DATA WAREHOUSING AND MINING (KB = Kiran Bhowmick Ma'am)

- Kindof static model. Have to make it adapt to new circumstances.

- Warehousing is related to storing, reloading and retrieving large amounts of data and maintaining it.

- Practicals in python.

- Need to write algorithms in mining instead of just calling them.

- Practicals would be interesting and tiring.

- Extensive syllabus as 2 topics.

- Diff chapters from diff books.

- 7 Units. Need to study all.

- Han, Kamber Data Mining is BIBLE

- Paulraj Pnniah is good for DW

COURSE OUTCOMES:

- Under fundamentals of DW and principles and algos of DM

- Design Sample DW with dimesnional modelling

- Understand diff DM techniques.

- Pracs will have HOD and KB

- Subject will be taught by KB

DATA MINING:

- Extracting data not openly visible

- Find patterns

- Make predictions.

- Use the vast amounts of data for productive purposes

- We are drowning in data, but starving for knowledge!

- Also known as knowledge discovery. Not knowing what we might find.

- Bussiness intelligence based on statistics.

Different ways of doing data Mining:

KDD: Knowledge discovery from data

KDD gives step wise instructions on how to extract knowledge from data you have.

Various steps in KDD:

1) Collect huge amounts of data from db's, ledgers, excel sheets, etc.

2) Store the data in data warehouse. Consists of cleaning and integration of data. (No changing of formats)

3) Select relevent data drom warehouse and perform transformation functions to bring it in one format.

4) Apply data mining models. Requires selection of the appropriate technique.

5) Evalute the pattern and do visualization.

Data Mining includes:

- Algorithms

- Statistics

- Visualization

- Pattern Recognition

- Machine Learning

- Databases, etc.

- Diff types of data to be mined:

- Streaming data

- Transacction data

- Text

- Multi-media ,etc.

Techniques:

- Data warehouse (OLAP)

- ML

- Statistics

- Visualization, etc.

DATA MINING DEFN:

- Find hidden data

- Fit to model

- Exploratory data abalysis

- Data driven discovery

- Deductive Learning

2 TASKS:-

Predictive:

- Classificication

- Regression

- Time series analysis

- Future states prediction

Descriptive:

- Clustering

- Summarizing

- Link Analysis

- Sequence discovery

TECHNIQUES:-

CLASSIFICATION:

- Has predefined groups/classes.

- Pattern recognition

- Grouping and prediction of data

REGRESSION:

- Has a range of prediction values

- No predefined classes.

TIME SERIES ANALYSIS:

- Can be used to predict stocks based on its history

- Prediction is done over a time period

- Determine similar patterns over time

- Classify behaviors.

CLUSTERING:

- Group data based on how similar they are

- Partitioning

- E.g Housing clusterization

SUMMARIZATION:

- Maps data into subsets with associated simple descriptions (summaries).

- Characterization

- Generalizatiom

LINK ANALYSIS:

- Find hidden relations among data

- Affinity analysis

- Association Rules

- Sequential analysis determines sequential patterns.

SEQUENCE DISCOVERY:

- Find patterns based on sequence

- e.g person goes to home page, then contact, then produccts, etc.

- - - - -- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -- - - - - -- - - - - - - - - - - - - - -- - - -- - - - - - - -

Data Mining: Types of Data

- What is a dataset?

-- It is a collection of data after warehousing has been done. Will be in te form of rows and columns, .xls, .csv, etc.

-- Rows represent one single instance/sample/object.

-- Columns represent attributes/fields/dimensions. E.g. Name, age, height, etc.

-- Data objects = instances, samples, data points, etc.

-- Attributes = fields or features.

- Data types and attributes types.

-- Nominal / categorical attributes

-- Binary attributes

-- Ordinal

-- Numerical

NOMINAL:

- values unrelated to numbers. Maybe Names or special symbols.

- Also called categorical values as each value represents a category

- Nominal values may be numberic. e.g cusomerID. You never require to find the mean of customerID

- Cannot find mean or median as these attributes are not quantitative.

- Finding mode is possible

- Generally finite

BINARY:

- Nominal attributes that have only 2 values - 0 and 1

- Also referred to as Boolean is values are True and False

- SYMMETRIC BINARY ATTRIBUTES:

- Both states are equally valuable. Have same weight

- E.g. Gender

- ASYMMETRIC BINARY ATTRIBUTES:

- Both states are not equally valuable. Don't have same weight.

- E.g Medical test results or student attendence.

ORDINAL:

- Attribute values have a meaningful order or ranking

- E.g. Grades - A+, A, A-, B ..., or faculty ranks, etc

NUMERICAL:

- Quantitative

- Values - Integer or Real

- INTERVAL SCALED:

- They are measured on equal size of units

- E.g. Temperature scale, dates in a calendar

- Comparable

- Can find mean, median and mode

- True Zero point cannot be calculated/exits.

- Temps in degree Celcius or Farhenheit

- RATIO SCALED:

- Numeric values with inherent zero-point.

- One value can be a multiple of another.

- True Zero point exists

- Temp in Kelvin

- Discrete and Continuous attributes

- Discrete are finite number of values

- May or may not be integers

- Countably infinite

- Continuous - numeric atributes

- Can be Real numbers

- Measure of central tendency

- Mean, Median, Mode, Midrange

- Median - L1 + (N/2 - (Sigma[freq]i)/freq(median))\*width FIND THIS CORRECT FORMULA

- Unimodal - datasets with one mode

- Bimodal - datasets with 2 modes

- Multimodal - 2 or more modes

- No mode - All values are unique

- MIDRANGE:

- Measures central tendency of numeric data set

- Avg of largest and smallest value in the data set

- GROUPED MEAN:

- Find midpoint of the range.

- Multiply it by freq and find mean.

Graphical user interface, table

Description automatically generated

* GROUPED MODE:

Graphical user interface, text, application

Description automatically generated

* DISTRIBUTION OF DATA
  + You take values and plot them to see how they are ranging
  + If mean, median and mode are at same point. It is a normal distribution and symmetrical curve
* Positively Skewed: Mean > Median and Mode
* Negatively Skewed: Mean < Median and Mode

Diagram

Description automatically generated

* **DISPERSION OF DATA**
  + Range: Max – Min value
  + Quartiles: Points diving data into n equal parts. 2-quantiles [1 point] and 4-quantiles [3 points].
  + In 4-quantiles -> Each part is called a quartile.
    - Q1 = 25th percentile
    - Q2 = 50th percentile = MEDIAN
    - Q3 = 75th percentile
  + Inter quartile range => IQR = Q3 – Q1
* Measuring the Dispersion of Data
  + BOX PLOT
    - Min Value
    - Q1 = 25th
    - Q2 = Median
    - Q3 = 75th
    - Max Value
    - Whiskers: Two lines outside the box that extend to Min and Max values

Diagram

Description automatically generated

* OUTLIERS:
  + Minimum Values >= Q1 - 1.5 \* IQR
  + Maximum Values <= Q3 + 1.5 \* IQR
  + If any values cross this range, they are called outliers
  + LE
  + Chart, box and whisker chart

    Description automatically generated
* **LEARN BOXPLOT ANALYSIS**
* **Graphical user interface, text, application, email

  Description automatically generated**
* Low standard deviation means graph is thin.
* High SD means wider graph
* QUANTILE PLOT
  + First arrange data in ascending order
  + Find *f* value [ f(i) = x(i) – 0.5 / N]
  + Quantiles values are used for observation
* QUANTILE-QUANTILE PLOT
  + Quantile Values of Branch 1 is compared with Branch 2 and observations are made.
  + Allows the user to see where the shift/difference happened.
  + **Chart, scatter chart

    Description automatically generated**
* **HISTOGRAM ANALYSIS**
  + Bar Graphs
  + Some groups/classes/sets are seen
    - FREQUENCY HISTOGRAMS
      * Rectangles/Bars show the frequency
* SCATTER PLOT
  + Plot points like coordinates
  + Negatively Correlation. Inverse Relation
  + Positive Correlation. Direct Relation.
  + Uncorrelated. Totally unrelated.
* LOESS CURVE
  + If you can plot a smooth curve in scatter plot.
  + Used in regression models

**17/09/2021**



Comparing n number of attributes

* Each row is one object
* Each object will have n features
* Find how similar/dissimilar two objects with n attributes are?
* Similar objects will belong to same cluster.
* Nearest Neighbour Classification:
  + We try to find “how” closer is the object.
* Text

  Description automatically generated
* Graphical user interface, application

  Description automatically generated
* Row 2 contains dissimilarity between object 2 and object 1
* Row 3 contains dissimilarity between object 3 and object 2
* Graphical user interface, text, application, email

  Description automatically generated
* FIND CALCULATING DISSIMILARITY MATRIX
* FIND BINARY VARIABLE MATRIX
* Table

  Description automatically generated
* FOR ASYMMETRIC COMAPRISION:
  + Numerator = no of (1,0) + no of (0,1)
  + Denominator = No of (1,1) + Numberator
* Text, letter

  Description automatically generated
* Text

  Description automatically generated
* For Ratio Scaled variables we apply a **logarithmic** transmission
* Diagram

  Description automatically generated with medium confidence
* **Take avg to merge matrixes if there are no missing values**