Rohan Nyati 500075940 (148) Batch - 5
(AI & ML)

Assignment - 1

Q.1) What are the different types of data that can be mined?

Solⁿ) The different types are:-

1) Flat files
2) Relational Databases
3) Data Warehouses
4) Transactional Databases
5) Multimedia Databases
6) Spatial Databases
7) Time Series Databases
8) World Wide Web (WWW)

Q.2) Describe five number Summary in detail.

Solⁿ) The five number is a set of descriptive statistics that provides information about a dataset. It consists of the five most imp. sample percentiles.

i) The Sample minimum (smallest observation)
ii) The lower quartile or first quartile
iii) The median (the middle value)
iv) The upper quartile or third quartile
v) The sample maximum (largest observation)

Inaddition to the median of asingle set of data there are two related statistics called the upper and lower quartile. If data are placed in order, then the lower quartile is central to the lower half of the data and the upper quartile is central to the upper half of the data. These qualities are used to calculate the interquartile range, which helps to describe the spread of the data and determine whether or not ~~any~~ any data points are outliers.

Q.3) What are the different data visualization methods used in data mining?

Sol^n) The different types of data visualization methods are:-

1.) Charts - The easiest way to show the development of one or several sets in chart. Charts vary from bar and line charts that show relation b/w element over time.

2.) Plots - Plots allow to distribute two or more data set over 2D or even 3D space to show the relation b/w these sets and parameters on the plot

3.) Maps - Maps are popular way to visualize data used in different industries.

4) <u>Diagram and Matrices</u> - Diagrams are usually used to demonstrate complex data relationship and links and include various types of data on one visualization. They canbe hierarchical, multi-dimensional, tree-like.

Q.4] How to Calculate a distance b/w two points using:-

1) Euclidean Distance -

$$d(p,q) = \sqrt{\sum_{p=1}^{n} (q_i - p_i)^2}$$, where

p, q = two points in euclidean space
q_i, p_i = euclidean vectors, starting from the origin of the space.
n = n-space

2.) Manhattan Distance -

The distance b/w two points measured along axis at right angles in a place p_1 at (x_1, y_1) and place p_2 at (x_2, y_2), it is

$$\longrightarrow |x_1 - x_2| + |y_1 - y_2|$$

3) Supremum Distance-

Distance b/w two points as the maximum diff over any of their axis value. In a 2-D grid for instance, if we have two points (x_1, y_1) and (x_2, y_2)

then, $\text{Max}(|y_2 - y_1|, |x_2 - x_1|)$

Q.5 Write a Short note on Data Reduction and Transformation.

Sol^n It is a technique used to obtain a reduced representation of the data set that is much smaller in ~~value~~ volume than the original data but still contains all the critical information.

The following methods are used for data reduction.

i) Data Cube Aggregation
ii) Dimensionality Reduction
iii) Data Compression
iv) Numerosity Reduction
v) Discretization and Concept hierarchy generation.

Data Transformation is a process in which data is transformed from one format to another format which is more suitable for data mining.

Some data transformation strategies are:-

1) Smoothing
2) Aggregation
3) Generalisation
4) Normalisation
5) Attribute Construction