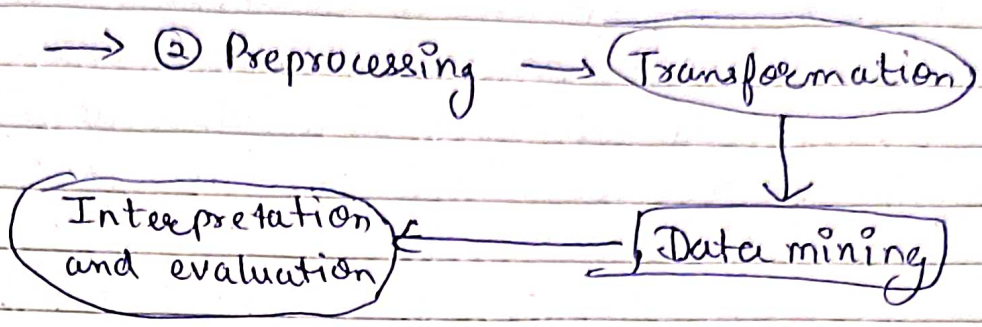
Data Mining

What is data Mining ?

Data mining is concerned with the analysis of data and the use of software techniques for finding patterns & regularities in sets of data.

Stages in data Mining



Types of learning.

Inductive learning is the model building process where the environment i.e database is ~~anay~~ analyzed with a view to finding patterns. Similar objects are grouped in classes and ① ~~Supervised learning~~ rules formulated whereby it is posible to predict the class of unseen objects.

This inductive learning has two main strategies.

① Supervised learning → This is learning from examples where a teacher helps the system construct a model by defining classes & ~~simply~~ supplying examples of each class. The system has to find a description of each class i.e the common properties in the example. Once the description has been formulated the description and the class form a classification rule which can

be used to predict the class of previously unseen objects.

(2) Unsupervised learning :- This is learning from observation and discovery. The data mine system is supplied with objects but no classes are defined so it has to observe the examples & recognition patterns by itself. This system results in a set of class descriptions one for each class discovered in the environment.

(#) A datawarehouse is a repository of information collected from multiple sources stored under a unified schema and which usually resides on a single site

★ Data Mining Applications.

The new database applications include handling spatial data (such as maps), engineering design data (such as design of buildings, system component), hypertext & multimedia data (including text, image, video and audio data), time related data (such as historical records), and the World wide Web
All these applications require efficient data structures & scalable methodes for handling complex object structures, text & multimedia & database schema's with complex structures & dynamic changes.

① Object oriented Databases.

② Object Relational Database.

③ Spatial Databases used in geographic db (maps), VLSI chip design databases, and medical & satellite image databases.
Geographic databases have a number of applications ranging from forestry and ecology planning to vehicle navigation & dispatching systems.

④ Temporal Databases & Time Series Databases.

They both store time real related data. A temporal database usually stores relational data that include time related attributes.
A time series database stores sequences of values that change with time such as data collected regarding the stock exchange. This data can be mined to uncover trends that could help us plan ~~invest me~~ investment strategies.

★ Data Mining fun^{n}

① Classification.
② Association
③ Clustering – classes are unknown
④ Temporal or sequential → purely based on Time
⑤ Spatial
⑥ Time series

① Classification →
Data mine tools have to infer a model from the database, and in the case of supervised learning this requires the user to define one or more classes.
We have predicted attributes & predicting attributes. A combination of values for the predicted attributes define a class.
Once the classes are defined based on predicting attributes, the system should infer rules that govern the classification & hence the system should be able to find the description of each class.

② Associations.
Association fun^n applied on a collection of items & a set of records returns affinities or patterns that exist among the collection of items.
The specific percentage of occurences is called the confidence factor of the rule.
A typical ~~example~~ application, identified by IBM, that can be built using an association function is Market Basket Analysis. This is where a retailer runs an association operator over the points of sale transaction log, which contains among other infor., transaction identifiers and product identifiers. The set of products identifiers listed under the same transaction identifiers ~~& product~~ constitutes a record. The output of ~~transaction~~ association fun^n is, in this case, a list of product affinities.

③ Sequential / Temporal patterns

Sequential pattern fun^n analyse a collection of records over a period of time for example to identify trends. Where the identity of a customer who made a purchase is known an analysis can be made of the collection of related records of the same structure. The records are related by the identity of the customer who did the repeated purchases.

A Sequential pattern operator could be used to discover for example the set of purchases that frequently precedes the purchase of a microwave oven.

Sequential pattern mining fun^n are quite powerful & can be used to detect the set of customers associated with some frequent buying patterns.

④ Clustering

Clustering and segmentation are the processes of creating a ~~par~~ partition so that all the members of each set of the partition are similar according to some metric. A cluster is a set of objects grouped together because of their similarity or proximity. Objects are often decomposed into an exhaustive and /or mutually exclusive set of clusters.

Clustering according to similarity is a very powerful technique, the key to it being to translate some intuitive measures of similarity into a quantitative measures.

There are a number of approaches for forming clusters. One approach is to form rules which dictate membership in the same group based on the level of similarity betⁿ members. Another approach is to build set fun that measure some property of partition as fun of some parameter of the partition.

✱ Data Mining problems/issues.

① Limited Information

A database is often designed for purposes different from data mining and sometimes the properties or attributes that would simplify the learning tasks are not present nor they can be requested from real world.

~~Fr~~ Inconclusive data causes problem because if some attribute essential to knowledge about the application domain are not present in the data it may be impossible to discover significant knowledge about a given domain.

For example → We cannot diagnose malaria from a patient database if that database does not contain the patients red blood cell count.

② Noise and Missing values.

Databases are usually contaminated by errors so it cannot be assumed that the data they contain is entirely correct. Attributes which rely on subjective or measurement judgements can give rise to errors such that some example may even be misclassified. Error in

either the values of attributes or class information are know are noise. Obviously where possible it is desirable to eliminate noise from the classification inf. as this affects the overall accuracy of the generated rules.

Missing data can be treated by discovery systems in a number of ways.

(1) Simply disregard missing values.
(2) Omit the corresponding records
(3) Infer missing values from known values
(4) treat missing data as a special value to be included additionally in the attribute domain.
(5) or average over the missing values using Bayesian techniques.

(3) Uncertainty
Uncertainty refers to the severity of the error & the degree of noise in the data.

(4) Size, updates & irrelevant fields.
Databases tend to be large & dynamic in that their contents are ever changing as info. is added modified or removed. The problem with this from the data mining perspective is how to ensure that the rules are up-to-date & consistent with the most current inf.

Data Preprocessing

There are a number of data preprocess techniques.

① Data cleaning can be applied to remove noise & correct inconsistencies in the Data. Datat integeration merges data from multiple sources into a coherent data store such as data ware house or a data cube.

② Data Transformations, such as normalization may be applied for example normalization may improve the accuracy & effi. of mining algorithms involving distance measurements

③ Data reduction can reduce the data size by aggregating, eliminating redundant features, or clustering for instance, These data processing techniques, when applied prior to mining, can substantially improve the overall quality of the patterns mined and/ or the time required for the actual mining.

* Data Cleaning

Real world data tend to be incomplete, noisy & inconsistent. Data cleaning routines attempt to fill in missing values, smooth out noise while identifying outliers and correct inconsistencies in the data.

(A) Missing Values - → To fill in the missing values for this attribute we use the fol. method.

(1) Ignore the tuple → not very effective unless the tuple contains several attributes with missing values.

(2) Fill in the missing value manualy → In general this approach is time consuming & may not be feasible given a large data set with many missing values.

(3) Use a global constant to fill in the missing values → Replace all missing values by the same constant such as a label like "Unknown".

(4) Use the attribute mean to fill in the missing value.

(5) Use attribute mean for all samples belonging to the same as the given tuple.

(6) Use the most probable value to fill in the missing values. This may be determined with regression, inference - based tools using a Bayesian formalism or decision tree induction.

1.9.2 (B)

Noisy Data → Noise is a random error or variance in a measured variable.

The data smoothing techniques to remove noise from the data are.

① Bining → Bining methods smooth a sorted data value by consulting its neighborhood, that is the values around it. The sorted values are distributed into a number of buckets or bins. As these methods consult the neighbourhood of values, they perform local smoothing.
Smoothing by bin medians in which each bin value is replaced by the bin median. In smoothing by bin boundaries, the mini. & max values in a given bin are identified as the bin boundaries. Each bin is then replaced by closest boundary.

② Clustering → Outliers may be detected by clustering where similar values are organized into groups or clusters and the values that fall outside of the set of the clusters may be considered outliers.

③ Combined computer & human inspection → Outliers may be identified through a combination of computer & human inspection. Patterns whose surprise content is above threshold are output to a list. ~~This is much~~ A human can then sort through

the patterns in the list to identify the actual garbage ones.

(4) Regression → Data can be smoothed by fitting the data to a fun^n, such as with regression. Linear regression involves finding the best line to fit two variables, so that one variable can be used to predict the other. Multiple linear regression is an extension of linear regression, where more than two variables are involved & the data are fit to a multidimensional surface. Using regression to find a mathematical equation to fit the data helps smooth out the noise.

Many methods for data smoothing are also methods for data reduction involving discretization.

★ Data Integration and Transformation

Data integration combines data from multiple sources into a coherent data store, as in data warehousing. These sources may include multiple databases, data cubes or flat files. There are multiple issues to consider during data integration → Schema integration can be tricky. How can equivalent real world ~~entries~~ entities from multiple data sources be matched up. This is called as entity identification problem.

Database & data warehouses typically have metaData, that is data about the data. Such metaData can be used to help avoid errors in Schema Integration.

Redundancy is another important issue. An attribute may be redundant if it can be 'derived' from another table. Inconsistencies in attribute or dimension naming can also cause redundancies. Some redundancie can be detected by correlation analysis.

The correlation bet^n attributes A & B can be

$$r_{A,B} = \frac{\sum (A-\bar{A})(B-\bar{B})}{(n-1)\,\sigma_A\,\sigma_B}$$

n is the no. of tuples.
$\bar{A}$ & $\bar{B}$ are the respective mean values of A & B.
σ_A and σ_B are Standard deviation of A & B

If $r_{AB} > 0$, A & B are positively correlated, meaning that the values of A increase as the value of B increase.

$r_{AB} = 0$, means A & B are independent, no correlation.

$r_{AB} < 0$, then A & B are negatively correlated. where the values of one attribute increases the values of the other attribute decreases.

The third issue in data integration is the detection & resolution of data value conflicts.

The mean of A is $\bar{A} = \frac{\sum A}{n}$

Standard deviation of A is $\sigma_A = \sqrt{\frac{\sum (A-\bar{A})^2}{n-1}}$

* Data Transformation

In data Transformation, the data are transformed or consolidated into forms appropriate for mining.

It involves foll.

① Smoothing, which works to remove the noise from data such techniques include binning, clustering and regression.

② Aggregation, where summary or aggregation operations are applied to the data. This step is typically used in constructing a data cube for analysis of
③ the data at multiple granularities.

③ Generalization of the data, where low-level or "primitive" data are replaced by higher level concepts through the use of concept hierachies.

④ Normalization, where the attribute data are scaled so as to fall within a small specified range, such as -1.0 to 1.0 or 0.0 to 1.0.

⑤ Attribute construction → where new attributes are constructed & added from the given set of attributes to help the mining process

Smoothing is a form of data cleaning, Aggregation & generalization also serve as forms of data reduction.

An attribute is normalized by scaling its value so that they fall withing a small specified range such as 0.0 to 1.0. It is particularly useful for classification algorithms involving neural network or distance measurements such as nearest neighbor classification & clustering.

3 Normalization → Min Max Normalization, Z-score Normalization & normalization by decimal scaling.

★ Min-Max Normalization or Scaling

This is not to be given in our class

Data (v)
200
300
400
600
1000

min → 200, Max → 1000

$$V = \frac{x - min}{max - min}$$

Data	New data (Normalised Data)
200	0
300	0.125
400	0.25
600	0.5
1000	1

★ Z score Normalization

$$Z = \frac{x - \mu}{\sigma}$$

μ ⇒ mean
σ = Standard deviation

$$\sigma = \sqrt{\frac{\Sigma (x_i - \mu)^2}{n}}$$

* Z-Score Normalization - Mean Absolute Devi

$$Z = \frac{x - \mu}{\Lambda}$$

μ = Mean
Λ = Mean Absolute Deviati

Mean Absolute Deviati $= \Lambda = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| \cdots |x_n - \bar{x}|}{n}$

*

Normalization using Decimal Scaling.

Find value of j,

$$\text{Max}\left(\frac{v_i}{10^j}\right) \leq 1$$

The smallest Integer j such that $\text{Max}\left(\frac{v_i}{10^j}\right) \leq 1$

Data (v)
200
300
400
600
1000

$\frac{200}{10^3} \leq 1$ ← Hence for this $j = 3$

and same goes till last value.

Hence $j = 3$

$\frac{200}{10^3}$	0.2
300	0.3
400	0.4
600	0.6
1000	1