# Feature subset selection:-

→ Another way to reduce the dimensionality is to use only a selbset ~~selection~~ of the features.

→ while it might seem that such an approach would lose information, This is not the case if redundent & irrelavant features are present

→ Redundent features duplicate much or all of the information contained in one or more other ~~other~~ attributes. For example the Purchase Price of a Product and the amount of sales tax paid contain much of the same information

→ Irrelavant features contain almost no useful information for the data mining task at hand. For instance, student's ID numbers are irrelavant to the task of predicting student's gradepoint averages.

→ Redundant and irrelavant Features can reduce classification accuracy and the quality of the cluster that are found.

→ while some irrelavant and redundent attributes can be eliminated immediatly by using common sense or domain knowledge, selecting the best subset of features frequently requires a systamatic approch.

→ The ideal approch to feature selection is to try all possible subsets of features as input to the datamining algorithm of interest, and then take the subset that produces the best results. this method has the advantage of reflecting the objective and bias of the dataming algorithm that will eventually be used. unfortunatly since the number of subsets involving n attributes is $2^n$ such an approach is impractical in most situations.

There are three standard approaches to feature selection : embedded, filter & wrapped

1. Embedded approches:-

→ Feature selection occur naturally as part of the data mining algorithms, ~~the algori~~ specifically, during the operation of the data mining algorithm, the algorithm it self decides which attribute to use and which to ignore.

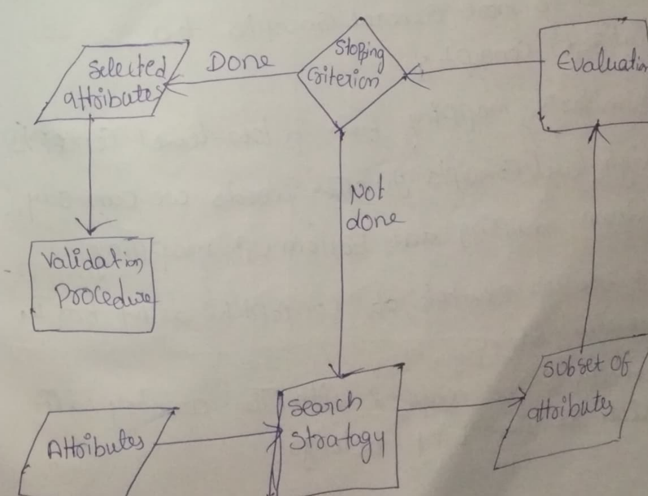
2. Filter approches:-

Features are selected before the data ming algorithm is run, using some approach that is independent of the data mining task

3. wrapped approches:-

Those methods use the target datamining algorithm as a black box to find the best subset of attributes. in a way similar to that of the ideal algorithm described above but typically with out enumarating all possible subsets

An Architecture for Feature subset selection:-



## Discritization & Binaryzation :-

Discritization is used to transform the attributes that are in continuous format

Data Discritization converts a large number of data values into smaller one, so that data evaluation and data management becomes very easy

eg:- we have an attribute of age with the following values

Age: 10, 11, 13, 14, 17, 19, 30, 31, 32, 38, 40, 42, 70, 72, 73, 75

| Attribute | age1 | age2 | age3 |
|---|---|---|---|
| | 10, 11, 13, 14, 17, 19 | 30, 31, 32, 38, 40, 42 | 70, 72, 73, 75 |
| After Discritization | Young | Mature | old |

## Hierarchy Generation data discritization and Concept:

A Concept Hierarchy represents a sequence of mapping with a set of more General Concepts to specialized concepts:

similarly mapping from a low-level concepts to high-level Concepts. In other words we can say top down mapping and bottom up mapping.

Let's see an example of concept hierarchy for the dimension location.

Each city can be mapped with the country with which the given city belongs

For example Delhi can be mapped to INDIA and INDIA can be mapped to Asia

## Top-Down Mapping :-

Top down mapping starts from top with General Concepts and mave to the bottom to the specialized Concepts

↓ ↓ ↓

## Bottom up mapping:-

Bottom up mapping starts from Bottom with specialized concept and move to the top to the Generalized Concepts

↑

## Concept Hierarchy Generation :-

| | | |
|---|---|---|
| General | All | |
| Region | Asia | Europe |
| Country | India | Canada |
| City | Delhi | toronto |
| College | TKRCET | University of toronto |

## Binaryzation:-

Binaryzation is used to transform both The discreate attributes and The Continuous attributes into binary attributes in data mining with respect to Feature selection.

Best Binarization approach is The one That "produces The best result for The data mining algorithm that will be used to analyze The data.

Simple techniques for binarization!

- → Assigning numarical value
- → Finding number of binary attribute require
- → Conversion into binary

eg:- say There is an categorical attribute with 'm' number of values.

→ Assigning numerical value:-

Number assigned will be between [0, m-1]

For ordinal attribute → assignment follows order

→ Finding number of binary attribute required.

say n be The no. of binary attributes

$n = [\log_2 m]$

Converting The number assigned into binary value.

eg: if number of binary attribute is $n = 3$ in numbers Then we can write Three bit binary number

2 - 010

if The number of binary attribute is $n = 4$ in numbers Then

2 - 0010

eg:- let us consider an example to learn

{awful, poor, ok, good, great} - ordinal form

| Attribute values | Integer value |
|---|---|
| awful | 0 |
| poor | 1 |
| ok | 2 |
| good | 3 |
| great | 4 |

Identifying number of binary attributes;

$n = [\log_2 m]$

$n = [\log_2 5] = 3$

Binary conversion

| Attribute Value | Integer values | $x_1$ | $x_2$ | $x_3$ |
|---|---|---|---|---|
| awful | 0 | 0 | 0 | 0 |
| poor | 1 | 0 | 0 | 1 |
| ok | 2 | 0 | 1 | 0 |
| good | 3 | 0 | 1 | 1 |
| great | 4 | 1 | 0 | 0 |

$\underbrace{\qquad\qquad}_{\text{Binarization}}$

If we have mutual relation overcoming the issues

No. of binary attribute = No. of values

| attribute values | Integer value | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|---|---|
| awful | 0 | 1 | 0 | 0 | 0 | 0 |
| Poor | 1 | 0 | 1 | 0 | 0 | 0 |
| ok | 2 | 0 | 0 | 1 | 0 | 0 |
| good | 3 | 0 | 0 | 0 | 1 | 0 |
| great | 4 | 0 | 0 | 0 | 0 | 1 |

Data transformation:-

→ Decimal scaling :-

→ It normalizes the values of an attribute by changing the position of their decimal points.

→ The no. of points by which the decimal point is moved can be determined by the absolute maximum value of attribute A.

→ A value v of attribute A is normalized to v' by computing

$$v' = \frac{v}{10^j}$$

where j is the smallest integer such that $\max(|v'|) < 1$

eg:- suppose values of an attribute p varies from −99 to 99

The maximum absolute value p = 99

For normalization the values we divide the numbers by 100 (i.e j = 2) (or) (no. of integers in largest number so that values come out to be as 0.98, 0.97 and so on.

## Measures of similarity & Dissimilarity

### Basics :-

**similarity:-**

Numarical measures of how alike two data objects are is higher when objects are more a like often falls in The Range $[0, 1]$

**Dissimilarity:-**

Numarical measures of how different are two data objects Lower when objects are more alike

Maximum dissimilarity is often 0

Upper limit varies.

### Datamatrix vs Dissimilarity matrix :-

suppose that we have n objects (eg: persons, items, Courses) described by p attributes (eg: age, height, weight or gender)

The objects are $x_1 = x_{11}, x_{12}, x_{13} \ldots x_{1p}$

$x_2 = x_{21}, x_{22}, x_{23} \ldots x_{2p}$ and soon

When $x_{ij}$ is The value for object $x_i$ of The $j^{th}$ attributes.

Then Datamatrix
$$\begin{bmatrix} x_{11} & \cdots & x_{1f} & \cdots & x_{1p} \\ \cdots & & \cdots & & \cdots \\ x_{i1} & \cdots & x_{if} & \cdots & x_{ip} \end{bmatrix}$$

### similarity / Dissimilarity for two objects with single attribute

p and q are The attribute values for two data objects

| Attribute type | Dissimilarity | similarity |
|---|---|---|
| Nominal | $d = \begin{cases} 0 & \text{if } p = q \\ 1 & \text{if } p \neq q \end{cases}$ | $s = \begin{cases} 1 & \text{if } p = q \\ 0 & \text{if } p \neq q \end{cases}$ |
| ordinal | $d = \frac{\lvert p-q \rvert}{n-1}$ (values mapped to integers 0 to n-1 where n is The no. of values) | $s = 1 - \frac{\lvert p-q \rvert}{n-1}$ |
| Interval (or) Ratio | $d = \lvert p-q \rvert$ | $s = -d$, $s = \frac{1}{1+d}$ (or) $s = 1 - \frac{d - \min\_d}{\max\_d - \min\_d}$ |

### Dissimilarities between Data objects with multiple Numeric attributes.

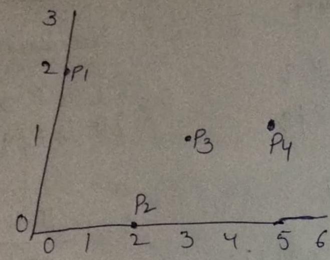
**Euclidean Distance.**

$$dist = \sqrt{\sum_{k=1}^{n} (p_k - q_k)^2}$$

where n is The no. of dimensions & $p_k$ and $q_k$ are respectively, $k^{th}$ attributes of data objects p and q.

→ standardization is necessary, if scales differs.

## Euclidean Distance:




| Point | x | y |
|---|---|---|
| P1 | 0 | 2 |
| P2 | 2 | 0 |
| P3 | 3 | 1 |
| P4 | 5 | 1 |

| | P1 | P2 | P3 | P4 |
|---|---|---|---|---|
| P1 | 0 | 2.828 | 3.162 | 5.099 |
| P2 | 2.828 | 0 | 1.414 | 3.162 |
| P3 | 3.162 | 1.414 | 0 | 2 |
| P4 | 5.099 | 3.162 | 2 | 0 |

Distance matrix

## Minkowski distance:-

Minkowski distance is a Generalization of Euclidiaan Distance, Given two objects p and q

$$dist = \left( \sum_{k=1}^{n} |p_k - q_k|^r \right)^{1/r}$$

where r is a Parameter, n is the number of dimensions and $p_k$ and $q_k$ are respectively, the $k^{th}$ attributes of data objects. p and q.

eg: r = 1 city block (manhattan, taxicab, $L_1$ norm) distance

A Common example of this the Hamming distance which is just the number of bits that are different between two binary vectors.

r = 2 Euclidean distance

r = ∞, "supremum" ($L_{max}$ norm, $L_\infty$ norm) distance

This is the maximum difference between any attribute of the vectors

$$d(x,y) = \lim_{r \to \infty} \left( \sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

Donot Confuse r with n i.e all these distances are defined for all number of dimensions

Minkowski Distance:

| Point | $x$ | $y$ |
|---|---|---|
| $P_1$ | 0 | 2 |
| $P_2$ | 2 | 0 |
| $P_3$ | 3 | 1 |
| $P_4$ | 5 | 1 |

| $L_1$ | $P_1$ | $P_2$ | $P_3$ | $P_4$ |
|---|---|---|---|---|
| $P_1$ | 0 | 4 | 4 | 6 |
| $P_2$ | 4 | 0 | 2 | 4 |
| $P_3$ | 4 | 2 | 0 | 2 |
| $P_4$ | 6 | 4 | 2 | 0 |

| $L_2$ | $P_1$ | $P_2$ | $P_3$ | $P_4$ |
|---|---|---|---|---|
| $P_1$ | 0 | 2.828 | 3.162 | 5.099 |
| $P_2$ | 2.828 | 0 | 1.414 | 3.162 |
| $P_3$ | 3.162 | 1.414 | 0 | 2 |
| $P_4$ | 5.099 | 3.162 | 2 | 0 |

Manhattan Distance formula:

$$D_m(x,y) = \sum_{i=1}^{n} |x_i - y_i|$$

$P_1$ at $(x_1, y_1)$ and $P_2$ at $(x_2, y_2)$

it is $|x_1 - x_2| + |y_1 - y_2|$

$= |0 - 2| + |2 - 0|$

$= 2 + 2$

$= 4$