Hierarchical Clustering

What it is & How it works

AI 05 Hyungjin Kim

Hierarchical Clustering

Hierarchy

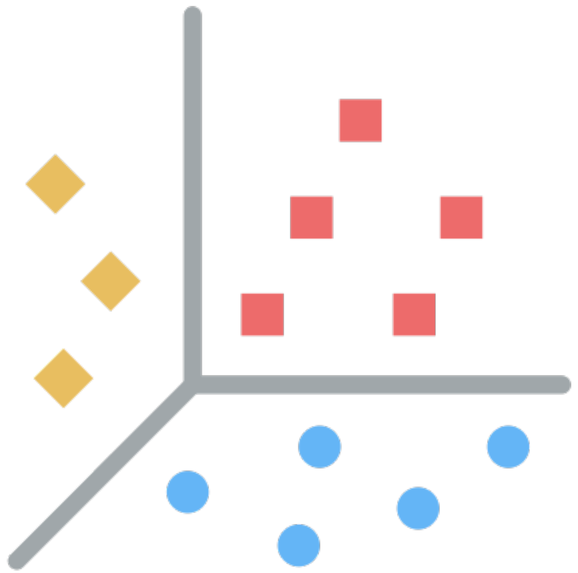
Clustering

any system of people or things **ranked** one above another

Hierarchical Clustering

Hierarchy Clustering

grouping data points based on specific characteristics

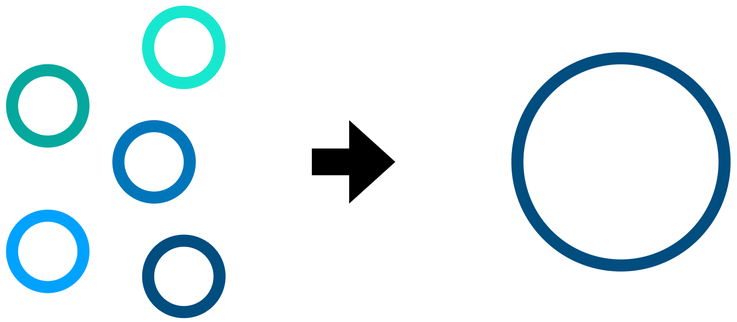


Hierarchical Clustering

Types of HC

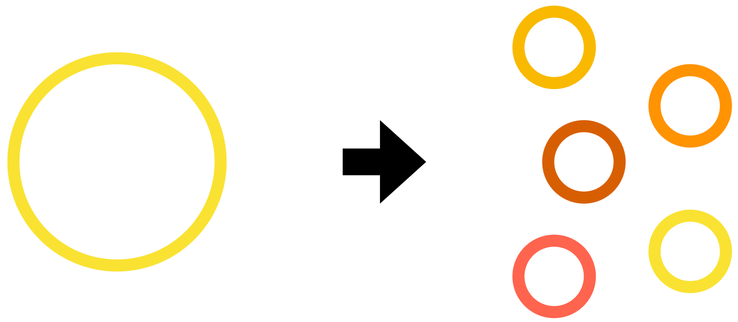
Agglomerative

bottom - up

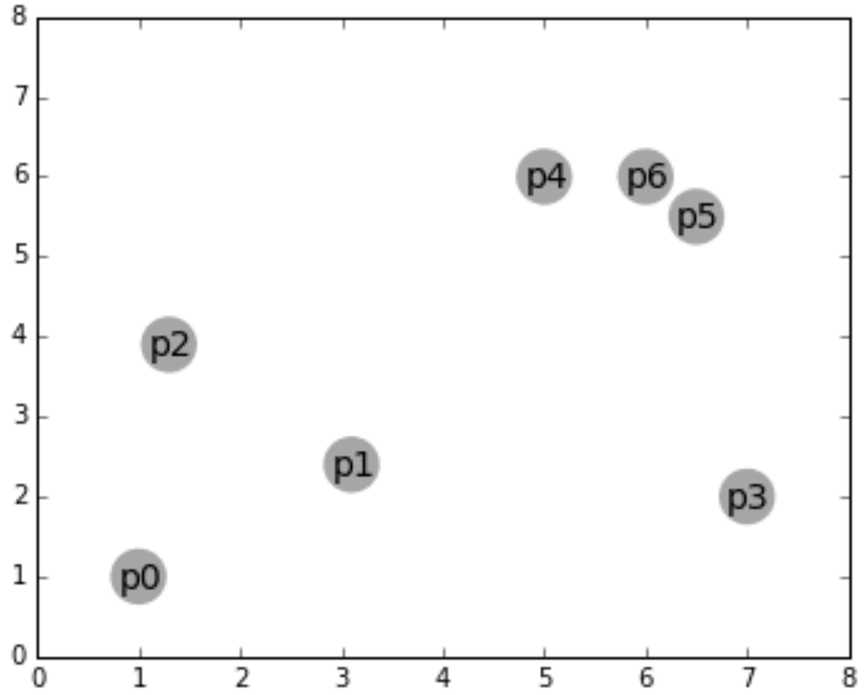


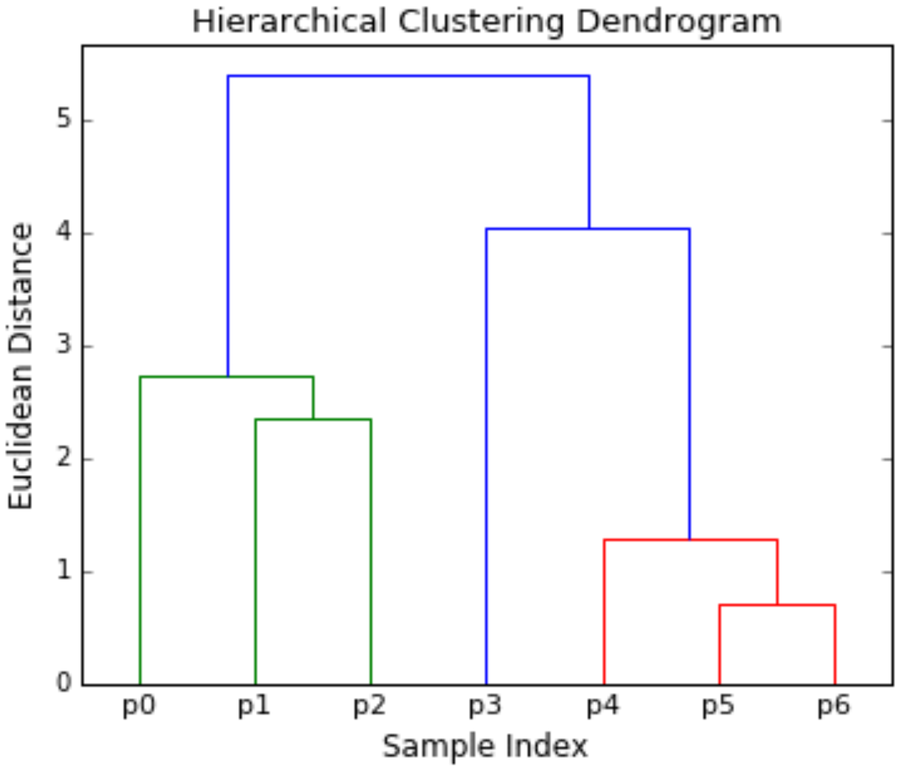
Divisive

top - down



Agglomerative Hierarchical Clustering

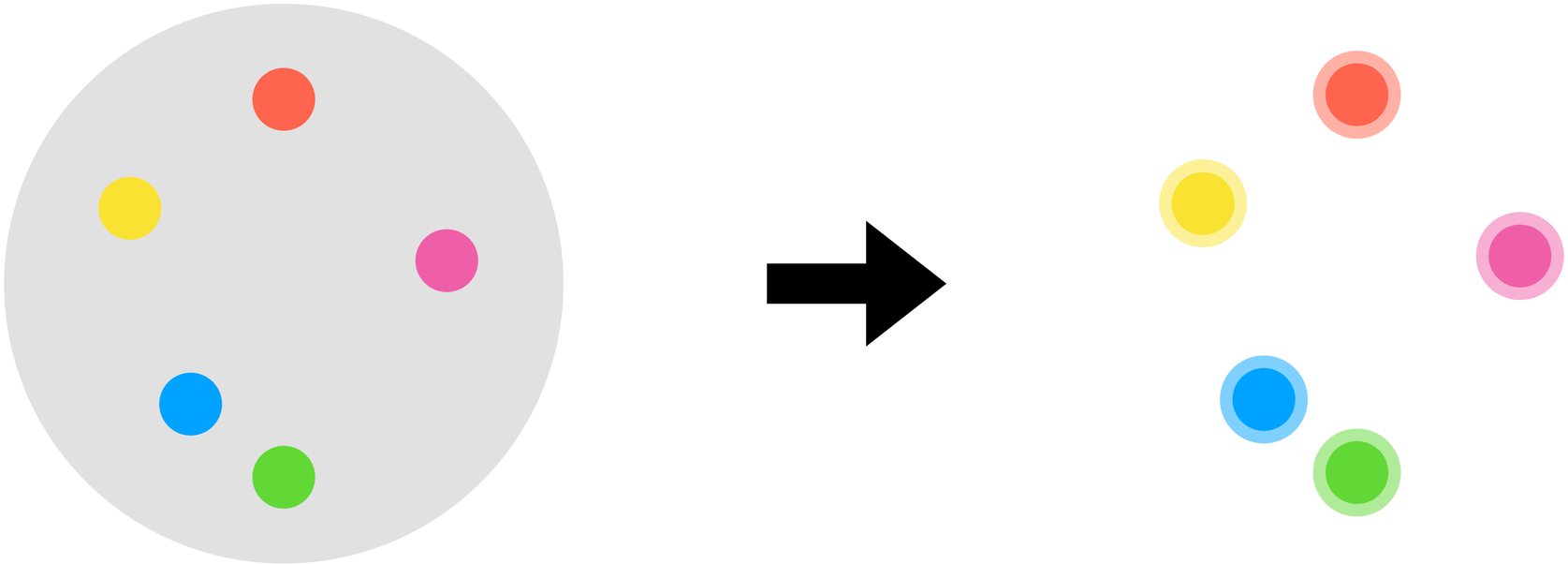




Merging clusters at each iteration

Divisive Hierarchical Clustering

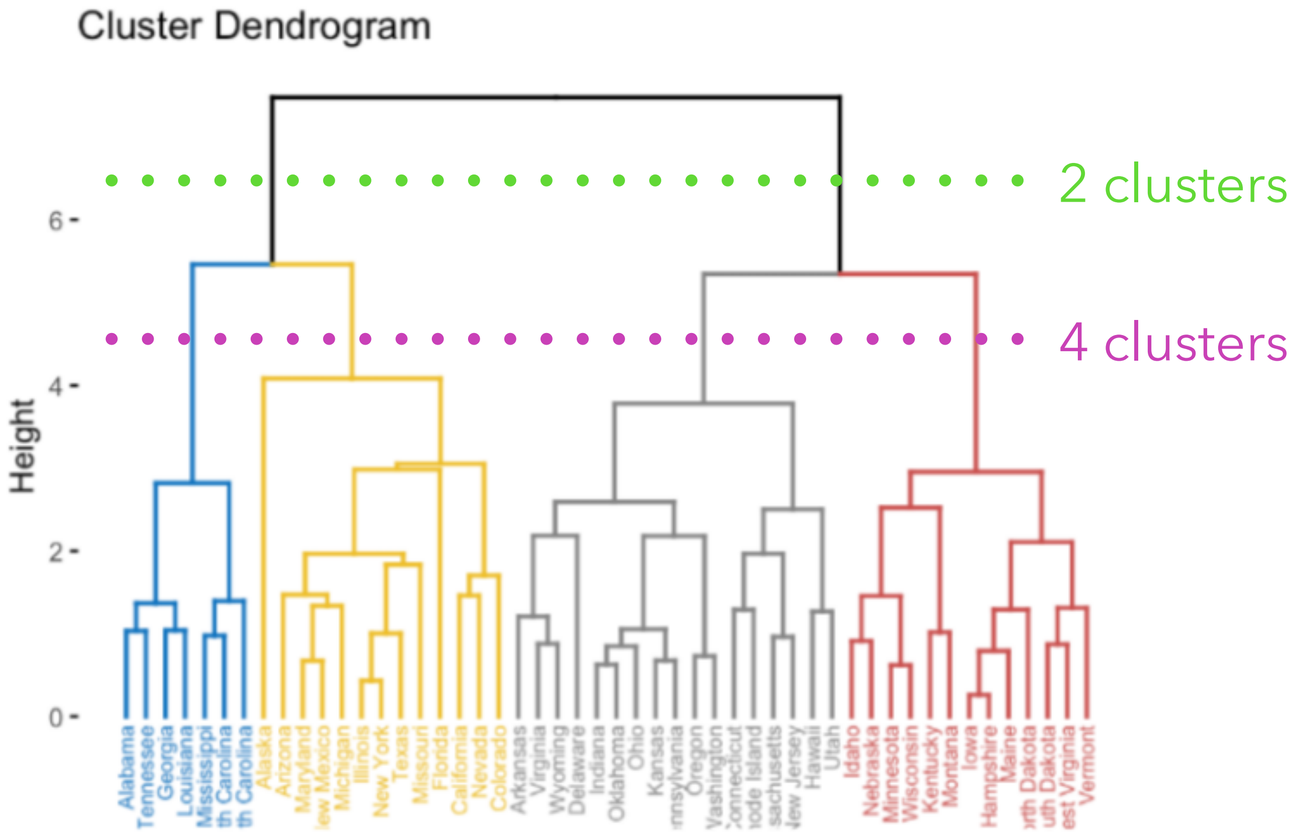
opposite way of agglomerative clustering



Splitting clusters at each iteration

Dendrogram

Result of HC (Agglomerative)



tree-like diagram that **records the sequences of merges or splits**

How it works?

Step by Step

Process of training

compute **Proximity Matrix**

let each data points be a cluster

Repeat

merge the two **closest** clusters

update the **Proximity Matrix**

Until only a single cluster remains

Step by Step

Process of training

compute **Proximity Matrix** tells us the distance between each data points (n x n)

let each data points be a cluster

Repeat

merge the two **closest** clusters

update the **Proximity Matrix**

Until only a single cluster remains

Distance Metrics

Euclidean distance

Squared Euclidean distance

Manhattan (city block) distance

Maximum distance

Step by Step

Process of training

compute **Proximity Matrix**

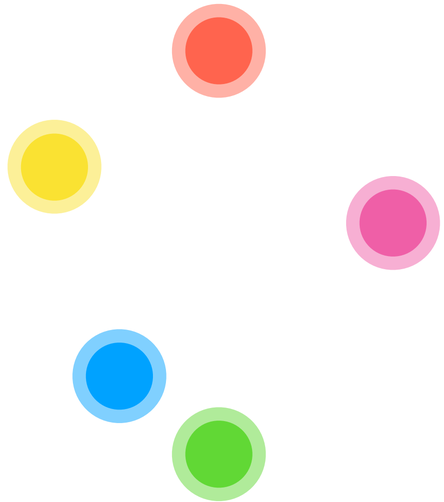
let each data points be a cluster

Repeat

merge the two **closest** clusters

update the **Proximity Matrix**

Until only a single cluster remains



Step by Step

Process of training

compute **Proximity Matrix**

let each data points be a cluster

Repeat

merge the two **closest** clusters

update the **Proximity Matrix**

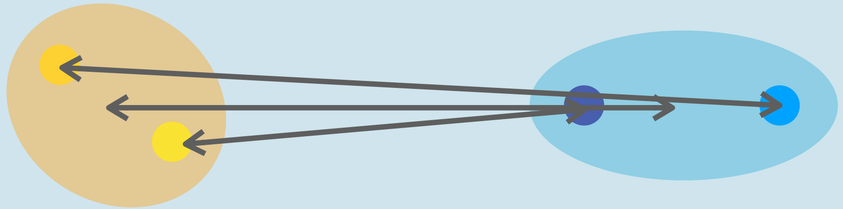
Until only a single cluster remains

Linkage Criteria

Min, Max, Group Average

Distance between Centroids

Ward's Method



Pros & Cons of HC

Should we use it or not?

Pros

No need to **pre-define** the number of clusters

Easy to understand

May correspond to **meaningful taxonomies**

Cons

High time & space complexity $O(N^3)$

Sensitive to outliers & noise

References

- https://www.analyticsvidhya.com/blog/2019/05/beginners-guide-hierarchical-clustering/
- https://todayisbetterthanyesterday.tistory.com/59
- https://www.displayr.com/what-is-hierarchical-clustering/
- https://www.displayr.com/strengths-weaknesses-hierarchical-clustering/
- https://cs.wmich.edu/alfuqaha/summer14/cs6530/lectures/ClusteringAnalysis.pdf
- https://towardsdatascience.com/understanding-the-concept-of-hierarchical-clustering-technique-c6e8243758ec
- https://en.wikipedia.org/wiki/Hierarchical_clustering
- https://dashee87.github.io/data%20science/general/Clustering-with-Scikit-with-GIFs/

Thank You 💙