ch-2 Data

→ Data mining is the process of extracting extra knowledge 2 hidden patterns from large amout of data

→ An attribute is a property or characteristic of an object that may vary from one object to another or one time to another For eg. eye colour

→ A measurement scale is a rule (function) that associates a numerical or symbolic value with attribute of an object

- Different types of attributes :

① Nominal : That provide info to distinguish one object from the other (=, ≠) Eg. zip code, employee ID, eye color, gender

② ordinal : They provide info to help order the data (<, >) Eg. {good, better, best}, grades, street no.

③ Interval : The difference b/w values are meaningful i.e a unit of measurement exists (+, -). Eg. calender dates, temp.

④ Ratio : Both difference 2 ratios are meaningful (*, /) Eg. temp. in Kelvin, counts, age, length

★ Nominal 2 ordinal are called categorical or qualitative Interval & ratio are called numerical or quantitative

→ Discete attribute has a finite or countably infinit set of values. Eg. zip codes, ID. Binary attributes has only 2 values. Eg. male, female or true/false

→ Contiuous values are real no's Eg. temp, height etc.

→ For asymmetric att, only the presence of non-zero value is imp.

- General characteristics of Data set :

① Dimensionality

___/___/___

(2) Sparsity

(3) Resolution : For eg. variation in atmospheric pressure on a scale of hours reflect the movement of storms. On a scale of months it is not detectable

- Types of Dataset

Tid	Refund	Marital status	Taxable Inc.	DB
1	Yes	single	125K	No
2	NO	Maried	100 k	No
3	NO	single	90 k	Yes

- Record Data

TID	ITEMS
1	Bread, Soda, Milk
2	Beer, Bread
3	Soda, Diaper, Milk

- Transaction Data

Xload	Yload	Dist.	Load	Thin
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
14.27	7.23	17.34	23	1.2

- Data Matrix

	team	coach	play	win	lost
Doc 1	3	0	5	0	2
Doc 2	0	7	0	0	3
Doc 3	0	1	0	2	0

- Document term matrix

→ It is a type of sparse matrix which tells the word count of a doc

- Types of ordered Data

(1) Sequential / Temporal Data is an extension of record data, where each record has a time associated with it.

(2) Sequence Data consists of data set i.e a sequence of individual entities (words / letter) Eg. genetic code

__/__/____

③ Time Series Data is a series of measurement taken over time.
Eg. avg. monthly temp. of Delhi from 1982 to 2020

- temporal autorelation is when two measurements are close in time, their values are similar

④ ~~Data~~ spatial Data is collection of data over various geographical locations eg. temp. all over india

- spatial autocorrelation: close ~~at~~ location, similar value

→ Precision is the closeness of repeated measurements ~~from~~ of the same quantity to one another. measured by std. deviation

→ A ~~systematic~~ systematic variation of measurements from the quantity being measured. measured by finding the difference b/w mean of measurements & actual quantity

→ Accuracy is the closeness of measurements to the true value of quantity being measured.

→ Outliers are either data objects that have different characteristics from the rest of data set OR values of attribute that are unusual w.r.t to typical values of that attribute.

→ ways to treat missing values:

① Eliminate data object / Attribute

② Estimate missing values

③ Ignore the missing values during analysis

→ Inconsistent values: Eg. given zip code & city does not match

→ Deduplication: process of dealing with ~~&~~ duplicates

DATA PREPROCESSING

① Aggregation: combining of two or more objects into single objects

→ Resulting data sets are smaller, which take less time & memory

__/__/__

→ It can act as change of scope or scale by providing high level view

→ The behaviour of group of objects is much more stable than indivisual objects

★ Disadvantage: Loss of interesting detail

(2) Sampling: selecting a subset of data to be analyzed.

→ It is too expensive or time consuming to ~~poost~~ process all the data

→ A sample is representative if it has appx. the same properties as original data set.

- Simple random sampling: There is an equal probability of selecting any particular item. It can be with or without replacement
- If the data has different types of objects, then it is added to prespecified groups & equal no. of records from each group are selected stratified sampling
- It is difficult to determine sample size, so adaptive/progressive sampling is used in which the size of sample is increased until sufficient: stop when accuracy levels off

(3) Dimensionality Reduction: lowering no. of attributes

→ Many DM algo work better with less dimensionality

→ can eliminate irrelevant features & reduce noise

→ creates more undestandable model, easily visualized

→ Less time & memory reqd.

★ Curse of Dimensionality: Phenomenon that ~~DM algo~~ many types of Data analysis become significantly harder with increase in dimensionality.

(4) Feature Subset Selection: subset of features/att. are used

→ Redundant & Irrelevant features can reduce accuracy & quality

- Embedded Approaches! Feature selection occurs natually as a part of DM algo. The algo itself decides which att to use ~~&~~ or ignore
- Filter Approaches! Features are selected before DM algo is run
- wrapper Approaches: The DM algo is used as black box to find best subsets

(5) Feature Creation: To create a new set of att from original dataset that captures imp. info. This is also known as Feature Extraction

(6) Discretization & Binarization! To transform a continuous att. into a categorical att is called discretization. Transformation of conti. or discrete att into one or more binary att is binarization

- Binarization technique! If there are m categorical values, then uniquely assign each value to an integer $[0, m-1]$. If the att is ordinal, order must be preserved. Convert into ~~to~~ binary, $n = \lceil \log_2(m) \rceil$

- Problem of discretization is to determine how many split points to choose & where to place them.

→ Equal width divide into user specified no. of interval with same width

→ Equal freq. tries to put same no. of objects

(7) Variable Transformation! Trans. applied to values of variable For eg. changing the magnitude of variable

→ Simple Functions: x^k, $\log x$, e^x, $\sqrt{x}$, $1/x$, $\sin x$ or $|x|$

→ Normalisation / Standardisation: Att values are normalized by scaling their values so that they fall in specified range

Eg: If our data is scattered, we can use z trans. so that our values lie b/w 0 & 1

$$\boxed{\text{zvalue} = \frac{x-\bar{x}}{\sigma}}$$

$\bar{x}$ = mean

σ = s.d

If we have outlier, use median instead of ~~mode~~ mean in calc. of s.d

- Dissimilarities b/w data objects

→ Euclidean distance : $d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2}$

→ Minkowski distance : $d(x,y) = \left(\sum_{k=1}^{n} |x_k - y_k| \right)^{1/r}$

$r=1$	Hamming distance
$r=2$	Euclidean "
$r=\infty$	Supremum "

- Properties of Euclidean Distance –

① Positivity

(a) $d(x,y) \geq 0$ for all x & y

(b) $d(x,y) = 0$ if only $x=y$

② Symmetry

$d(x,y) = d(y,x)$ for all x & y

③ Triangle Inequality

$d(x,z) \leq d(x,y) + d(y,z)$ ∀ x y z

→ Measures that satisfy all three properties are known as metric

- Similarities b/w data objects

1. $s(x,y) = 1$ only if $x = 1$ $(0 \leq s \leq 1)$
2. $s(x,y) = s(y,x)$ ∀ xy