Hierarchical Clustering

Kuldeep Dhole

Supervised By: Dr. Pollett

What is clustering?

The process of grouping a set of objects into classes of similar objects

What is Hierarchical clustering?

It is a method of cluster analysis which seeks to build a hierarchy of clusters.

Strategies:

- Agglomerative (Bottom-up)
- Divisive (Top - down)

Agglomerative approach is preferred.

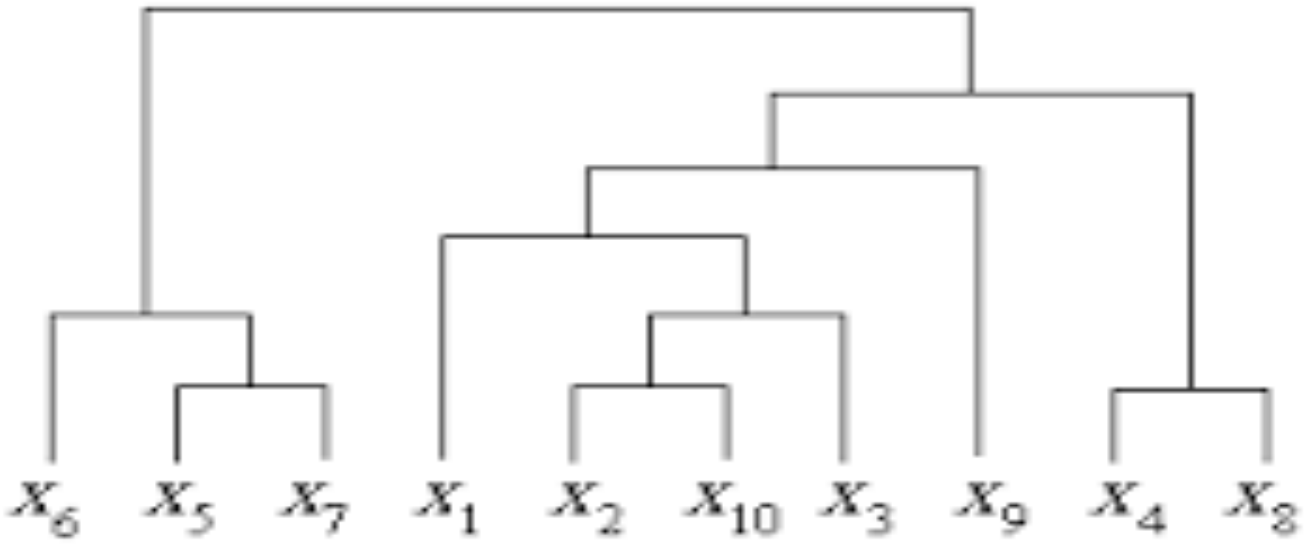
Agglomerative (Bottom Up) Approach

-Given n points p1, p2, … pn; We assume all as different clusters c1, c2, c3,.... cn

-Algorithm is pretty intuitive and simple:

```
num = #n
while (num > 1):
    find the pair of closest distance points pi,pj i.e. ci,cj
    form a new cluster ci,j = ci + cj
    remove ci and cj
```

Dendrogram



Closest Distance

-Let’s assume we are in 2D space, and we have points p1, p2
-Approaches to get closes distances:
  -Euclidean Distance:
   [(x1-x2)^2 + (y1-y2)^2]^0.5
   In multidimensional space,
   [(a1-a2)^2 + (b1-b2)^2 + (c1-c2)^2 + ……]^0.5
  -Manhattan Distance:
   |x1-x2| + |y1-y2|
   In multidimensional space,
   |a1-a2| + |b1-b2| + |c1-c2| + ……

Distance Calculations continued...

How to calculate distance between:
-a point p and cluster c2 OR
-a cluster c1 and cluster c2

Approaches:
- Single-link: We pick the point of closest distance from c2
- Complete-link: We pick the point of furthest distance from c2
- Centroid: We pick the centre of gravity from c2

Computational Complexity

-Initially, we need to compute distances of all pairs of n individual points => O(m n^2)

-In each loop of the algorithm, we compute the closest distance between most recently created cluster and other clusters.

-Using heap, we can achieve this by O(m n^2 log(n^2))

Applications

- Clustering search results for efficient navigation e.g. “jaguar” -> c1 = car, c2 = animal, c3 = apple inc

- Citation ranking e.g. google scholar