**Clustering**

You are given a data without labels;

You are dividing data into constituent clusters

Devise a algorithm to know no of clusters.

Some data may have different clustering which are equally nice.

Clustering is the beginning to Data Analysis.

1) No of clusters

2) Similarity/Dissimilarity Measures;

Two points are similar if the value of similarity is more and the value of dissimilarity is less.

Two points are dissimilar if the value of dissimilarity is more and the value of similarity is less;

By clustering we would like to get some information about labeling; your clusters are good. Your clusters indicating the whole population. [Reflects the division of populations]

Clustering Validity: Division in the dataset is called clusters.

Two assumptions are made:

1) No of clusters (C ) is known

2) Give points belong to (Not always true)

Dissimilarity Measurement: Distances

Similarity Measurement: cosine [ Angle between two vectors]

If Angle is 0 then it is similar.

You may want to get cluster which is having following some properties:

**1) Minimum within cluster distance criterion**

Let the no of clusters C be known

Let [] be the given set of points

Let be the Euclidean distance between and [ Dissimilarity measure]

Let partition be defies

a)

b) [ Taking S partition in C subset]

c)

which satisfies three above properties;

For a partition

Let ;

(No of points)

For each

Then

[Within cluster distance in the cluster];

Now choose such that

Loss associated with the optimal partition;

We want to get a partition of the dataset with some properties and that properties will optimize with all the partition.

K-Means Clustering

K-> No of Means

Several version of K-means available