Non-Hierarchical Clustering

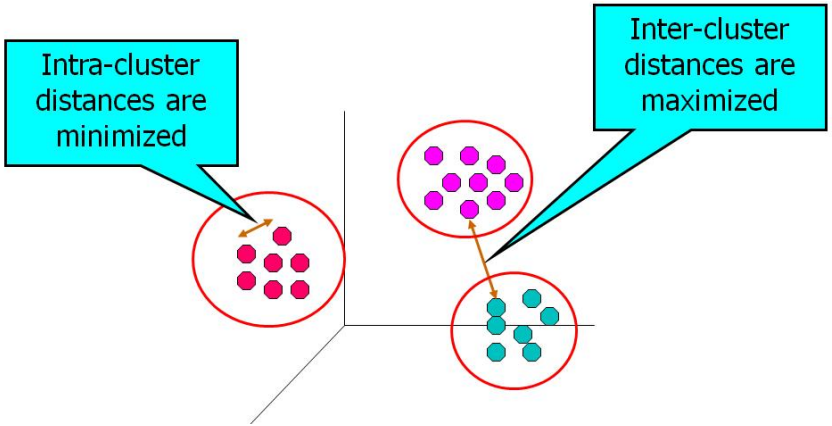
K Means Method

Contents

Cluster Analysis

Cluster analysis is a class of statistical techniques that can be used to classify objects or cases into groups called **Clusters.**

- A cluster is a group of relatively homogeneous cases or observations.
- The observations are dissimilar to objects outside the cluster, particularly objects in other clusters.



- Cluster Analysis is one of the unsupervised learning method.

K-Means Clustering

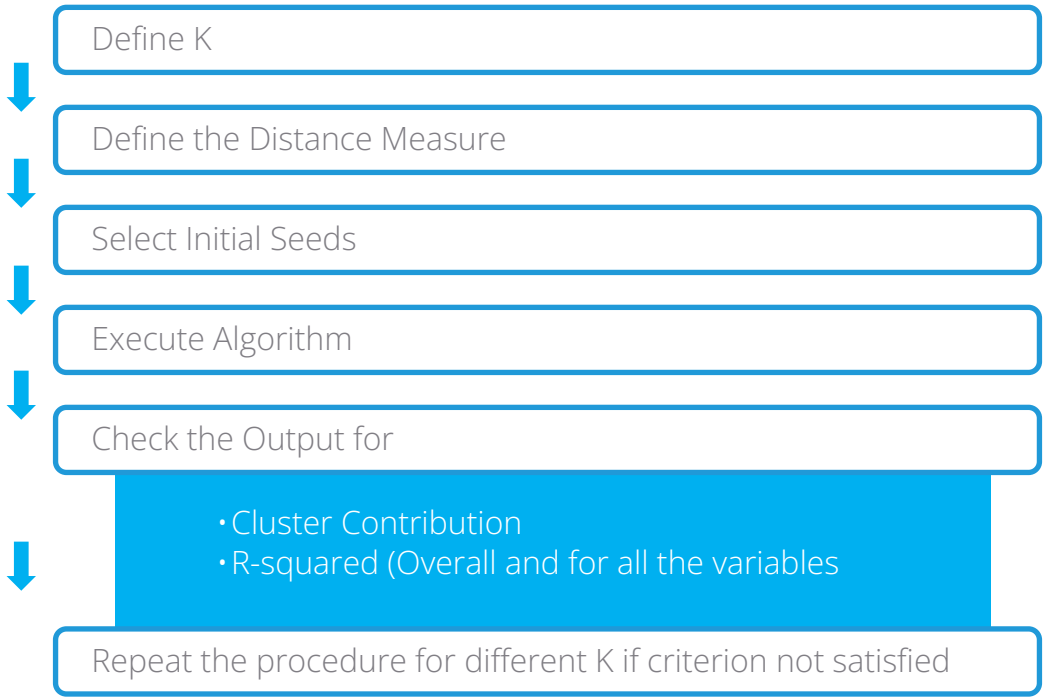
- K-Means Clustering is one of the most popular non-hierarchical clustering methods

- K –Means method is suitable for large data sets and widely used for customer segmentation in BFSI or retail domains

- The number of clusters (k) must be known a priori
  (Though in reality this may not be the case)

- Alternatively, cluster solutions can be observed for different k and evaluated to get the best possible cluster solution

K-Means Clustering – Steps



Distance Measures

- Clustering algorithms require a **mathematical measure to assess the similarity of a pair of observations or clusters**

Object	X1	X2			XP
1	a1	a2			ap
2	b1	b2			bp

- Manhattan Distance

The sum of the absolute differences in values of P variables

$$d(x , y) = | a1-b1 | + | a2-b2 | + ... + | ap-bp |$$

- Chebyshev Distance

The maximum absolute difference in values of P variables

$$d(x , y) = Max (| ai-bi |)$$

Distance Measures – Euclidean Distance

- **Squared Euclidean Distance: The** sum of squared differences between values of each variable

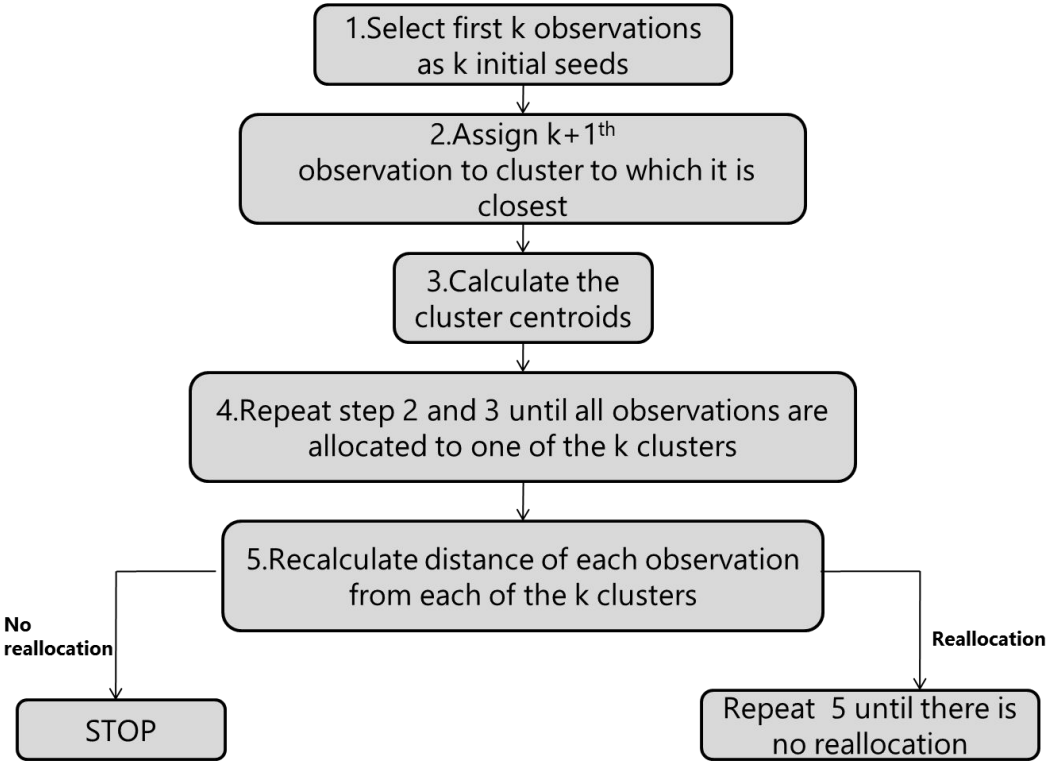
$$d(x , y) = (a1-b1)^2+(a2-b2)^2+....+(ap-bp)^2$$

- **The square root is defined as 'Euclidean Distance'**
- 'Euclidean Distance' is the most widely used distance measure in cluster analysis

Choice of Initial Seeds

- There are **different methods to decide initial seeds**, some of them are:
  - K-random observations
  - First K observations
  - Last K observations
  - Partition the data into k partitions randomly and then use the partition mean/ median as initial seeds

Algorithm


1.Select first k observations as k initial seeds
2.Assign k+1th observation to cluster to which it is closest
3.Calculate the cluster centroids
4.Repeat step 2 and 3 until all observations are allocated to one of the k clusters
5.Recalculate distance of each observation from each of the k clusters
No reallocation
Reallocation
STOP
Repeat 5 until there is no reallocation

Data Snapshot

Town Insurance

Variables

Observations

Town	x1	x2	x3
A	1.06	9.2	151
B	0.89	10.3	202
C	1.43	15.4	113
D	1.02	11.2	168
E	1.49	8.8	192
F	1.32	13.5	111
G	1.22	12.2	175
H	1.1	9.2	245

Columns	Description	Type	Measurement	Possible values
Town	Towns under study	character	-	-
x1	Loss Ratio	numeric	-	Positive values
x2	Premium Rates	numeric	-	Positive values
x3	Number of Policies	numeric	-	Positive values

Iteration 1

Data:

Town	x1	x2	x3
A	1.06	9.2	151
B	0.89	10.3	202
C	1.43	15.4	113
D	1.02	11.2	168
E	1.49	8.8	192
F	1.32	13.5	111
G	1.22	12.2	175
H	1.1	9.2	245

K=2

Step 1 : Initial Seeds

Initial Seeds

Town	x1	x2	x3
A	1.06	9.2	151
B	0.89	10.3	202

Iteration 1

Step 2: Find distance of Town C from A (Cluster1) and B(Cluster2)

Distance of C from A = $\sqrt{(1.43-1.06)^2+(15.4-9.2)^2+(113-151)^2}$ = **38.50**

Distance of C from B = √(1.43-0.89)2+(15.4-10.3)2+(113-202)2 = 89.15

Minimum Distance = **38.50**

Since distance between Town C and Town A is minimum, Town C will be combined with Town A

Updated cluster centroids (means) are:

Cluster	Members	X1	X2	X3
1	A,C	1.245	12.3	132
2	B	0.89	10.3	202

* Note : Values of x1,x2,x3 for 1st cluster are mean values of A & C

Iteration 1

Step 3: Find distance of Town D from Cluster1 and Cluster 2

Distance of D from Cluster 1	**36.018**
Distance of D from Cluster 2	**34.012**

Minimum Distance = 34.012

Here Town D will be combined with Town B (Cluster 2)

Updated cluster means are:

Cluster	Members	X1	X2	X3
1	A,C	1.245	12.3	132
2	B,D	0.955	10.75	185

Iteration 1

Step 4: Find distance of Town E from Cluster 1 and Cluster 2

Distance of E from Cluster 1	**60.102**
Distance of E from Cluster 2	**7.2862**

Minimum Distance = 7.2862

Here Town E will be combined with Cluster 2 (i.e. with Towns B & D)

Updated cluster means are:

Cluster	Members	X1	X2	X3
1	A,C	1.245	12.3	132
2	B,D,E	1.133333	10.1	187.3333

Iteration 1

Step 5: Find distance of Town F from Cluster 1 and Cluster 2

Distance of F from Cluster 1	**21.034**
Distance of F from Cluster 2	**76.409**

Minimum Distance = 21.034

Here Town F will be combined with Cluster 1 (i.e. with Towns A & C)

Updated cluster means are:

Cluster	Members	X1	X2	X3
1	A,C,F	1.27	12.7	125
2	B,D,E	1.133333	10.1	187.3333

Iteration 1

Step 6: Find distance of Town G from Cluster 1 and Cluster 2

Distance of G from Cluster 1	**50.003**
Distance of G from Cluster 2	**12.511**

Minimum Distance = 12.511

Here Town G will be combined with Cluster 2 (i.e. with Towns B,D & E)

Updated cluster means are:

Cluster	Members	X1	X2	X3
1	A,C,F	1.27	12.7	125
2	B,D,E,G	1.155	10.625	184.25

Iteration 1

Step 7: Find distance of Town H from Cluster 1 and Cluster 2

Distance of H from Cluster 1	**120.05**
Distance of H from Cluster 2	**60.767**

Minimum Distance = 60.767

Here Town H will be combined with Cluster 2 (i.e. with Towns B,D,E & G)

Updated cluster means are:

Cluster	Members	X1	X2	X3
1	A,C,F	1.27	12.7	125
2	B,D,E,G,H	1.144	10.34	196.4

Since all the towns are assigned to two clusters, to verify our clusters membership we go for the next iteration

Iteration 2

In iteration 2, initial seeds will be those two clusters which are obtained at the end of iteration 1

Step 1:

Initial seeds				
Cluster	**Members**	**X1**	**X2**	**X3**
1	A,C,F	1.27	12.7	125
2	B,D,E,G,H	1.144	10.34	196.4

Iteration 2

Step 2: Find the Distance of Town A from Cluster 1 (i.e from combined Towns A,C &F) and then Cluster 2(i.e from combined Towns B,D,E,G & H)

Distance of A from Cluster 1 = √(1.06-1.27)2+(9.2-12.7)2+(151-125)2 = 26.23536

Distance of A from Cluster 2 = √(1.06-1.44)2+(9.2-10.34)2+(151-196.4) = 45.41439

Minimum Distance = 26.23536

Since distance between Town A and Cluster 1 is minimum, Town A will be retained in Cluster 1

Iteration 2 – Summary

Initial seeds				
Cluster	**Members**	**X1**	**X2**	**X3**
1	A,C,F	1.27	12.7	125
2	B,D,E,G,H	1.144	10.34	196.4

No town is reassigned to different cluster

This is final cluster solution

Town	Distance from Cluster1	Distance from Cluster2	Cluster
A	26.24	45.41	1
B	77.04	5.61	2
C	12.30	83.55	1
D	43.03	28.41	2
E	67.11	4.67	2
F	14.02	85.46	1
G	50.00	21.48	2
H	120.05	48.61	2

Points For Good Cluster Solution

- **Standardize variables if scale differs widely**. Variables with high variance tend to influence cluster solution.
- Use **data reduction technique like factor analysis before cluster analysis if number of variables is high**
- **Run the algorithm for different choices of K and initial seeds**
- **Use dummy variables for nominal scaled variables**. K means algorithm is not suited for nominal scaled variables.
- **You may use hierarchical clustering for sample of the data to get preliminary information on K and distance values**

Statistics Associated with Cluster Solution

Cluster solution can be assessed using 'between clusters' variability and 'within clusters' variability.

Within Sum of Squares (WSS) is a measure to explain homogeneity within a cluster. WSS can be calculated for each cluster and then added to get Total WSS

Total WSS should be small

R-squared is computed as ratio of Between Clusters Variability to Total Variability.

R-squared should be large

Quick Recap

Cluster Analysis	• Statistical techniques that can be used to classify objects or cases into groups called Clusters • Cluster is a group of relatively homogeneous cases or observations
K-Means Clustering – Steps Involved	• Define k • Define Distance Measure • Define Initial Seeds • Assign cases to clusters based on distance measure • Repeat the Procedure Until no reassignment is required