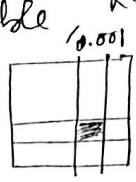
At (DOF, $\alpha(p)$)

i.e at (1, 0.001) from the χ^2 distribution table

its value is 10.828



Since χ^2 value which is computed as 508 is above this value, the hypotheses can be rejected

i.e Gender & Preferred Reading } are not independent

$\Rightarrow$ These two attributes are strongly co-related

If $\chi^2 >$ value at (DOF, $\alpha(p)$), strongly co-related

For Numeric, it is correlation co-efficient

i.e $r_{A,B} = \dfrac{\sum_{i=1}^{N}(a_i b_i) - N\bar{A}\bar{B}}{N\sigma_A\sigma_B}$

N → no. of tuples
A, B → Attributes (Numeric)
a_i → values of A
b_i → " of B

$$-1 \leq r_{A,B} \leq 1$$

If $r_{A,B} > 0$, A & B are positively correlated (i.e if A ↑, B also increases)

If $r_{A,B} = 0$, A & B are independent (NO correlation between them)

If $r_{A,B} < 0$, A & B are negatively correlated

i.e if A ↑, B ↓

Data Transformation

⇒ The Data are transformed from original state to an appropriate new set of consistent state so that the old value can be identified by the one of the new values.

Methods for Transformation

- Smoothing - Remove noise from data using techniques as Binning, Regression, clustering
- Aggregation - Summerization / Aggregation opn. applied
- Generalization - Hierarchies concept applied
- Normalization - Attributes values are scaled to fall within a small specified range as -1.0 to 1.0 or 0 to 1.0
- Attribute / feature construction - new attributes are constructed / added to existing attributes to help mining process.

Smoothing, Aggregation — Same as on Data Cleaning / Generalization process.

|| form of Data Cleaning

same as forms of Data Reduction.

Normalization –

useful for classification algorithms involving neural networks

✓ Normalizing inputs that help speed up the learning phase

useful for Distance measurements

Methods –

- Min-Max normalization → Range $[new\_min_A, new\_max_A]$
- Z-Score Normalization (zero-mean Normalization) → $(-\infty, \infty)$
- Decimal scaling → $(-1, 1)$

Min-Max normalization

$$v' = \left(\frac{v - min_A}{max_A - min_A}\right)(new\_max_A - new\_min_A) + new\_min_A$$

Z-Score Normalization (zero-mean Normalization)

$$v' = \frac{v - \mu_A}{\sigma_A}$$

$$\mu_A = \frac{1}{N}\sum_{i=1}^{N} x_i \rightarrow \text{mean}$$

$$\sigma_A^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2 \rightarrow \text{variance}$$

~~standard deviation~~

Decimal scaling

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

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

Q) Using Min-Max, Normalize the income from $12,000 to $98,000 into the range [0.0, 1.0].

Ans: Q) Suppose minimum, maximum values of the attribute "Income" are $12000, $98,000 resp. which is mapped to the range [0.0, 1.0]. Find the transformation of income $73,600.

Ans:-

$min_A = 12000$, $max_A = 98000$

$new\_min_A = 0$, $new\_max_A = 1$

value which is to be mapped, $v = 73600$

Mapped value on the range [0.0, 1.0] is v'

$$\Rightarrow v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

$$= \frac{73600 - 12000}{98000 - 12000}(1 - 0) + 0$$

$$= 0.716$$

Q) Normalize the following data by min-max & by z-score

200 300 400 600 1000 [within [0, 1] in min-max]

Ans:

a) $min_A = 200$, $new\_min_A = 0$

$max_A = 1000$, $new\_max_A = 1$

a) $v = 200$

$$\Rightarrow v' = \frac{200 - 200}{1000 - 200}(1 - 0) + 0 = 0$$

b) $v = 300$

$$\Rightarrow v' = \frac{300 - 200}{1000 - 200}(1 - 0) + 0 = 0.125$$

c) $v = 400$

$$v' = \frac{400-200}{800} = 0.25$$

d) $v = 600$

$$v' = \frac{600-200}{800} = 0.5$$

e) $v = 1000$

$$v' = \frac{1000-200}{800} = 1$$

Using z-score :- $v' = \frac{v-\mu_A}{\sigma_A}$

$$\mu_A = \frac{200+300+400+600+1000}{5} = 500$$

$$\sigma_A^2 = \frac{1}{5}\left[(200-500)^2 + (300-500)^2 + (-100)^2 + (100)^2 + (500)^2\right]$$

$$= \frac{400000}{5} = 80000$$

$$\Rightarrow \sigma_A = 283$$

a) $v = 200$

$$\Rightarrow v' = \frac{v-\mu_A}{\sigma_A} = \frac{200-500}{283} = -1.06$$

b) $v = 300$

$$\Rightarrow v' = \frac{300-500}{283} = -0.706$$

c) $v = 400$

$$\Rightarrow v' = \frac{400-500}{283} = -0.35$$

d) $v = 600$

$$\Rightarrow v' = \frac{600-500}{283} = 0.35$$

e) $v = 1000$

$$\Rightarrow v' = \frac{1000-500}{283} = 1.76$$

Data Reduction -

Why data reduction ?

1) Data set may contain Terabytes of data
2) Since huge data set,
   Data Analysis, Data mining } may be complex/tedious
3) Since complex,
   It may take a very long time to analyse / mining.

Data Reduction -

→ applied to obtain a reduced representation of data set that is much smaller in volume & maintain the integrity of the original data

Data Reduction Strategies —

- Data Cube Aggregation
- Attribute Subset Selection
- Dimensionality reduction
- Numerosity reduction
- Discretization

Data Cube Aggregation — Here aggregation operations are applied to the data to form a data cube.

→ Data cube is defined by dimensions & facts

→ It stores multidimensional aggregated information.

Data can be aggregated so that it can be viewed on the abstract level without loss of information & also it takes less volume.

→ Data Cubes provide faster access to precomputed/Summerized data

→ The cube created at the lowest level of abstraction is called Base cuboid.
At the highest level of abstraction, called Apex cuboid.

Attribute Subset Selection :-

→ It reduces the datasize by removing irrelevant or redundant attributes or dimensions.
→ It finds a minimum set of attributes that can be able to represent the original attributes approximately (as close as possible).
→ So that it will be easier to understand a pattern.
→ Heuristic methods are used that reduce the search space for attribute subset selection.
(for n attributes, 2^n subsets can be formed.
If we go on searching, this will be expensive & may take large search space)
eg:- Greedy Approach (until getting the best choice at that time)

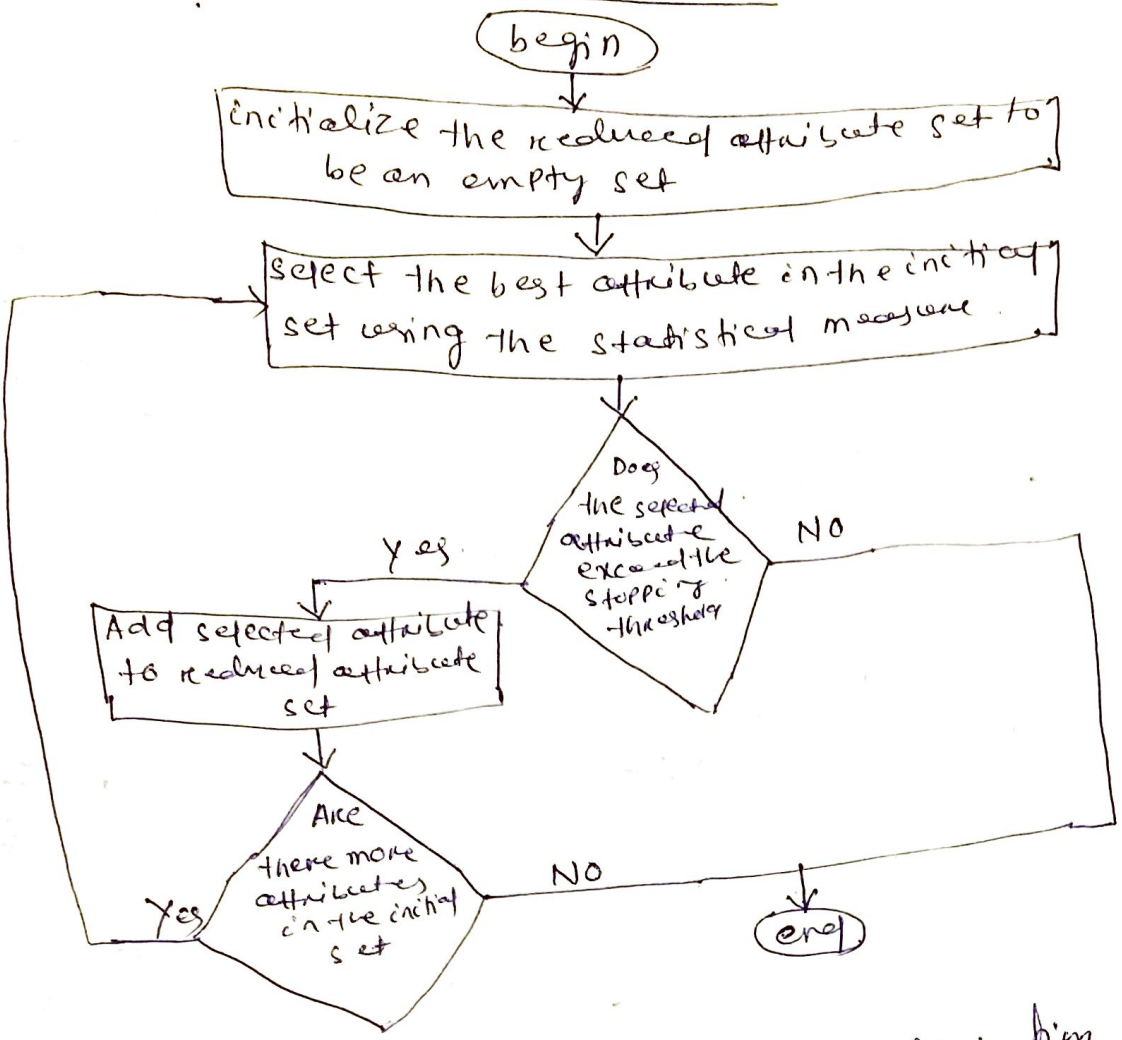
Various heuristic methods include the following techniques for Attribute Subset Selection

- Stepwise forward Selection
- Stepwise backward Elimination
- Combination of backward Elimination & forward selection
- Decision Tree induction

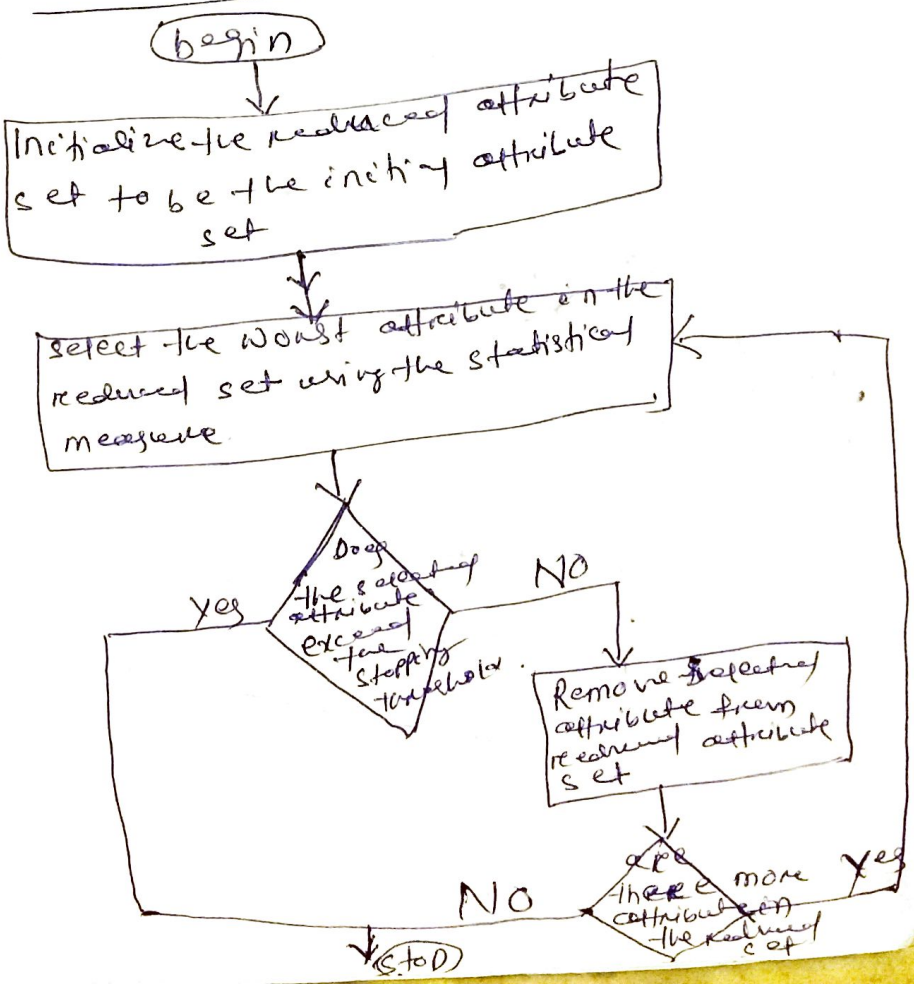
Stepwise forward Selection: It starts with empty attribute sets. When best is determined it is added to attribut set. & so on.

Combination of backward Elimination & forward selection: It starts with the full set of Attributes. The worst one is deleted & so on

Stepwise forward selection



Stepwise backward Elimination



→ In Decision tree induction, a tree is constructed from the given data.

→ Set of attributes not appearing in tree are all irrelevant.

→ Set of attributes appearing form the reduced subset of attributes.

Dimensionality Reduction :-

→ Data are reduced/compressed by applying Data encoding or transformation.

→

- Lossless
  - original data can be reconstructed without any loss of information
- Lossy
  - Approximation of original data is reconstructed
  - (Adv- Logical manipulation / processing requ

2 popular lossy dimensionality reduction.

- wavelet transforms
- Principal Component Analysis

Numerosity Reduction :-

→ It reduces data volume by choosing smaller forms of data representation.

Techniques used

Parametric Method

1) It is assumed that data fits some model
2) The model estimates the data
3) Only the data parameters are stored instead of actual data.

eg: Regression, log linear model

- Regression → can be applied to high dimensionality data
- log linear model → upto 10

Non-parametric Method

1) no model is assumed.
2) It stores reduced representation of data in the form of either Histograms or Clustering or Sampling.

Data Discretization :-

→ It is a method of replacing numerous values of a continuous attribute by a small no. of interval labels

→ Thus it reduces & simplifies the orginal data

Types / Category

Supervised
vs.
Un-Supervised
Discretization

⇓

If the discretization process uses class information, it is called Supervised discretization otherwise un-supervised.

i.e if process proceeds from top to down

Top-down (Splitting)
vs.
Bottom-up (Merging)
Discretization

⇓

If the process starts by first finding one or few interval to split the entire attribute range then repeats this recursively on the resulting interval it is called Splitting

Typical Methods -

1) Binning -
Top-Down Split, unsupervised

2) Histogram Analysis -
Top Down split, unsupervised

3) clustering Analysis -
Either top down or bottom up & merge split
unsupervised.