data to discrete values.

concept hierarchy.

Clustering - Unsupervised