Data Science Workshop

Clustering Analysis

Toulon, February 2019

Emmanuel Génard – *genard@fr.ibm.com*
Data Scientist & Cloud Developer Advocate Europe,
IBM Business Solution Center Nice, France

ISEN
ALL IS DIGITAL!
MÉDITERRANÉE

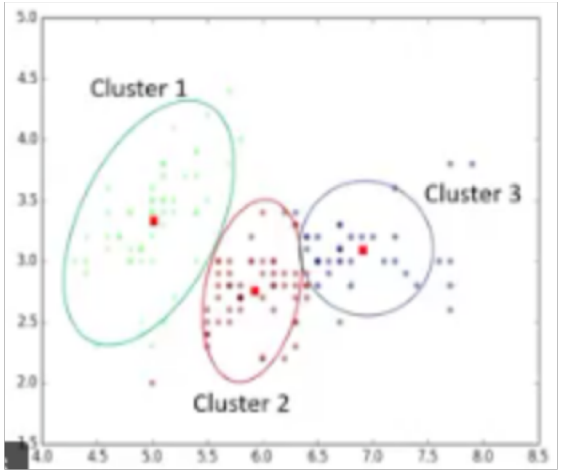
Objectives

- What is Cluster Analysis?
- What are the types of Clustering techniques?
- Why do we use Clustering?

What is a cluster?

A group of objects that are **similar to other objects** in the cluster, and **dissimilar to data points** in other clusters.

Clustering can group data only **« unsupervised »**, based on the similarity of customers to each other.



Why do we cluster?

Clustering : given a collection of data objects group them so that:

- Similar to one another within the same cluster
- Dissimilar to the objects in other clusters

Clustering results are used:

- As a **stand-alone tool** to get insight into data distribution
  Visualization of clusters may unveil important information
- As a **preprocessing step** for other algorithms
  Efficient indexing or compression often relies on clustering

Clustering Applications

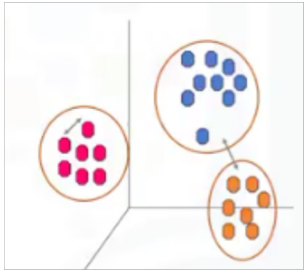
- **Retail and Marketing:** find associations among customers based on their demographic characteristics and use that information to identify buying patterns of various customer groups. Recommendation systems to find a group of similar items or similar users, and use it for collaborative filtering, to recommend things like books or movies to customers.
- **Banking:** analysts find clusters of normal transactions to find the patterns of fraudulent credit card usage. Also, they use clustering to identify clusters of customers, for instance, to find loyal customers, versus churn customers.
- **Insurance:** In the Insurance industry, clustering is used for fraud detection in claims analysis, or to evaluate the insurance risk of certain customers based on their segments.

Clustering Applications

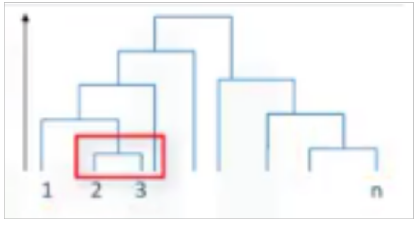
- **Media and Publication:** clustering is used to auto categorize news based on its content, or to tag news, then cluster it, so as to recommend similar news articles to readers.
- **In Medicine:** it can be used to characterize patient behavior, based on their similar characteristics,
  so as to identify successful medical therapies for different illnesses. Or,
- **Biology:** clustering is used to group genes with similar expression patterns, or to cluster genetic markers to identify family ties.

Clustering Algorithms

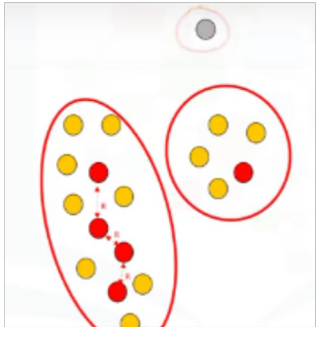
- **Partitioned-based Clustering**
  - Relatively efficient
  - K-Means, K-Median, Fuzzy c-Means



- **Hierarchical Clustering**
  - Produces Trees of Clusters
  - Agglomerative, Divisive



- **Density-Based Clustering**
  - Produces arbitrary shaped clusters
  - DBSCAN

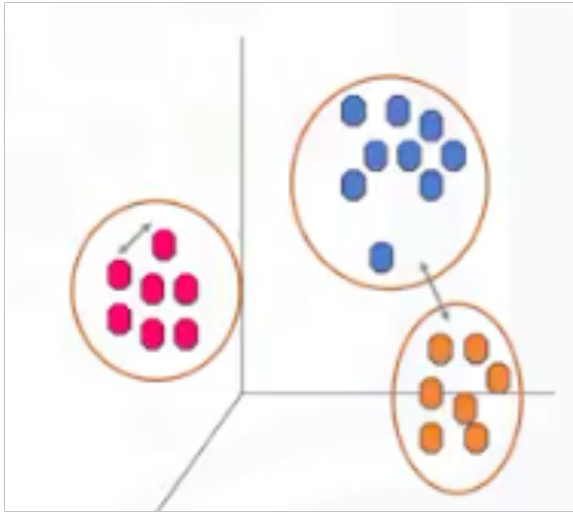


K-Means Clustering?

Basic Idea:
It is a type of **partitioning clustering**, that is it **divides** the data into ***k*** non-overlapping subsets (or clusters) without any cluster-internal structure or labels.

Objects within a cluster are very similar and objects across different clusters are very different or dissimilar.

k-Means tries to minimize the “intra-cluster” distances and maximize the “inter-cluster” distances.

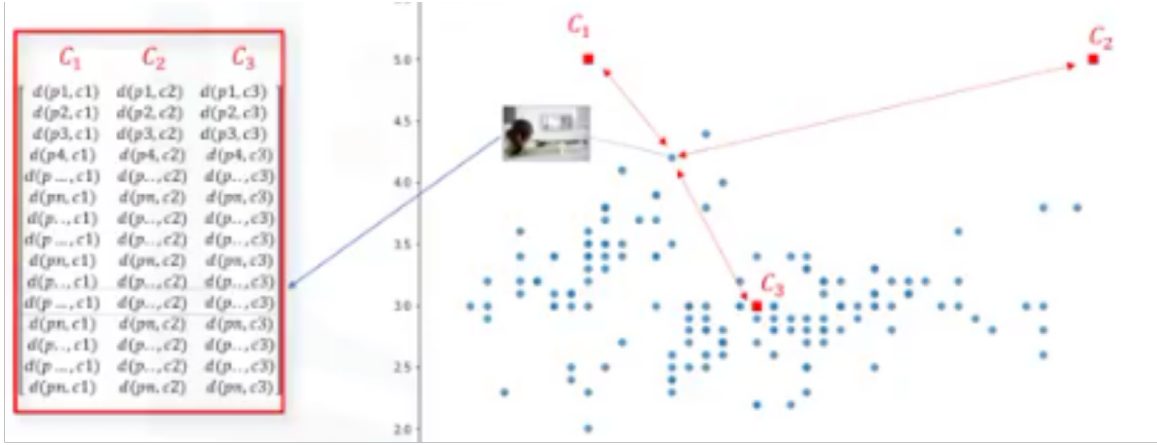


K-Means Clustering Algorithm

1. Define the number of clusters and clusters' centroids

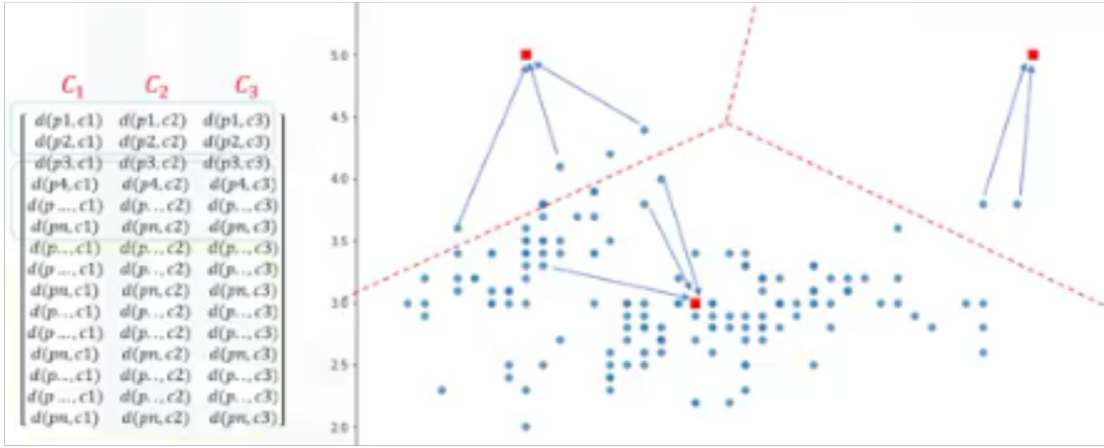
K-Means Clustering Algorithm

1. Define the number of clusters and clusters' centroids
2. **Calculate distance from the centroids**



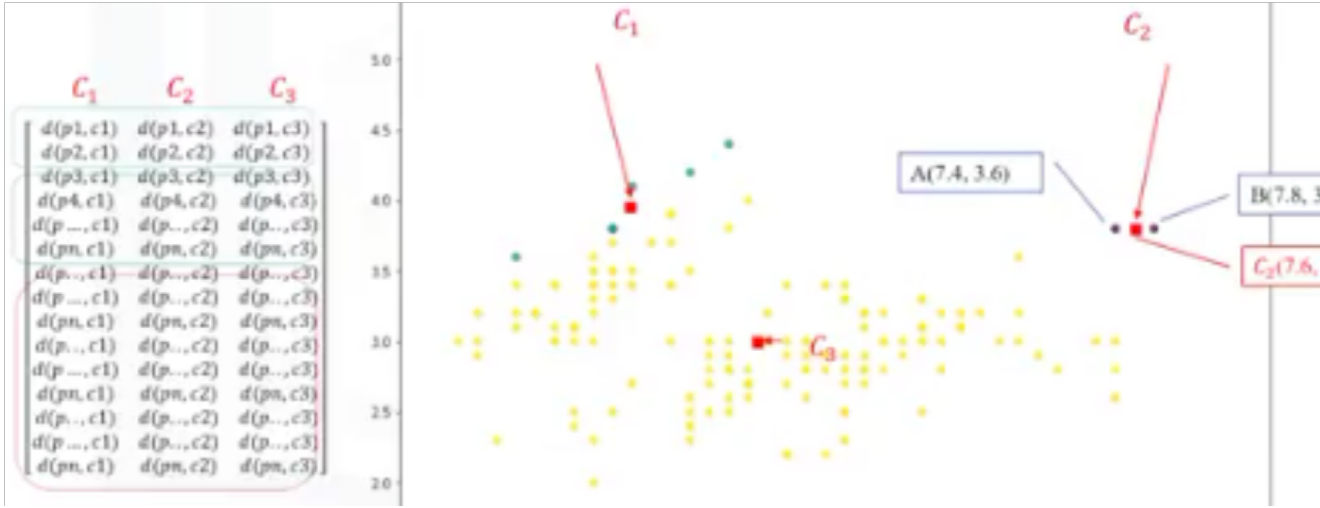
K-Means Clustering Algorithm

1. Define the number of clusters and clusters' centroids
2. Calculate distance from the centroids
3. **Form the clusters by assigning each points to the closest centroid**



K-Means Clustering Algorithm

1. Define the number of clusters and clusters' centroids
2. Calculate distance from the centroids
3. Form the clusters by assigning each points to the closest centroid
4. **Recalculate the new centroids for each cluster**



K-Means Clustering Algorithm

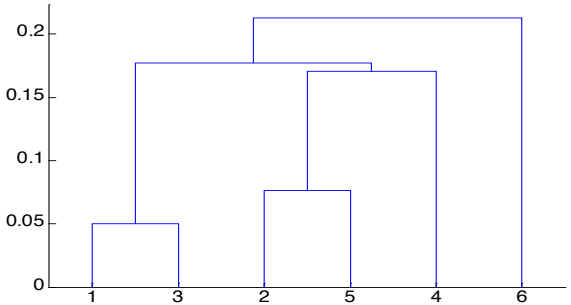
1. Define the number of clusters and clusters' centroids
2. Calculate distance from the centroids
3. Form the clusters by assigning each points to the closest centroid
4. Recalculate the new centroids for each cluster
5. **Iterate to improve clustering** (repeat until there are no more changes)

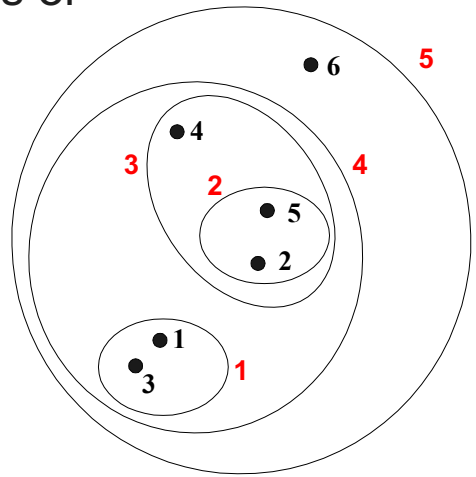
Hierarchical Clustering?

Basic Idea:
Builds a hierarchy of clusters where each node is a cluster consisting of the clusters of its daughter nodes.

- Produces a set of ***nested clusters*** organized as a hierarchical tree

- Can be visualized as a **dendrogram**
  A tree-like diagram that records the sequences of merges or splits





Approaches of Hierarchical Clustering

Two main types of hierarchical clustering

- **Agglomerative:**
  - Start with the points as individual clusters
  - At each step, merge the closest pair of clusters until only one cluster (or **k** clusters) left

- **Divisive:**
  - Start with one, all-inclusive cluster
  - At each step, split a cluster until each cluster contains a point (or there are **k** clusters)

Traditional hierarchical algorithms use a similarity or distance matrix

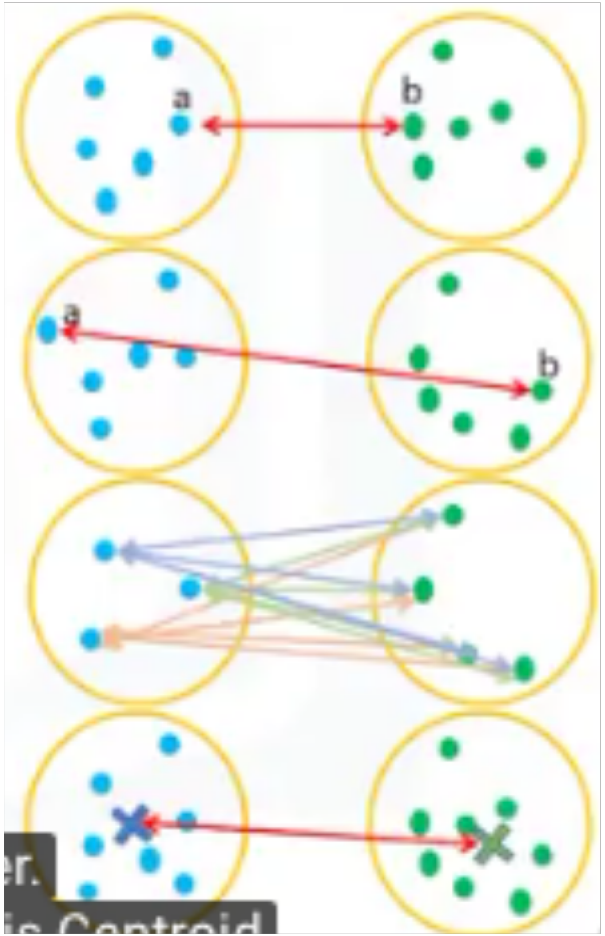
- Merge or split one cluster at a time

Hierarchical Clustering, agglomerative algorithm

1. Create *n* clusters, one for each data point
2. Compute the **Proximity Matrix**
3. **Repeat**
   - » Merge the two closest clusters
   - » Update the proximity matrix
4. **Until** only a single cluster remains

Distance between clusters

- **Single-Linkage Clustering**
  - Minimum distance between clusters
- **Complete-Linkage Clustering**
  - Maximum distance between clusters
- **Average Linkage Clustering**
  - Average distance between clusters
- **Centroid Linkage Clustering**
  - Distance between cluster centroids



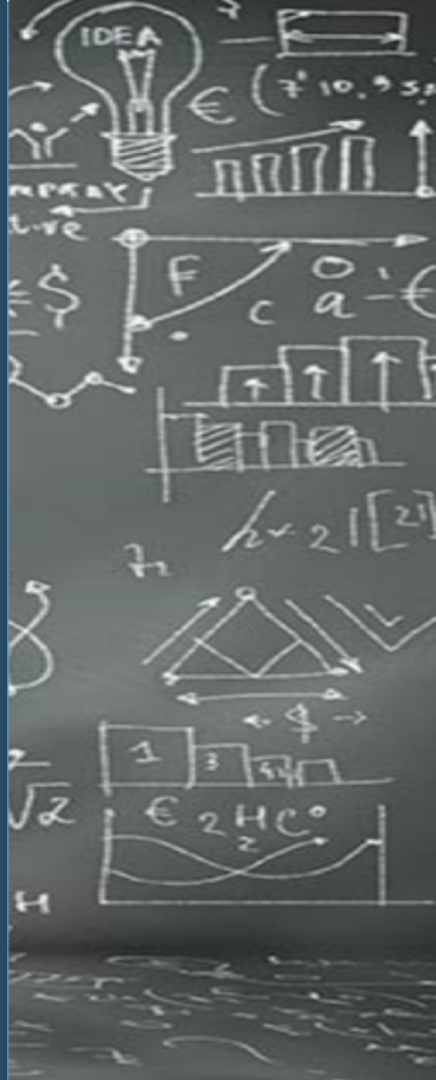
Pros and Cons of Hierarchical Cluster Analysis?

Advantages	Disadvantages
Doesn't require number of clusters to be specified	Can never undo previous steps throughout the algorithm
Easy to implement	Generally has long runtimes
Produces a dendogram, which helps with understanding the data	Sometimes difficult to identify the number of clusters by the dendogram

Hierarchical Clustering vs. K-Means

K-Means	Hierarchical Clustering
Much more efficient	Can be slow for large datasets
Requires the number of clusters to be specified	Does not require the number of clusters to run
Gives only one partitioning of the data based on the predefined number of clusters	Gives more than one partitioning depending on the resolution
Potentially returns different clusters each time it is run due to random initialization of centroids	Always generates the same clusters

Thank You



Useful Links & Resources

External

Getting Started:
Service Homepage
Feature Requests / Suggestions

Case Studies:
OmniEarth
Aerialtronics
BlueChasm
iTrend

Tutorials & Best Practices:
Training models with Watson Studio
Getting started with Watson + Core ML
Stacking Multiple Custom Models
Create a Calorie Counting App
Watson Visual Recognition & Twilio
Best Practices for Custom Models

Code Patterns:
Classify vehicle damage
Analyze industrial equipment for defects
Create an Android calorie-counter app

External continued

Books:
Redbook: Building Cognitive Application using IBM Watson Services vol3 – Watson Visual Recognition

Blogs:
IBM Watson on Medium

Internal

Slack Channel: #ibmvisual-recognition
Service Roadmap
IBMer key limit increase request form
ZACS portal
Digital Sales Play
Content Request & Feedback Form