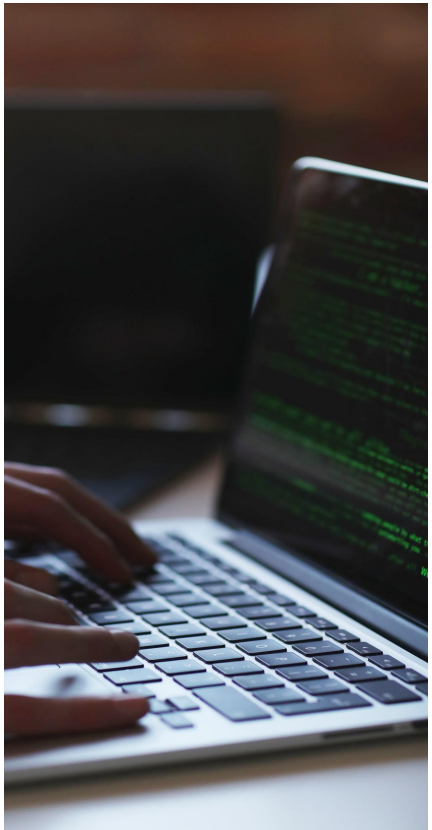
Course-Net
Empowering IT People

#EmpoweringITPeople

Introduction to Clustering

Course-Net
Empowering IT People

#EmpoweringITPeople



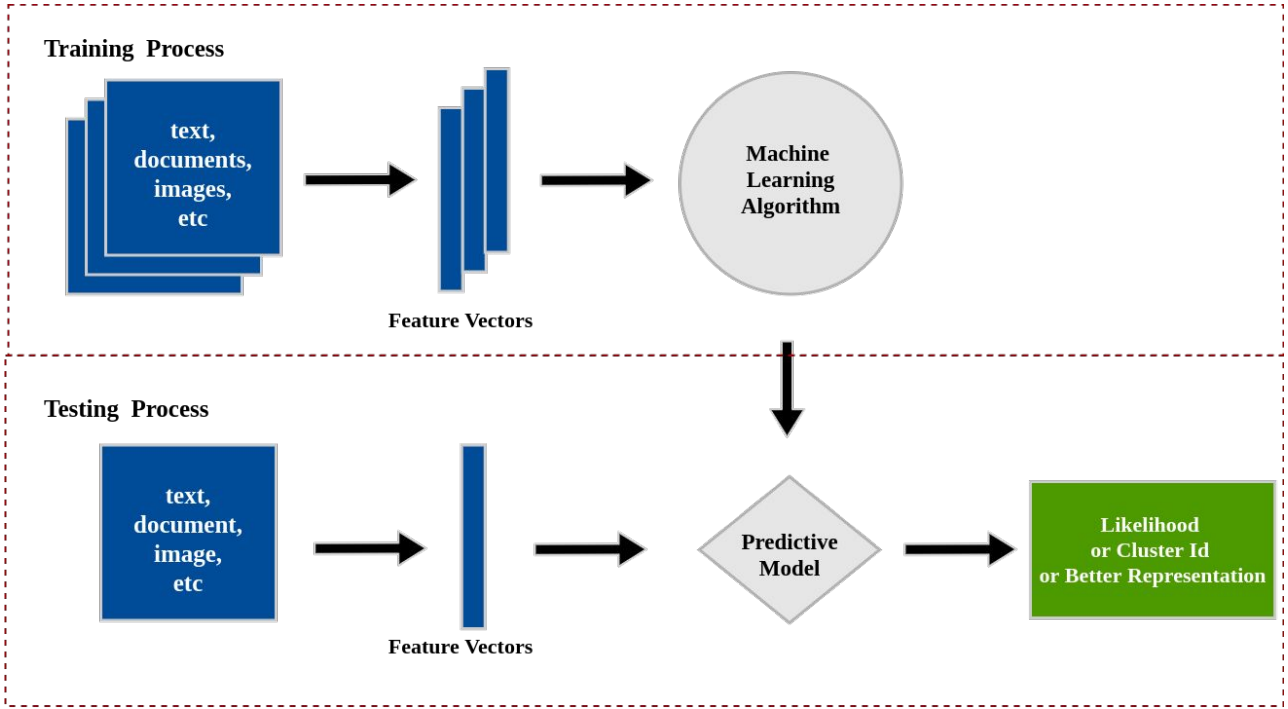
APA YANG AKAN KITA PELAJARI

Clustering and applications

- What is Unsupervised Learning
- Clustering Intuition
- Clustering Algorithm
- Applications of Clustering
- Clustering Cases

Unsupervised Learning:
The learning method
finding new possibilities
or/and patterns without
any targets from the
given data

Unsupervised Learning Process



Course-Net
Empowering IT People

#EmpoweringITPeople

Apa itu Clustering Mengapa Kita Pelajari di Data Science ?

don’t know what the groups resulting from an analysis might be, you should use clustering technique, in order to discover new possibilities and patterns.

Clustering Algorithm

1. K-Means
2. Density Based Clustering
3. Hierarchical Clustering
4. Gaussian Mixture Model

K-means Cases in Industries

Identify Subscriptions

Loyalty Customers Identifications

Virus Classes Clustering

A Glance on Clustering Data Example

CustomerID	Gender	Age	Annual Inc...	Spending ...
1	Male	19	15	39
2	Male	21	15	81
3	Female	20	16	6
4	Female	23	16	77
5	Female	31	17	40

X(Independent) :

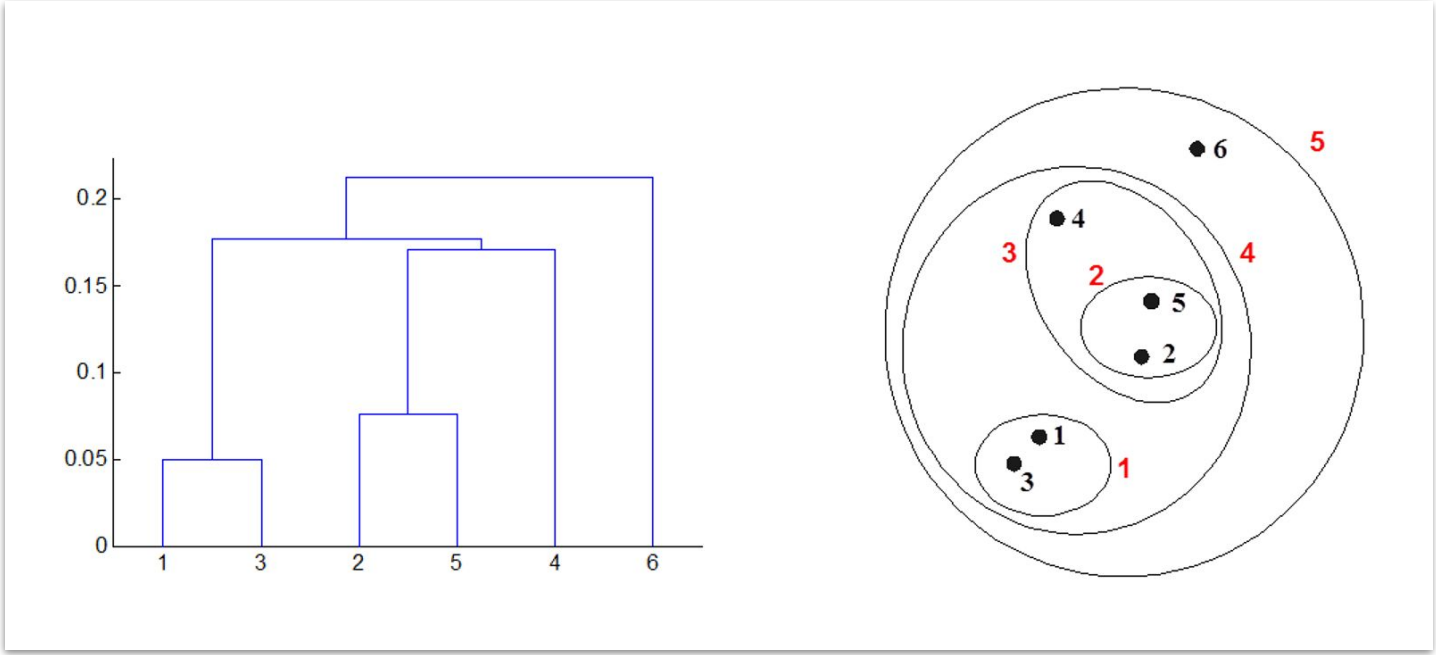
- CustomerID
- Gender
- Age
- Annual Income
- Spending Score

Mall visitors dataset

Course-Net
Empowering IT People

#EmpoweringITPeople

Hierarchical Clustering Intuition


0.2
0.15
0.1
0.05
0
1
3
2
5
4
6
6
5
4
3
2
4
5
2
1
3
1

Hierarchical Types

1. **Agglomerative (Bottom-Up)** -> Each data is a cluster of its own, further pairs of clusters are merged as one moving up hierarchy
   - Single Linkage (Min)
   - Complete Linkage (Max)
   - Average Linkage (Mean)
1. **Divisive (Top-Down)** -> All data points in dataset belong to one cluster and split is performed recursively as one move down the hierarchy

Agglomerative Steps

1. Compute distance matrix between data points
2. Merge each data point as a cluster depends on the parameter types
3. Repeat:

   Update the distance matrix to represent the new cluster between each data point and the remains

1. **Until : single cluster remains**

Distance Matrix

1. Manhattan Distance

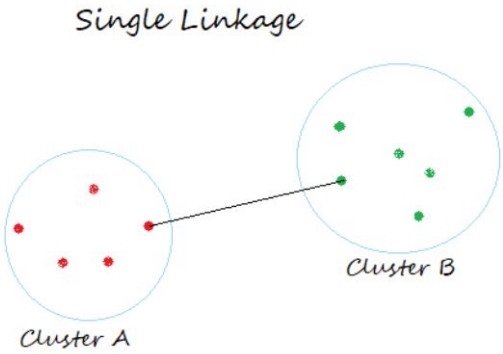
$$d(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

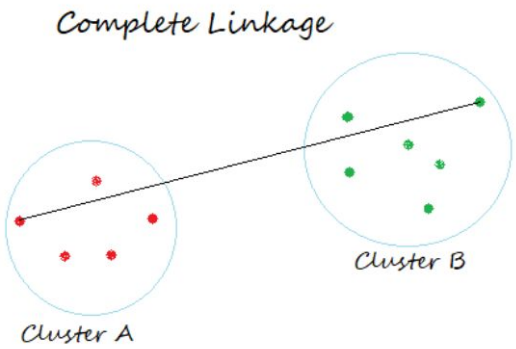
1. Euclidean Distance

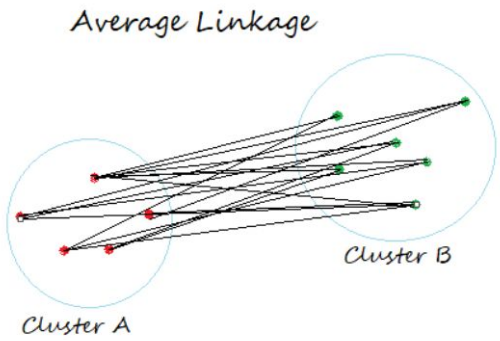
$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

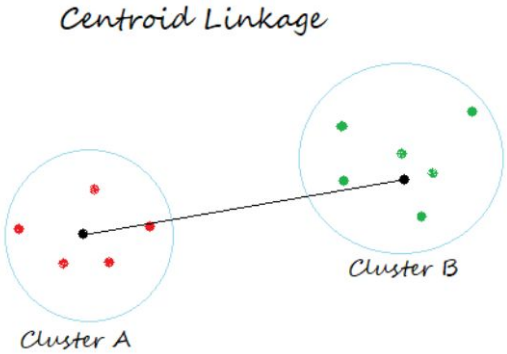
1. Minkowski Distance

$$D(X, Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{1/p}$$









https://medium.com/@tribinty/algorithm-agglomerative-hierarchical-clustering-31d2cea14d9

Course-Net
Empowering IT People

#EmpoweringITPeople

K-means Clustering Intuition

K-Means Steps

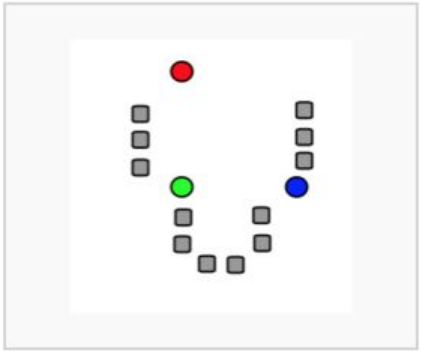
1. Select a K number of clusters (min K = 2)
2. Select the centroids at random K points
3. Assign each data point to their closest centroid

Repeat :

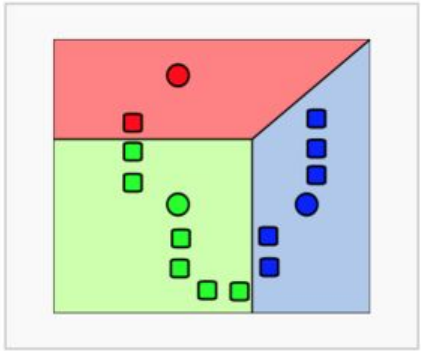
Determine and place the new centroid of each new cluster

1. Reassign each data point to the new closest centroids
2. Until : Converges (Centroids doesn't changes)

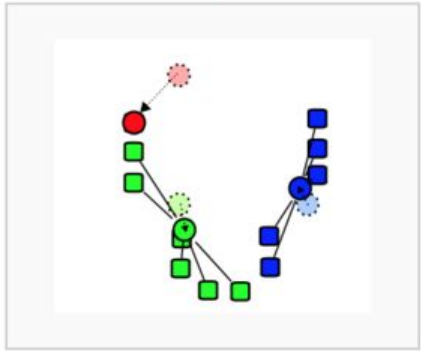
Demonstration of the standard algorithm



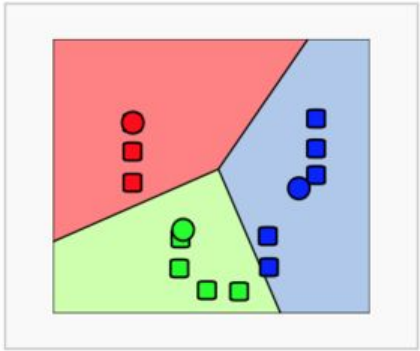
1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).



2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



3. The centroid of each of the *k* clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.