Module 3: Data Pre-processing

Basic Data Mining Tasks / Techniques

They place each purchase into one of four classes

i. Authorized

ii. Ask for further identification before authorization

iii. Do not authorize

iv. Do not authorize but contact police

1. Classification 2. Clustering 3. Association Rules

Classification is the separation or ordering of objects or things into classes. i.e. It maps data into predefined groups or classes.

There are two types of classifications

Supervised

When the classes are created without looking at the data, such type of classification is called Apriori classification.

Unsupervised

When the classes are created after looking at the data, such type of classification is called Posteriori classification.

Ex: Airport Security Screening Station

Credit Card Companies determine authorize customers

List different data mining techniques. Explain KDD process in detail. Dec 2018 [10 Marks]

1

Basic Data Mining Tasks / Techniques

2. Clustering It is similar to classification

The difference is in classification classes are predefined

In clustering classes are not predefined

Clustering always try to determine the similarity among the data and accordingly form groups.

i.e. the most similar data is grouped into cluster.

So, we can say Clustering is a special type of classification.

Ex: The advertising is for a special sale on children's clothes. Then they will target only to those who have children.

Attribute 1

another attribute can be advertising only to those who are located near the show room.

Attribute 2

Ex: A company wishes to group its customers and company mgmt does not have any specific label for these groups.
Based on the outcome of grouping they are trying to target marketing and advertising campaigns to the different groups.

Sample Data :

Income	Age	Children	Marital Status	Education
15,000	35	3	Married	Highschool
20,000	25	0	Single	Highschool
25,000	40	1	Divorced	College
30,000	20	0	Single	College
35,000	60	0	Married	Highschool
50,000	30	0	Married	Graduate
60,000	45	5	Married	Graduate
70,000	50	2	Divorced	College

We can cluster the given dataset on different attributes.

2 Dataset

Group of those who have children

Group of homes that are close

Basic Data Mining Tasks / Techniques

3. Association Rules

It is used for finding the relationship among the product.

Super Market transactions are analysed

Market Basket analysis

It assist retail stores to assist in
marketing,
advertising,
floor placement and inventory.

Predicting faults in telecommunication N/W.

Association rules are dependent on discovery of frequent set.
Therefore most of the algos try to determine frequent item set.

The selection of Association Rule is based on 2 values

1. Support (S)

   It is a probability that a transaction contains {X U Y}.

   i.e It is the percentage of transaction in which item X and Y occurred together.

2. Confidence (C)

   It is a conditional probability that transaction having X also contains Y.
   i.e. It is the probability that if the L.H.S. appears in a transaction then also the R.H.S. will.

Basic Data Mining Tasks / Techniques

3. Association Rules

Transaction:

Transaction	Item Bought
T1	A, B, C
T2	A, C
T3	D, C
T4	B, E, F

For A ⟶ C [Milk ⟶ Bread]

A ⟶ B = No. of tuple containing both A and B / total no of tuples

Support: Out of 4, 2 transactions are supporting A ⟶ C

Therefore 2/4*100 = 50%

Confidence:

Wherever A is purchased C is also purchased

Confidence = 100%

Conf(A ⟶ B)= Sup(AUB)/Sup(A)

= No of tuple contains A& B/ No of tuples contain A

Class Assignment:

Transaction	Item Bought
100	F, A, B, D
200	D, A, C, E, B
300	C, A B, E
400	B, A, D

Association Rule

- An implication expression of the form $X \rightarrow Y$, where X and Y are itemsets
- Example:
  {Milk, Diaper} → {Beer}

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Rule Evaluation Metrics

- Support (s)
  - Fraction of transactions that contain both X and Y
- Confidence (c)
  - Measures how often items in Y appear in transactions that contain X

Example:

$$\{\text{Milk}, \text{Diaper}\} \Rightarrow \text{Beer}$$

$$s = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{|\text{T}|} = \frac{2}{5} = 0.4$$

$$c = \frac{\sigma(\text{Milk}, \text{Diaper}, \text{Beer})}{\sigma(\text{Milk}, \text{Diaper})} = \frac{2}{3} = 0.67$$

Cust#	Movies				
C1	*Inside Out*	*Halloween*	*Coco*	*A Star Is Born*	*Captain Marvel*
C2	*Inside Out*	*Halloween*	*A Star Is Born*	*Captain Marvel*	
C3	*Inside Out*	*Coco*	*A Star Is Born*	*Captain Marvel*	
C4	*Venom*	*Inside Out*	*Halloween*	*Captain Marvel*	
C5	*Coco*	*A Star Is Born*	*Captain Marvel*		

Calculate the following (20 points, 5 points each)

Supp(*Inside Out*→*Halloween*) = %;
Conf(*Inside Out*→*Halloween*) = _____%;

Supp(*A Star Is Born*→*Captain Marvel*) = _____%;
Conf(*A Star Is Born*→*Captain Marvel*) = _____%;

Which of these two rules do you think is more interesting? Why? (5 points)

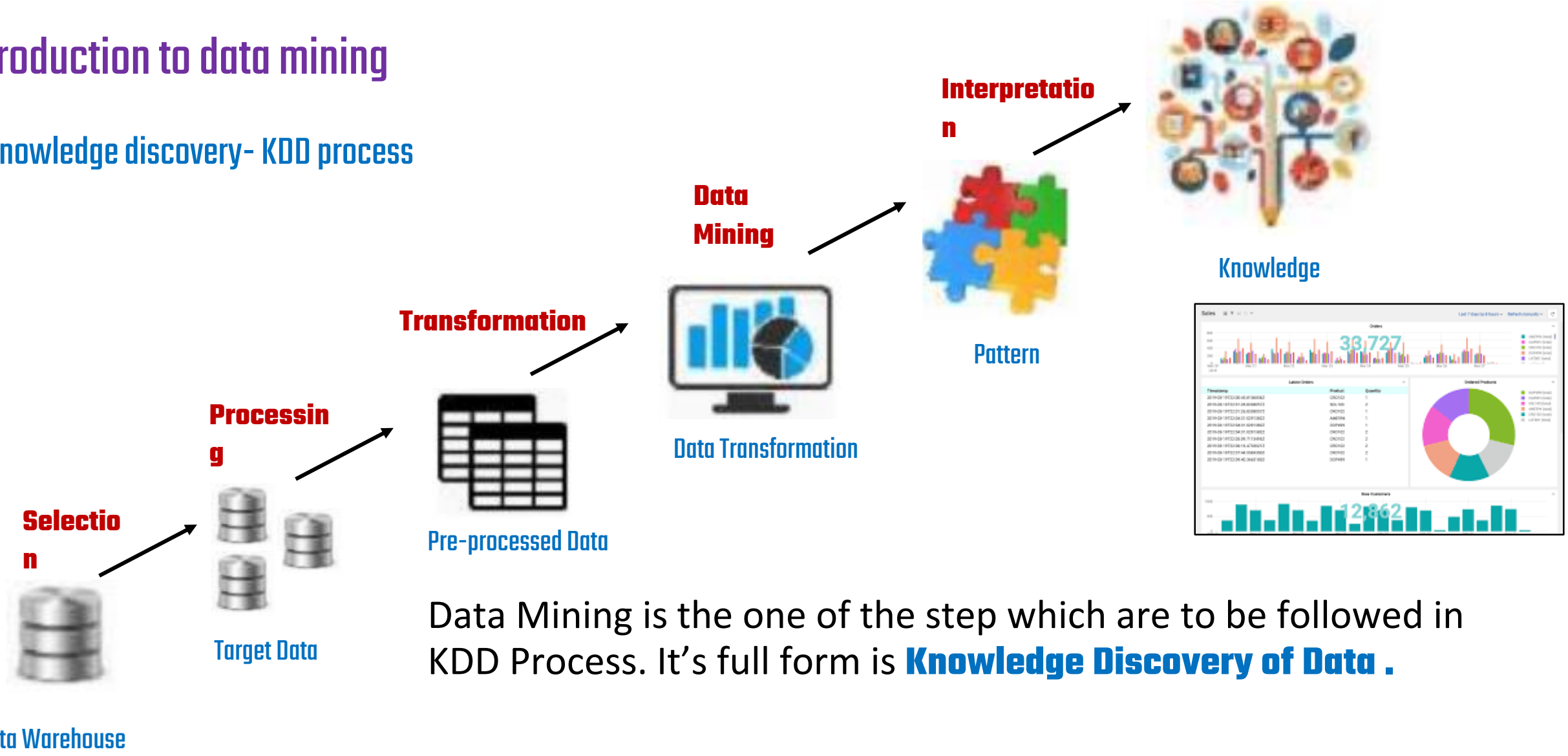
Transaction 1: Frozen pizza, cola, milk
Transaction 2: Milk, potato chips
Transaction 3: Cola, frozen pizza
Transaction 4: Milk, potato chips
Transaction 5: Cola, pretzels

TransactionId	Items
1	{A,C,D}
2	{B,C,D}
3	{A,B,C,D}
4	{B,D}
5	{A,B,C,D}

Trans id	Products purchased
1001	Laptop, Av-software, Speakers, Microphone
1002	Laptop, Speakers, wireless mouse, External Hard disk
1003	LED Television, Speakers, Av-software

Introduction to data mining

Knowledge discovery- KDD process



Data Mining is the one of the step which are to be followed in KDD Process. It's full form is **Knowledge Discovery of Data .**

- To do KDD we need first Data, a huge amount of data to be discovered. Which is available at Data Warehouse.
- **1.Selection:** Here we are selecting the relevant data.
- For Ex: For credit card customer profiling, we extract the type of transaction for each type of customer and we may not be interested in details of the shop where the transaction takes place.
- **2.Processing:** It is the data cleaning stage where unnecessary information is removed.
- For Ex: It is unnecessary to note the sex of a patient when studying pregnancy.
- Since data is collected from various data sources, so there may be corruption of data or missing data. So, these problems are removed at this step.
- **3.Transformation:** Data is gathered from various data sources, so it may have different data types or formats. This data is transformed into common format for further processing at this step.
- **4.Data Mining:** This step applies algorithms to the transformed data to generate the desired results.
- **5.Interpretation:** The usefulness of results is dependent on how the data mining results are presented. For that various visualization and GUI strategies are used.

Types of Data

- ✓Numerical data - continuous variables on a roughly linear scale e.g. marks, age, weight.
  - ❖Units can affect the analysis.
    - ❖Eg:Distance in metres is obviously a larger number than in kilometres. May need to scale or normalise data
- ✓Binary data – many variables are binary e.g. gender, married or not, u/g or p/g.

Types of Data

- ✓Nominal data – similar to binary but can take more than two states e.g. colour, staff position.
- ✓Ordinal data – similar to nominal but the different values are ordered in a meaningful sequence.
- ✓Ratio-scaled data – nonlinear scale data

Preprocessing

1. Basic Data Mining Tasks / Techniques
    1. Classification
    2. Clustering
    3. Association Rules

Data Preprocessing techniques

2. Knowledge discovery- KDD process
    - Selection
    - Processing
    - Transformation
    - Data Mining
    - Interpretation

Data Preprocessing

It's a data mining technique that transforms raw data into a more understandable, useful and efficient format.

Why is data preprocessing required?

Real world data is generally: Since data is coming from large set of different sources of data.

1. **Incomplete:** Certain attributes or values or both are missing, or only aggregate data is available.
   Ex: Occupation =" "
2. **Noisy:** Data contains errors or outliers
   Ex: Salary=-1
3. **Inconsistent:** Data contains differences in codes or names etc.
   Ex: B.Tech.
   B.E. Rating in the format A,B,C and 1, 2, 3 together.

No quality data, no quality mining results.

Quality decisions must be based on quality data.

Data Preprocessing

To improve the quality of the data into the warehouse

1. Data Cleaning
2. Data integration
3. Transformations
4. Data Reduction
5. Data Discretization

It cleans the data by filling in the missing values, smoothing noisy data, resolving the inconsistency and removing outliers.

typing error

naming conventions

Ways to handle missing data during cleaning-

1. Manual Entry of missing data
2. Using attribute mean
3. Using most probable value
   - Decision Tree
   - Regression
4. Using global constant — **NA / Unknown**
5. Ignore the tuple

	Mark
1	20
2	40
3	36.66
4	50

(20+40+50)/3 = 36.66

Data Preprocessing

1 Data Cleaning

Data Smoothing

Binning This method smooth a sorted data value by consulting its "neighbourhood", that is, the values around it.

Step 1: Sort the data

Step 2: Decide the number of bin to be formed

Step 3: Partition into equal-frequency bins

Step 4: Apply binning method

a. Smoothing by bin means

b. Smoothing by bin boundaries

Example of Binning: Price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Step 2: Decide the number of bin to be formed

Max = 34 Min = 4

$$\frac{\text{Max - Min}}{N} = \frac{34-4}{9} = 3.33 \approx 3$$

Step 3: Partition into equal-frequency bins

Bin1: 4, 8, 15

Bin2: 21, 21, 24

Bin3: 25, 28, 34

Step 4: Apply binning method

a. Smoothing by bin means

Bin1: 4, 8, 15 $\frac{4+8+15}{3} = 27/3 = 9$

Bin2: 21, 21, 24 $66/3 = 22$

Bin3: 25, 28, 34 $87/3 = 29$

Bin1: 9, 9, 9

Bin2: 22, 22, 22

Bin3: 29, 29, 29

Data Preprocessing

1 Data Cleaning

Data Smoothing

Binning

Step 4: Apply binning method

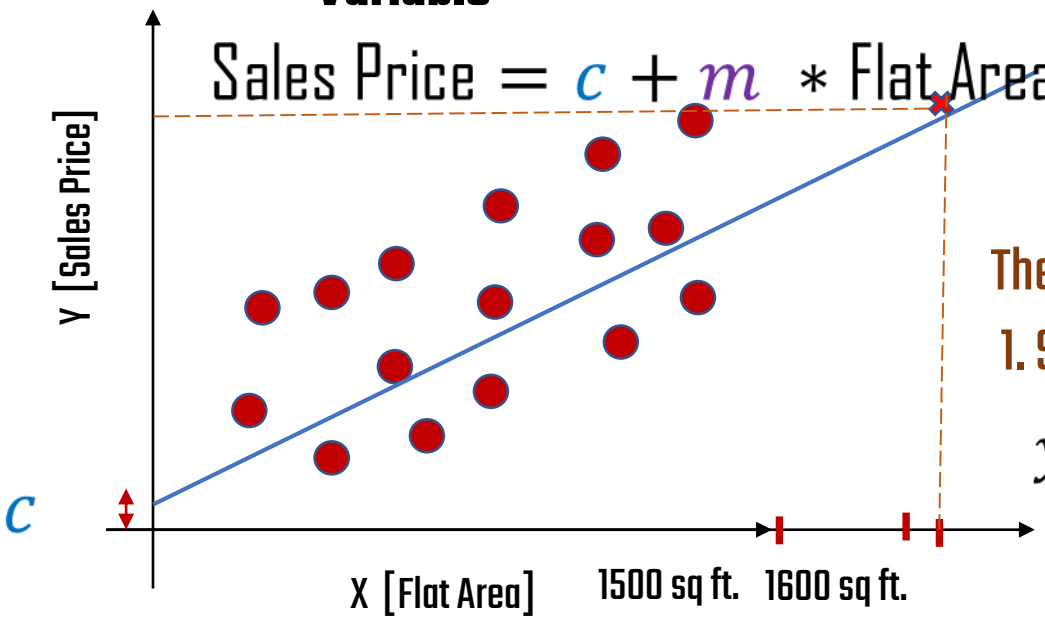
a. Smoothing by bin means

b. Smoothing by bin boundaries

Regression

Dependent Variable

Independent Variable



Example of Binning: Price (in dollars): 4, 8, 15, 21, 21, 24, 25, 28, 34

Bin1: 4, 8, 15

Bin2: 21, 21, 24

Bin3: 25, 28, 34

a. Smoothing by bin boundaries

Bin1 4, 8, 15

:

4 7

Bin2: 21, 21, 24

Bin3: 25, 28, 34

Bin1: 9, 9, 9

Bin2: 22, 22, 22

Bin3: 29, 29, 29

Bin1: 4, 4, 15

Bin2: 21, 21, 24

Bin3: 25, 25, 34

There are two types of Regression

1. Simple Regression

$$y_1 = \beta_0 + \beta_1 x_1$$

$$y = c + m\,x$$

2. Multiple Regression

$$y_1 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

Clustering

There are three approaches to perform smoothing –

1. **Smoothing by bin means :** In smoothing by bin means, each value in a bin is replaced by the mean value of the bin.
2. **Smoothing by bin median :** In this method each bin value is replaced by its bin median value.
3. **Smoothing by bin boundary :** In smoothing by bin boundaries, the minimum and maximum values in a given bin are identified as the bin boundaries. Each bin value is then replaced by the closest boundary value.

Sorted data for price(in Rupees) : 2, 6, 7, 9, 13, 20, 21, 24, 30

Partition using equal frequency approach:
Bin 1 : 2, 6, 7
Bin 2 : 9, 13, 20
Bin 3 : 21, 24, 30

Smoothing by bin mean :
Bin 1 : 5, 5, 5
Bin 2 : 14, 14, 14
Bin 3 : 25, 25, 25

Smoothing by bin median :
Bin 1 : 6, 6, 6
Bin 2 : 13, 13, 13
Bin 3 : 24, 24, 24

Smoothing by bin boundary :
Bin 1 : 2, 7, 7
Bin 2 : 9, 9, 20
Bin 3 : 21, 21, 30

Data Preprocessing

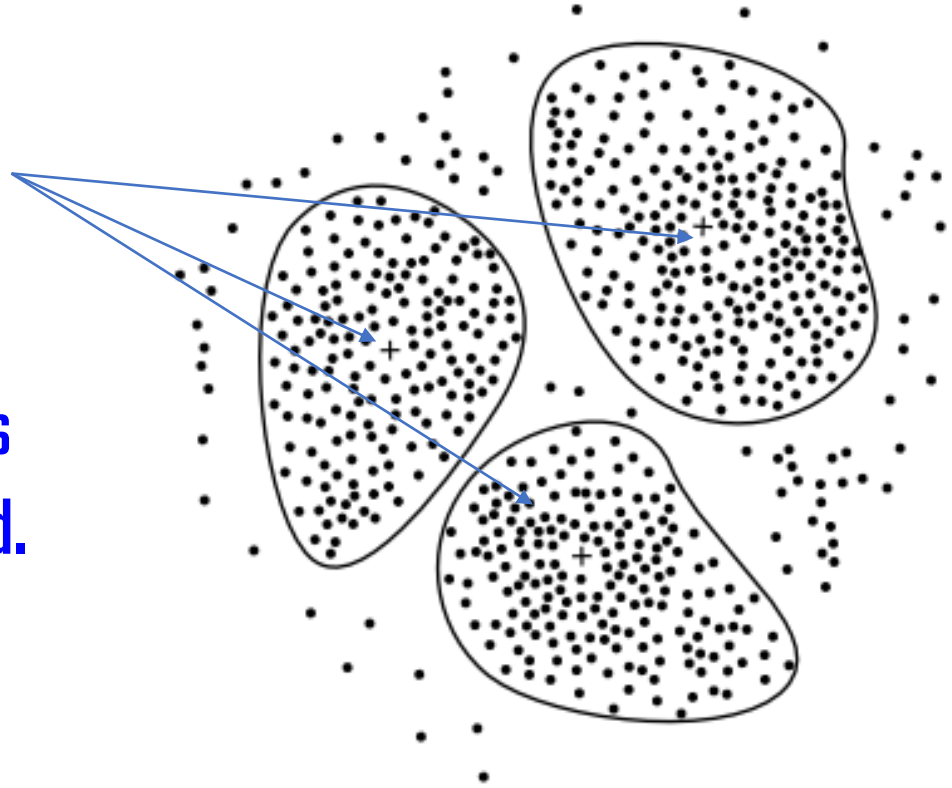
1 Data Cleaning

Clustering It is the forming of groups of similar data.

And the data which is not similar to the data, is called outlier and we remove it.

Cluster centroid is marked with a "+"

Values that fall outside of the sets of clusters are treated as outliers and they are removed.



A 2-D plot of customer data with respect to customer locations in a city, showing three data clusters.

Data Preprocessing

1. Data Cleaning
2. Data integration
3. Transformations
4. Data Reduction
5. Data Discretization

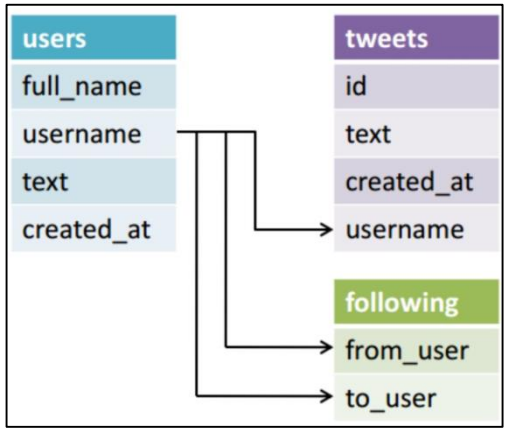
2 Data integration

Data Integration: mergeing the data from different sources to form a data source like Data Warehouse. So, the sources can be Flat files MDDB, Data Cubes etc.

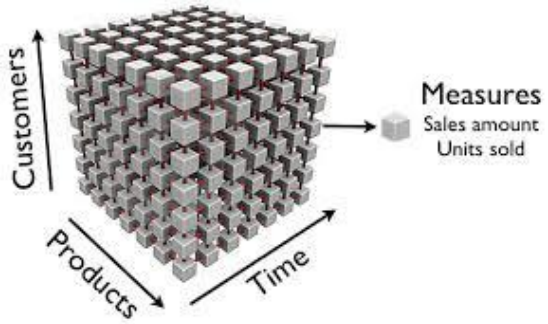
✓ Merging the data from different sources to form a data source

	Rout No.	Miles	Activity
Record 1	1-95	12	Overlay
Record 2	1-495	05	Patching
Record 3	SR-301	33	Crack seal

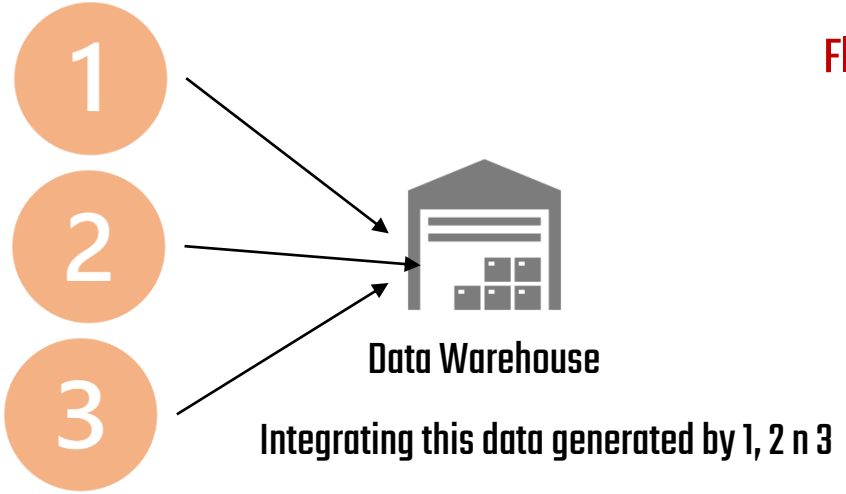
Flat files



MDDB



Data Cubes



Integrating this data generated by 1, 2 n 3

Issues in integrating data

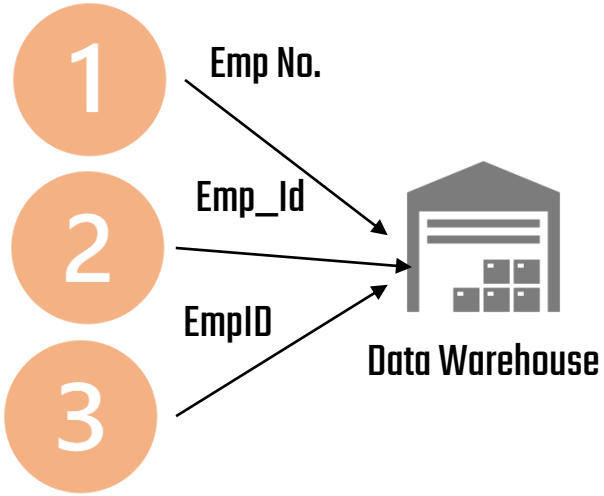
1. Schema integration and object matching
2. Redundancy
3. Detection and resolution of data value conflicts

Discuss issues to consider during data integration. May 2018 [10 Marks]

Data Preprocessing

Issues in integrating data

1. Schema integration and object matching



Emp No.	Name
001	ABC
002	XYZ
003	PQR

1

Emp_Id	Name
350041-1	MNO
350041-2	UVW
350041-3	DEF

2

EmpID	Name
AA-011	GHI
AA-012	STU
AA-013	WXY

3

Correctly modify the values

2. Redundancy

Emp No.	Name	DOB	Age
001	ABC	----	----
002	XYZ	----	----
003	PQR	----	----

Redundant data

3. Detection and resolution of data value conflicts

Two different tables may show same type of data using different values

Price
100 ₹

Price
2.5$

Price
50 ₪

Data Preprocessing

3 Transformations

The data are transformed in ways that are ideal for mining the data.

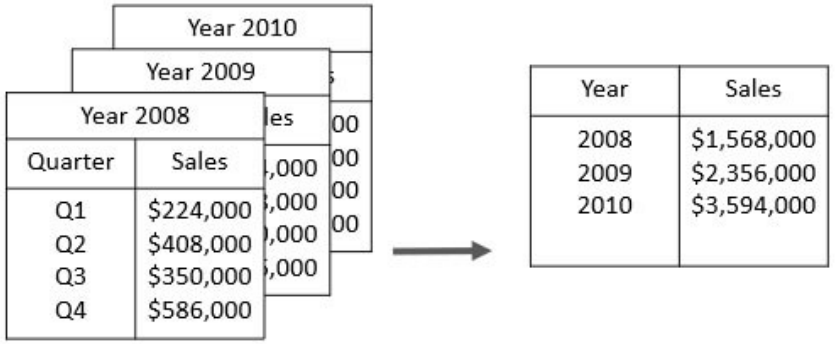
1. Smoothing

Removing the noise from the data.

Methods used are Binning, Regression and Clustering

2. Aggregation

here summary or aggregation operations are applied to the data

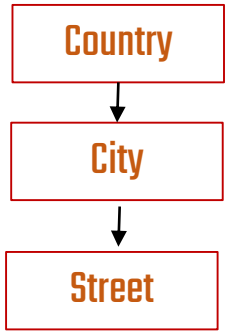


Aggregated Data

Quarterly sales data may be aggregated to compute annual total amounts.

3. Generalization

Here low-level or "primitive" (raw) data are replaced by higher-level concepts through the use of concept hierarchies.



like street, can be generalized to higher-level concepts, like city or country

4. Normalization

- Normalization is used to scale the data of an attribute so that it falls in a smaller range, such as -1.0 to 1.0 or 0.0 to 1.0. It is generally useful for classification algorithms.
- Normalization is generally required when we are dealing with attributes on a different scale.
- Methods of Data Normalization –
  - Decimal Scaling
  - Min-Max Normalization
  - z-Score Normalization(zero-mean Normalization)

To normalize the data by this technique, we divide each value of the data by the maximum absolute value of data

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

- *Let the input data is: -10, 201, 301, -401, 501, 601, 701*
- *To normalize the above data,*
- *Step 1: Maximum absolute value in given data(m): 701*
- *Step 2: Divide the given data by 1000 (i.e j=3)*
- *Result: The normalized data is: -0.01, 0.201, 0.301, -0.401, 0.501, 0.601, 0.701*

the attribute data are normalized by scaling their values, so that they fall within a small specified range

3 Transformations

2. Min-max normalization

linear transformation is performed on the original data. Minimum and maximum value from data is fetched and each value is replaced according to the following formula.

V Original value of attribute A

200
1 min_A
3
150
10
40,
500 max_A

new value $$v' = \frac{v - min_A}{max_A - min_A}$$

some how, we are converting them into one range between 0 to 1

$$v' = \frac{200 - 1}{500 - 1} = \frac{199}{499} = 0.398$$

3. z-score normalization

Zero mean normalization

- The Z-Score value is one of the Normalization Techniques in Data Mining that determines how much a data point deviates from the mean.
- It calculates the standard deviations that are below or above the mean. It might be anywhere between -3 and +3 standard deviations.

mean of Attribute

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

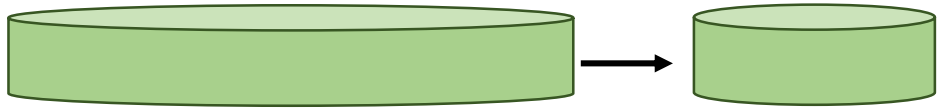
Standard Deviation of Attribute

Data Preprocessing

1. Data Cleaning
2. Data integration
3. Transformations
4. Data Reduction
5. Data Discretization

4 Data Reduction

Data from the AllElectronics data warehouse for analysis

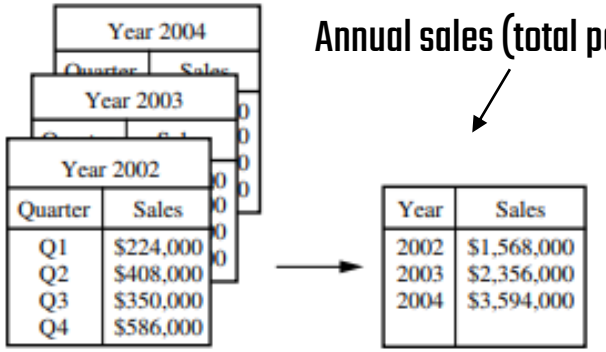


✓ 1. Data cube aggregation

✓ 2. Dimensionality reduction

✓ 3. Numerosity reduction

4. Discretization and concept hierarchy generation

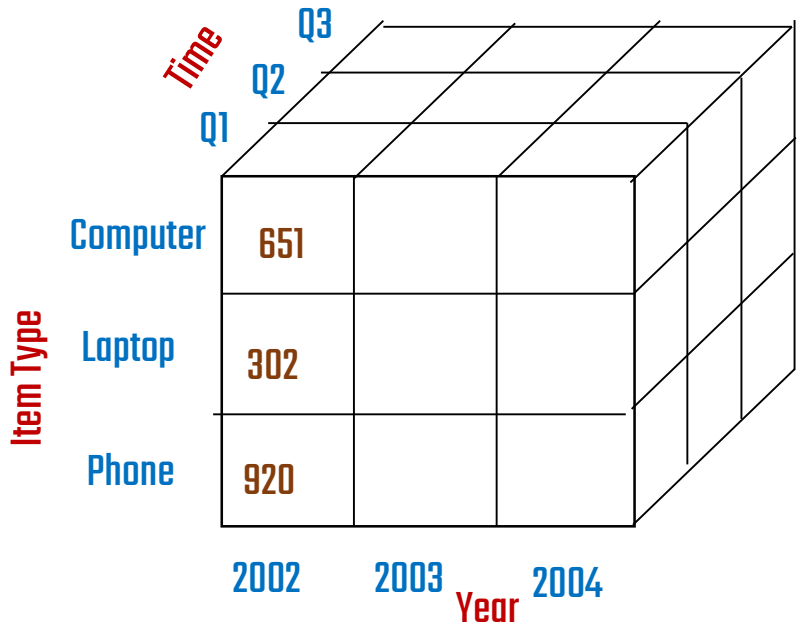


Annual sales (total per year)

Thus the data can be aggregated so that the resulting data summarize the total sales per year instead of per quarter.

All Electronics sales per quarter for the years 2002 to 2004

Now, we will represent it in the form of Data Cube.



Data cube aggregation: here we can find out that in this year how sale is done

2. Principal Components Analysis

4 Data Reduction

2. Dimensionality reduction

Data encoding or transformations are applied so as to obtain a reduced or "compressed" representation of the original data.

lossless

If the original data can be reconstructed from the compressed data without any loss of information, the data reduction is called lossless.

lossy

If we can reconstruct only an approximation of the original data, then the data reduction is called lossy.

1. Wavelet Transforms

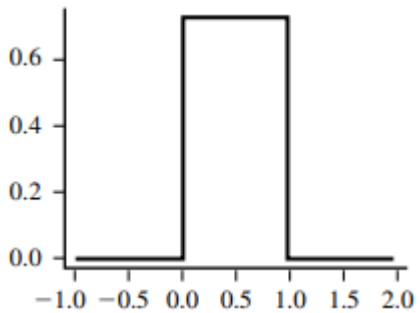
Digital signal processing and image processing

Fourier Transformation

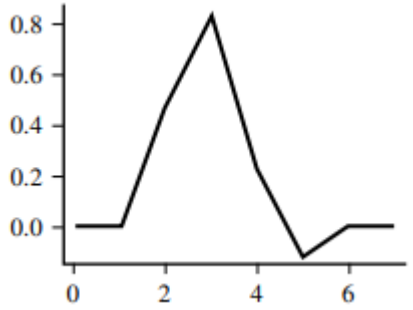
Data vector X ⟶ vector X ' wavelet coefficients

The technique also works to remove noise without smoothing out the main features of the data

Popular wavelet transforms are:



Haar-2



Daubechies-4

Data Preprocessing

4 Data Reduction

2. Dimensionality reduction

2. Principal Components Analysis

[Karhunen-Loeve or K-L method]

Dimensions (**D**) means number of independent variables

Here we convert $N^D \longrightarrow n^D$ Where $n \leq N$

PCA finds the new set of variables smaller than the original set of variables.

Link for PCA

2. Nonparametric methods

a. Histogram

b. Clustering

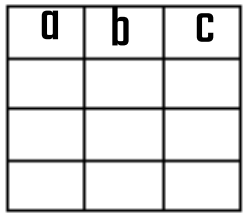
c. Sampling

3. Numerosity reduction

Can we reduce the data volume by choosing alternative, 'smaller' forms of data representation?

1. Parametric method

Only the data parameters need to be stored, instead of the actual data.



Log-linear models

$$y = x_1 + ax_2 + bx_3 + cx_4$$

If there are 10 ^10 data size is there then we can represent this only storing x1,x2, x3 and x4 only

Data Preprocessing

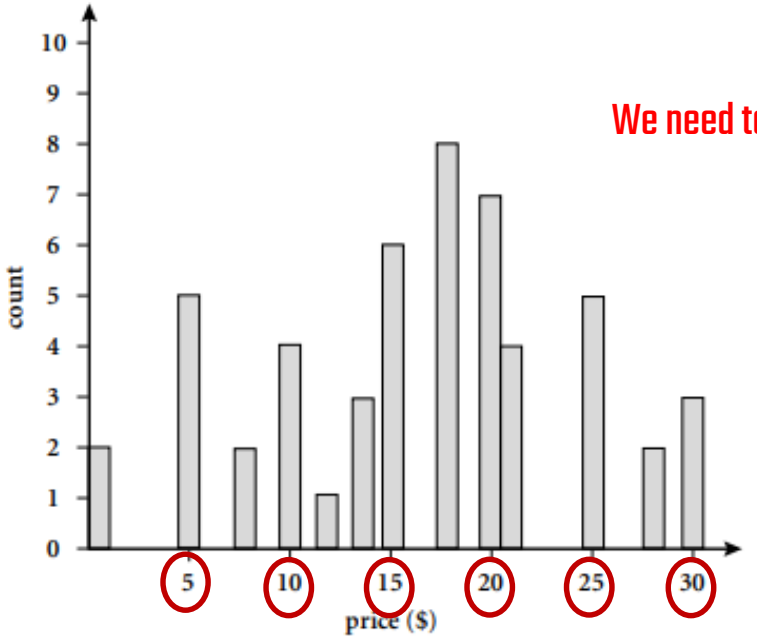
4 Data Reduction

3. Numerosity reduction

2. Nonparametric methods

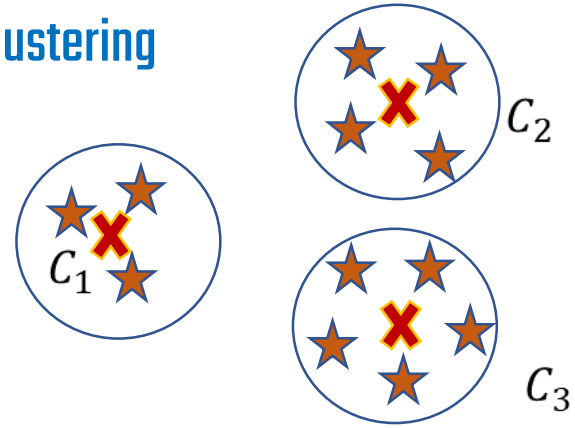
a. Histogram

The following data are a list of prices of commonly sold items at AllElectronics (rounded to the nearest dollar). The numbers have been sorted: 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, 20, 20, 20, 20, 20, 20, 21, 21, 21, 21, 25, 25, 25, 25, 25, 28, 28, 30, 30, 30.

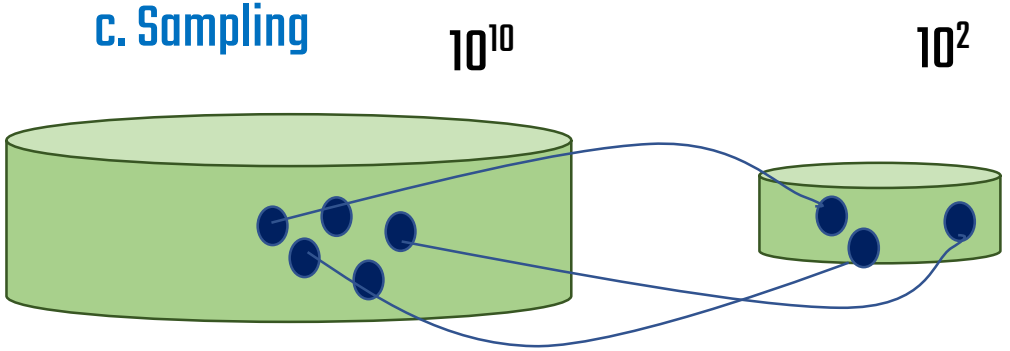


We need to store only mean value of each bucket.

b. Clustering



c. Sampling



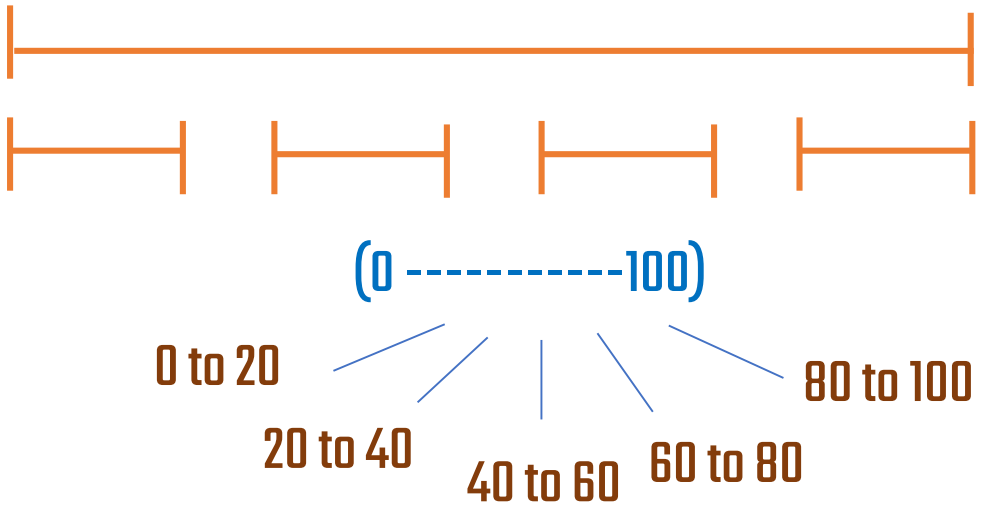
Data Preprocessing

4 Data Reduction

4. Discretization

It divides the range of attribute into intervals so as to reduce number of values for a given continuous attribute.

Continuous data



Concept hierarchy generation

convert low level concept to into higher level concept

Example:

We have an attribute as age with the following value

Age: 10, 11, 13, 14, 17, 19, 30, 31, 38, 40, 42, 70, 72, 73, 75

We will group the data into high level concept