T.Y.B.Tech (CSE)

Data Warehousing And Data Mining

Theory Assignment No – 1

Name: Aniruddha Shende

Roll number: PE04

Batch: E1

Panel: E

Batch: E1 PE04_Aniruddha Shende

Name :- Aniruddha Arun Shende
Roll no:- PE04
Batch :- E1
Panel :- E

Subject :- Data Warehousing & Data Mining

DWDM

THEORY ASSIGNMENT NO-1

Batch-1

Q.1.a: Enlist different types of attributes used in any Data Mining project. Discuss them from A Data Mining Perspective with examples.

Ans 1.a: The different types of attributes used in any Data Mining project are:-

1. Qualitative attributes:-
   - a. Nominal
   - b. Binary
   - c. Ordinal
2. Quantitative:-
   - a. Numeric
   - b. Discrete
   - c. Continuous

Nominal attributes:- The values of a nominal attribute are symbols or names of things. Each value represents some kind of category, code, state & so nominal attributes are also reffered to categorical.

Eg:- The attribute marital status can take on the

example values like single, married. Another example of nominal attribute is occupation, with values such as teacher, dentist, programmer, farmer & so on.

Binary attributes:- A binary attribute is a kind of nominal attribute with only 2 categories or states: 0 or 1, where 0 typically means that the attribute is absent & 1 means that it is present.
Eg:- The result of a medical test has 2 possible outcomes, where 1 means positive & 0 means negative

Ordinal attributes:- An ordinal attribute is an attribute with possible values that have a meaningful order or ranking among them, but magnitude between successive values is not known.
Eg:- Customer satisfaction can have the following ordinal categories:- somewhat dissatisfied, neutral, satisfied, very satisfied.

Numeric Attributes:- A numeric attribute is quantitative, that is, it is a measurable quantity, represented in integer or real values.
(a) Interval - Scaled Attributes:- Interval scale attributes are measured on a scale of equal-size units. Eg:- Temperature attribute is interval-scaled.
(b) Ratio Scaled Attributes:- A ratio-scaled attribute is a numeric attribute with an inherent zero-point.
Eg:- Years of experience in a company, etc.

Discrete Attributes :- A discrete attribute has a finite or countably infinite set of values, which may or may not be represented as integers.
Eg:- Medical test, drink size, etc.

Continuous Attributes :- If an attribute is not discrete, we say it is continuous. It falls in a continuous sequence. They are generally represented by floating point variables.
Eg:- Persons height, etc.

Q.2.a. Analyze data reduction Techniques. Discuss Numerosity Reduction suitable examples.

Ans. 2.a. Following are the different Data Reduction Techniques:

(a) Data Cube Aggregation:- This technique is used to aggregate data in a simpler form.

(b) Dimensionality Reduction:- Whenever we come across any data which is weakly important, then we use attribute required for analysis. It reduces data size as it eliminates outdated / redundant features.

(c) Data Compression:- The Data compression technique reduces the size of file by using different encoding mechanisms like Huffman Encoding & Run Length Encoding.

(d) Numerosity reduction:- Numerosity reduction techniques replace the original data volume by alternative smaller forms of data representation

(e) Discretization & Concept Hierarchy generation:- Techniques of data discretization are used to divide attributes of the continuous nature into data with intervals.

Different methods for Numerosity reduction are:-

① Regression / log-linear model (parametric)
② Histograms, clustering, sampling (non-parametric)

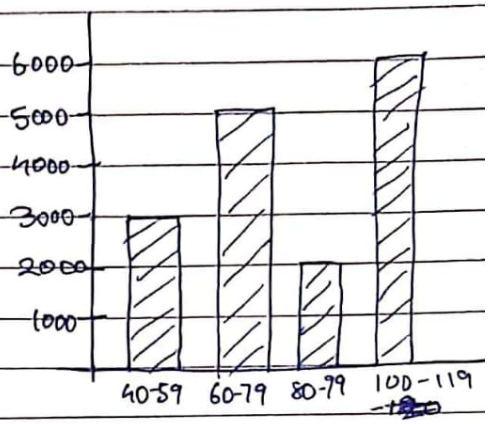
Eg:- For a large dataset, we can plot Histogram, and use binning to approximate data reduction distribution & is a popular form of data reduction.
Clustering can also be used in the same way & it also helps to detect outliers in data

Q.2.b. Illustrate histogram analysis with suitable examples.

Ans.2.b. Histogram is graphical representation for summarizing the distribution of a given attribute, X.
A histogram for an attribute A partitions the data distribution of A into disjoint subsets, or buckets. Typically, the width of each bucket is uniform.
Each bucket is represented by a rectangle whose height is equal to the count or relative frequency of the values at the bucket.

For eg:- The following data are a list of AllElectronics prices for commonly sold items

Unit price ($)	Count of items
40	275
43	300
47	250
⋮	⋮
117	270
~~120~~ 119	350



From the above plot we can get a clear idea about the sales by plotting a histogram. We can also plot a histogram of an image.

The following data are a list of prices of commonly sold items (rounded to nearest dollar):-

[1, 1, 5, 5, 5, ... 28, 28, 28, 30, 30, 30].