Artificial Intelligence
Chapter : Clustering

Marouane Ben Haj Ayech

Outline

- Presentation
- Kmeans
- Hierarchical clustering
- Clustering in practice
- Clustering in real world

Presentation

- Prediction

Prediction task	Description	Output Nature	Examples
Clustering	Grouping data points into clusters based on similarity or patterns, often used for unsupervised learning.	Unlabeled classes or clusters	- Customer segmentation based on purchase behavior.

- Learning

Learning Type	Dataset Type	Prediction Tasks	Learning models
Unsupervised	Unlabeled	Clustering	K-Means Gaussian Mixture Models (GMM) Hierarchical Clustering DBSCAN

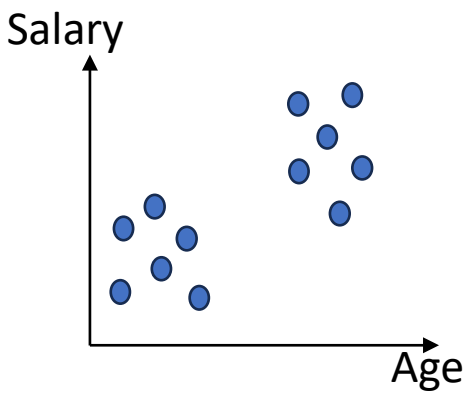
Presentation

Clustering problem

x=employee=(Age, Salary) → y=cluseter id ∈ {0,1}

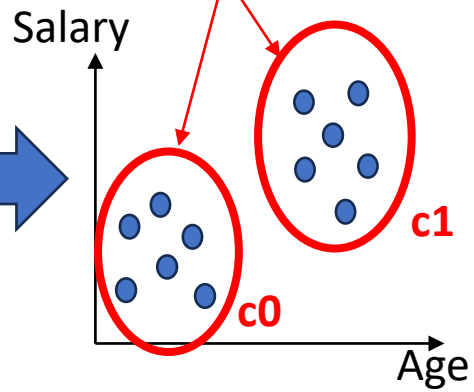
input output

Unlabeled training dataset



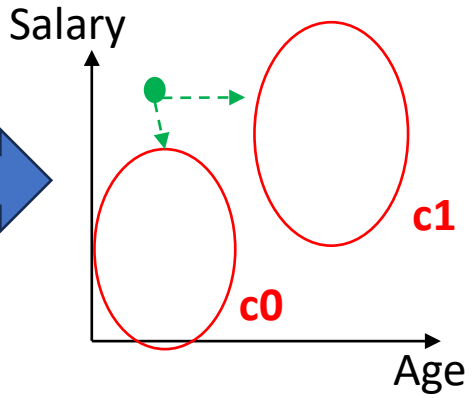
Learning process

Model



Prediction process

Prediction of the nearest cluster id for a new employee



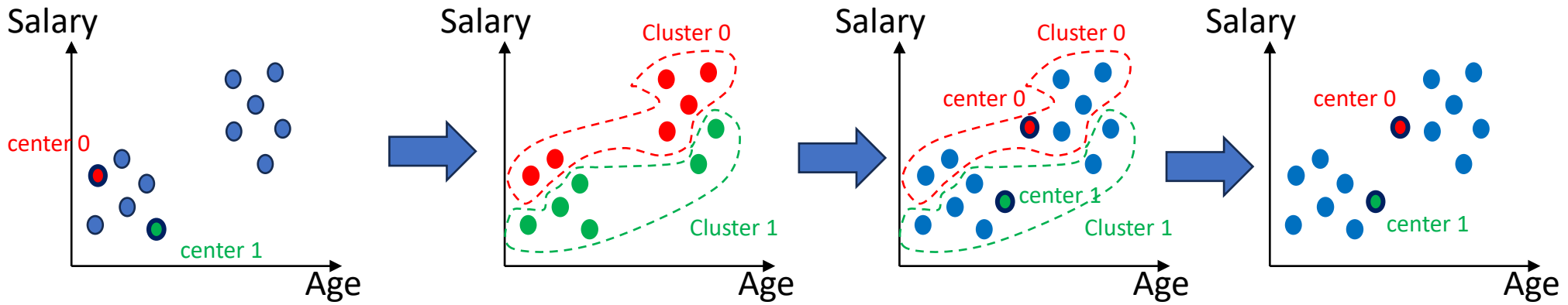
Kmeans

Technique	Learning Process	Prediction Process	Hyperparameters
Kmeans	- Initialize cluster centers - Iteratively until max_iter : (1) Estimate clusters (2) Update cluster centers (3) Check convergence based on a cost function	Find the nearest cluster center for a new data point	- K : number of clusters - max_iter : max number of iterations - epsilon: convergence threshold
	Model		
	- Cluster centers		

Kmeans

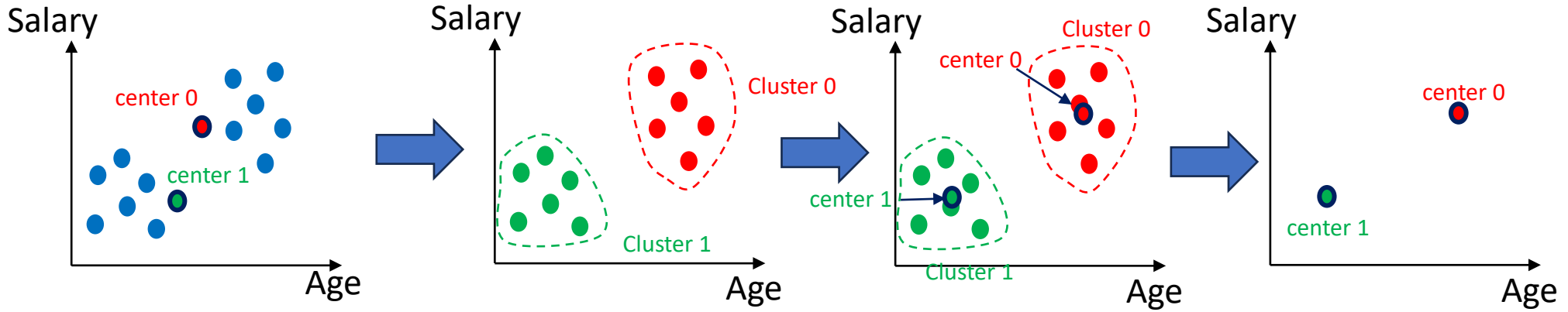
- **Learning process**

Itération 0



Itération 1

Modèle estimé des clusters



Kmeans

Hyperparameters

- K : number of clusters
- max_iter : max number of iterations
- epsilon : convergence threshold

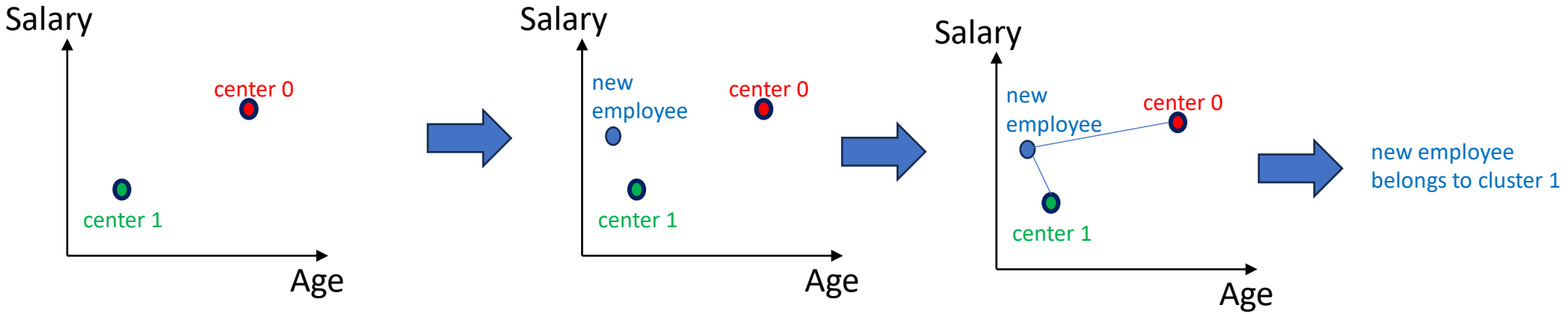
Learning Algorithm

- Randomly initialize K cluster centers as data points
- For t From 0 To max_iter-1:
  - Step 1 : Determine the clusters : find the nearest cluster center for each data point
  - For x in X :
    - cluster_id = argmin distance(x, cluster_center_i)
  - Step 2 : Update cluster centers :a cluster center is the mean of data points that belongs to the cluster
  - For each cluster_center_i :
    - cluster_center_i = mean(cluster_i)
  - Step 3 : Check convergence : compute cost function and check if it doesn't change enough
    - Compute cost_function (total sum of distances between each data point and its nearest cluster center)
    - If |cost_function(t) - cost_function(t-1)|<epsilon Then break

Kmeans

- **Prediction process**

Modèle estimé des clusters



- Let w be a new data point
- Determine which cluster x belongs to : find the nearest cluster center to x

  cluster_id = argmin distance(x, cluster_center_i)

Hierarchical clustering

Hierarchical Clustering Type	**Learning Process**	**Prediction Process**	**Hyperparameters**
Agglomerative	-Initialization : Each data point is a cluster -Iteratively merges clusters based on similarity. -Produces a dendrogram to visualize the hierarchy of clusters	-Compute distances from a new data point to each cluster center -Determine the nearest cluster	- affinity (euclidean, …) - linkage (ward, single, complete, average) - distance_threshold (None or float) - n_clusters (None or int) -compute_full_tree : auto or bool
	Model		
	Dendogram (No parameters)		

Clustering in practice

- Kmeans implementation
  - Define Kmeans as a Python class
  - Define hyperparameters as attributes
  - Define parameters as attributes
  - Define learning process as fit() method
  - Define prediction process as predict() method

Clustering in practice

- Clustering on simple data
  - We apply clustering on simple dataset using a pre-implemented Kmeans in Scikit-learn library.
- Clustering on real data
  - We apply clustering on pseudo real dataset based on the following steps:
    - Data preprocessing
    - Normalization
    - Data reduction
    - Kmeans