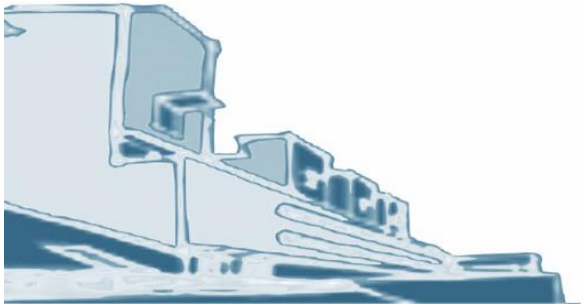
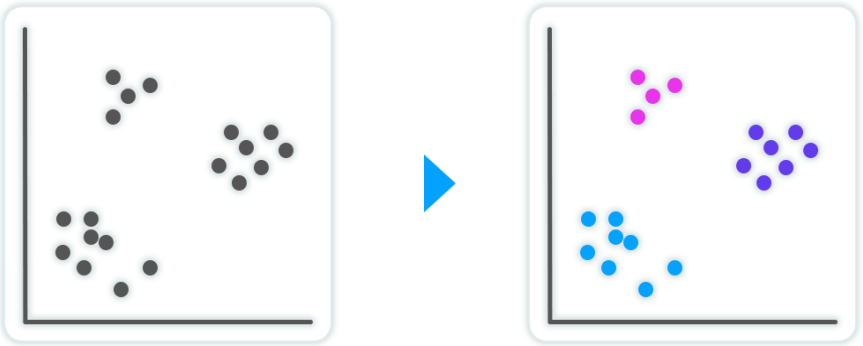
AC – Aprendizagem Computacional / Machine Learning

P2 – Clustering

Jorge Henriques
jh@dei.uc.pt

Departamento de Engenharia Informática
Faculdade de Ciências e Tecnologia





1 2 90
UNIVERSIDADE D COIMBRA

dei engenharia informática

Contents

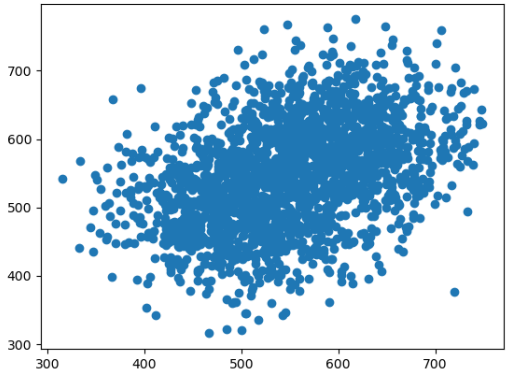
- Clustering
  - Main concepts
  - Different techniques
    - kmeans, hierarchical/Agglomerative, subtractive clustering, DBSCAN
    - Specific parameters for each technique
  - Evaluate each distinct technique / dataset
    - Metrics: silhouette, sum of squared errors, Dunn index

- **Answer:**
  - Which technique is more adequate for each dataset {1,2,3,4,5,6} ?
  - Specific parameters for each solution?

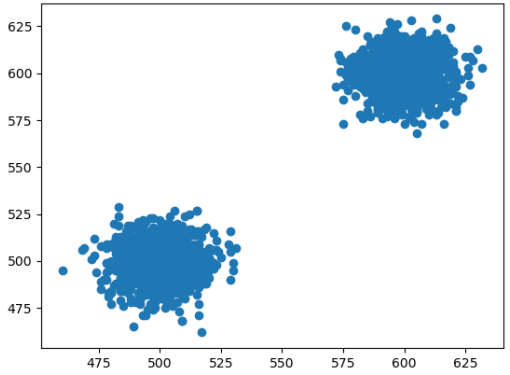
Contents

2.1| Datasets

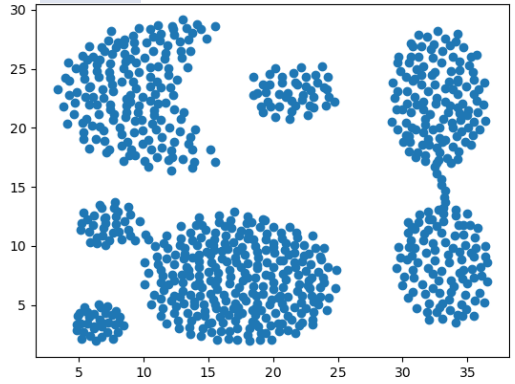
#1



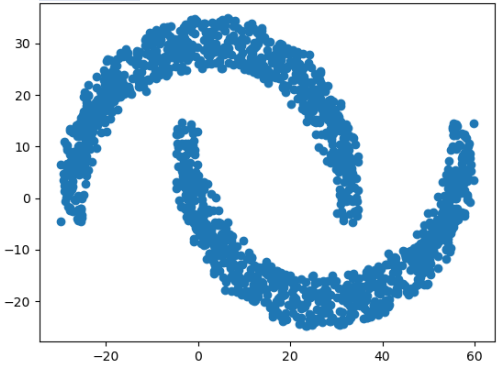
#2



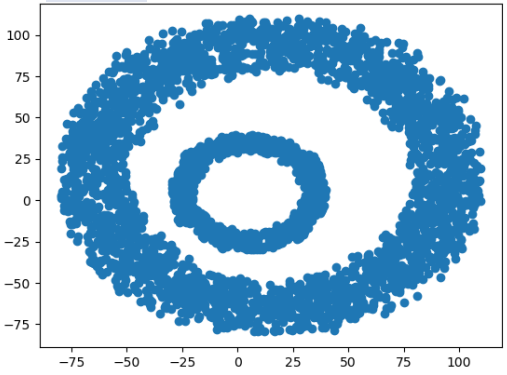
#3



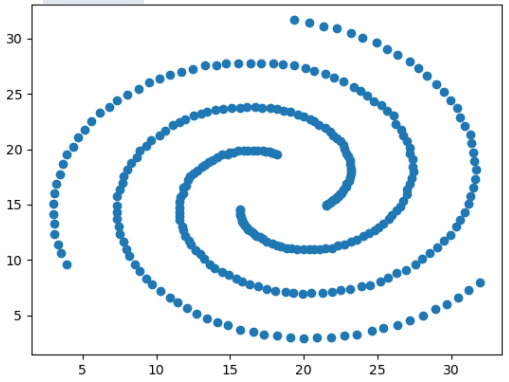
#4



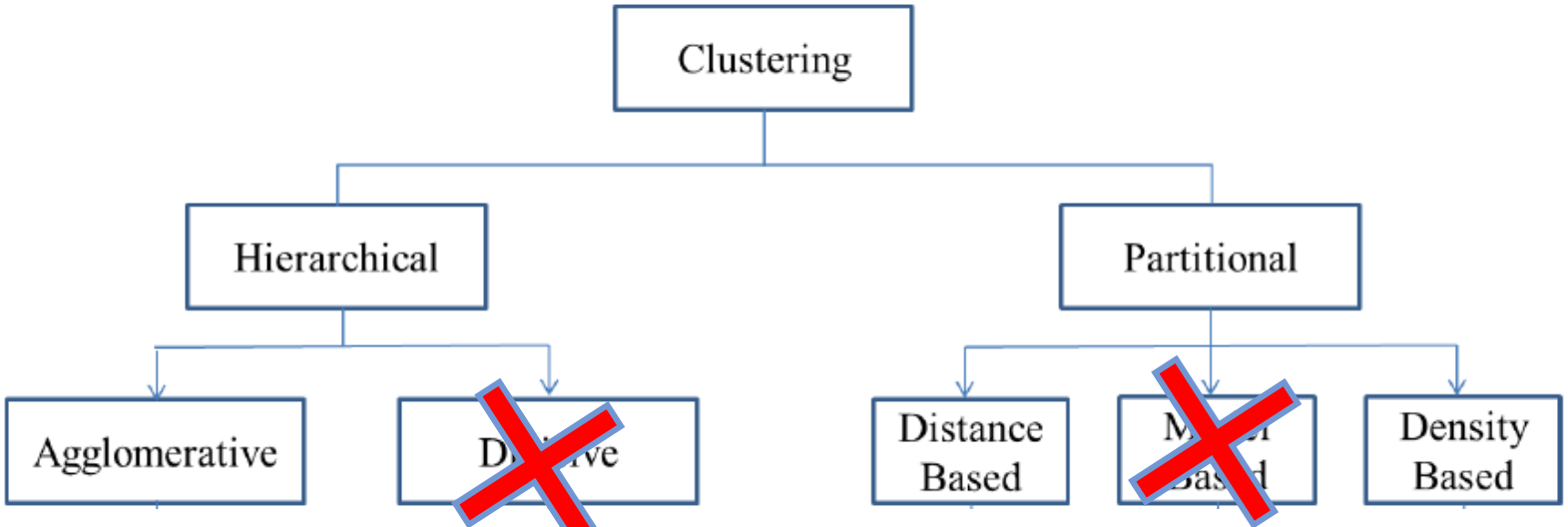
#5



#6



- 2.2| Techniques



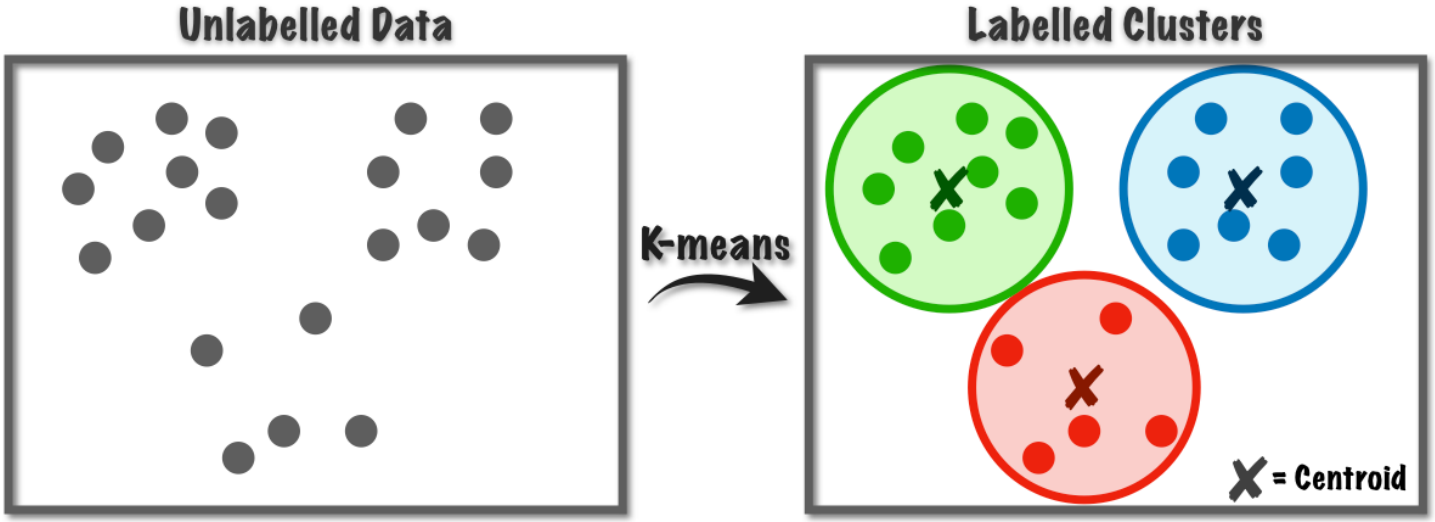
2. Agglomerative

1. Kmeans

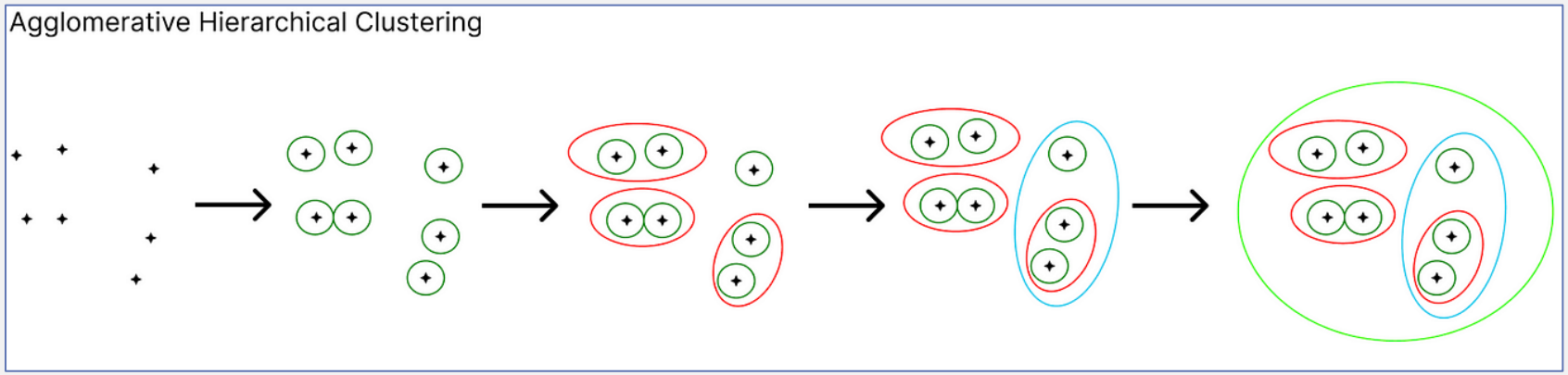
3. Subtractive Clustering

4. DBSCAN

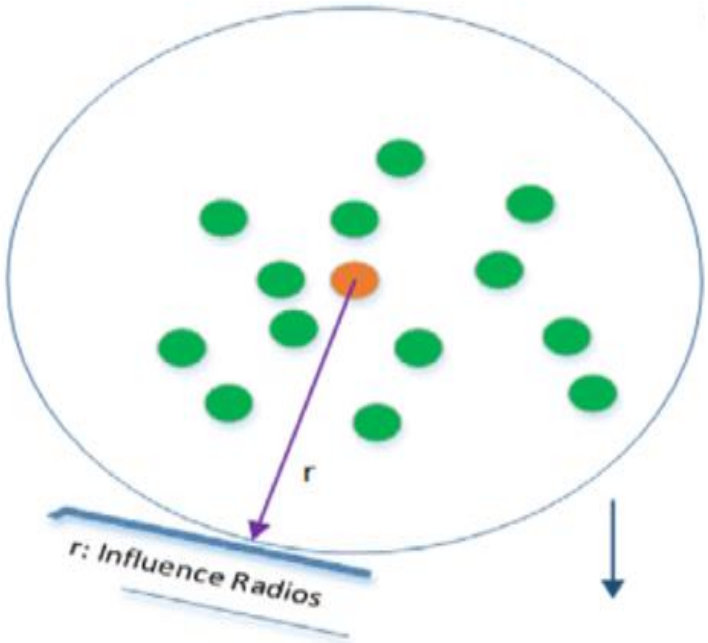
- 2.2| PARTITIONAL (distance) - **KMEANS**
  - Parameters
    - NK | number of clusters
    - Each cluster is defined by the respective **centroid**



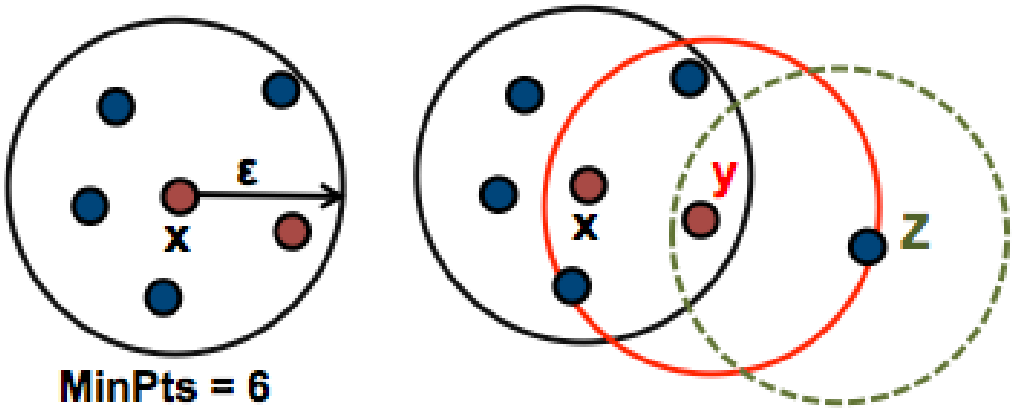
- 2.2| HIERARCHICAL - **agglomerative**
  - Parameters
    - Linkage method | single, complete, average, centroid
    - Final number of clusters **NK**



- 2.2| PARTITIONAL (density) - **Subtractive Clustering**
  - Parameters
    - **ra** |influence radius, controlling the influence of a point
    - **rb** |influence radius, controlling how much the point density is reduced.



- 2.2| PARTITIONAL (density) - **DBSCAN**
  - Parameters
    - **Radius eps** | Define the neighbour
    - **MinPts** | Minimum number of points



Contents

- 1| Code from scratch

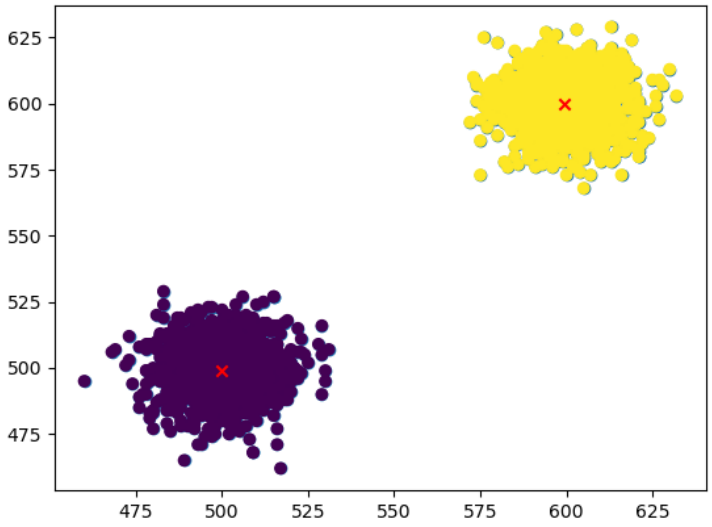
Kmens	Was implemented !
Subtractive Clustering	Has to be implemented !
Agglomerative	Library
DBSCAN	Library

- 2| For each dataset
  - Which is the best technique / parameters ?

    - Implement the techniques using **scikit / matlab** and the specific parameters
    - Visualization of relevant information
    - Assess the performance of each method/dataset

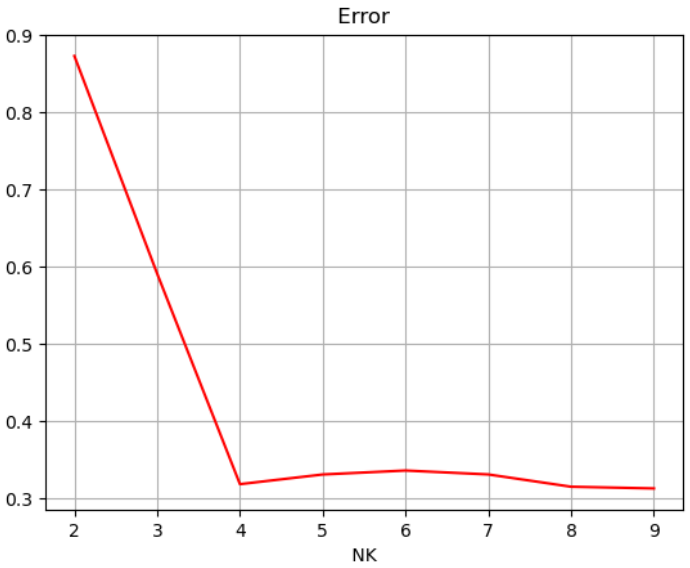
  - Select a method (and parameters) for each dataset

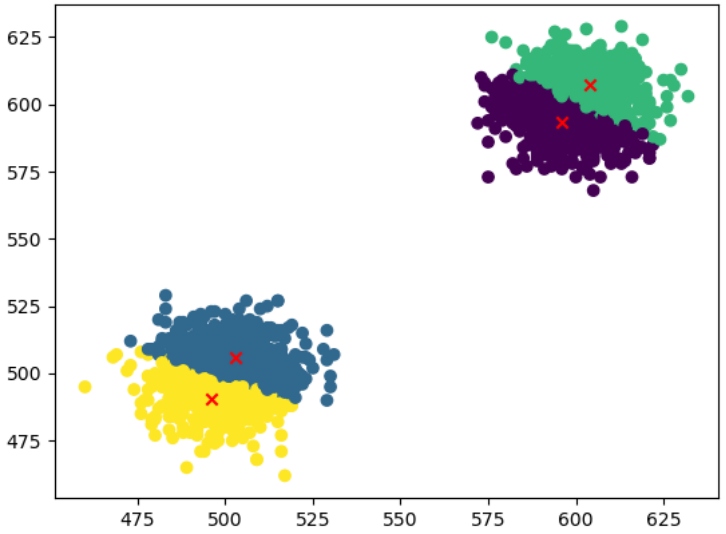
- 2.1| Techniques
  - Kmeans parameters
    - NK=2
    - Dataset=2



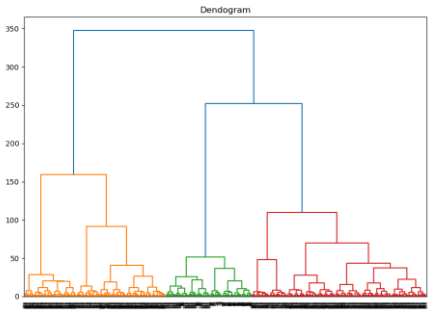

```
from sklearn.cluster import Kmeans

clusterK = KMeans(n_clusters=NK)
clusterK.fit(X)
centerK = clusterK.cluster_centers_
labelsK = clusterK.labels_

#-------------------------------Evaluation
score_average = silhouette_score(X, labelsK)
error         =  clusterK.inertia
```

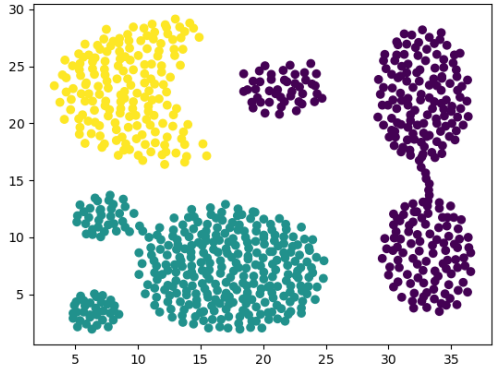
- 2.1| Techniques
  - Kmeans ? How many clusters / centroides?
    - Dataset =2
    - Is **NK = 4** an adequate solution ?





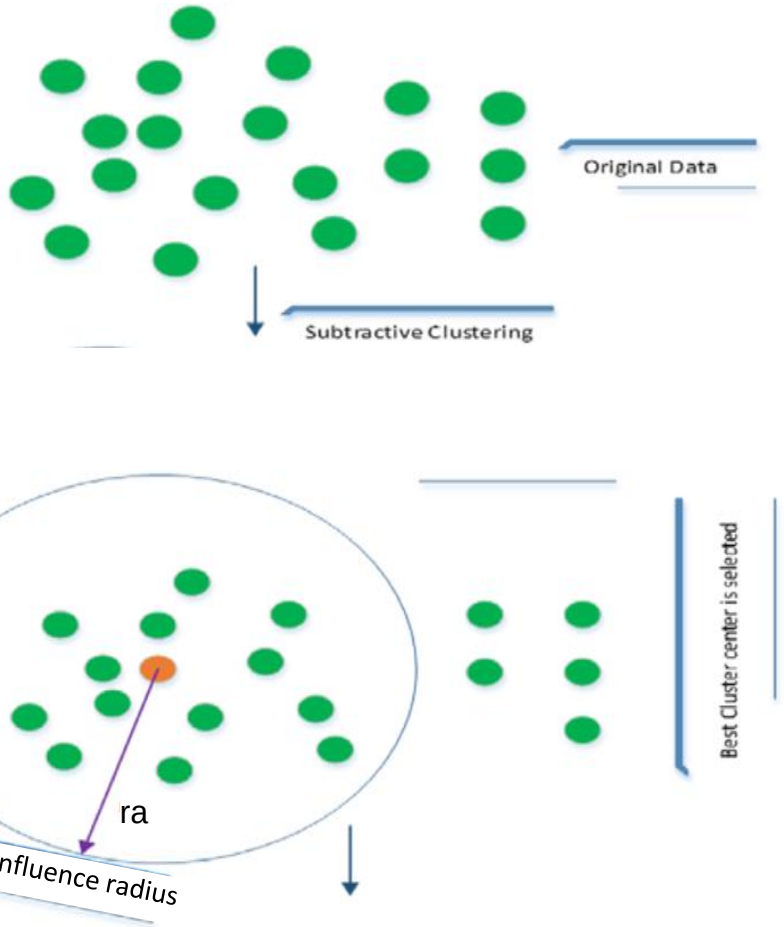
- 2.2| Techniques - **Hierarchical - agglomerative**
  - Dataset=3
  - Linkage method = single
  - Nk=3



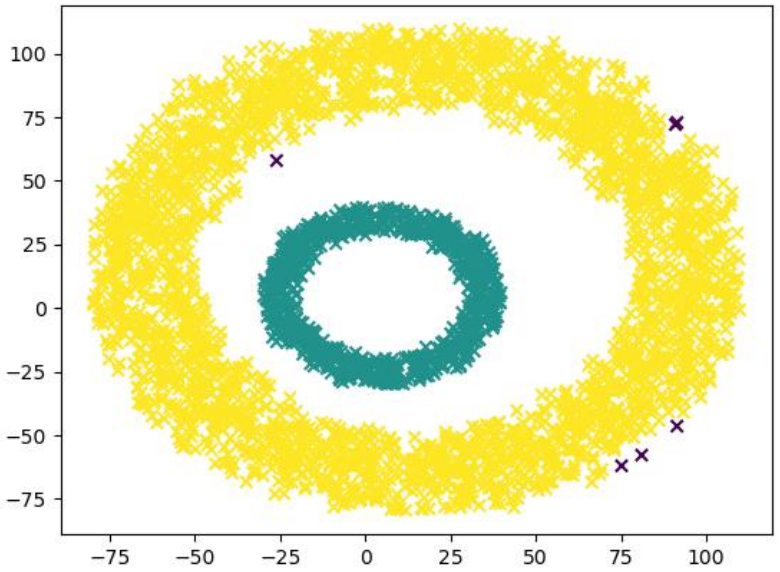



```
from scipy.cluster.hierarchy import linkage
from sklearn.cluster import AgglomerativeClustering
import scipy.cluster.hierarchy as shc

linkage_matrix = linkage(X, method='single', metric='euclidean')
clusterH = AgglomerativeClustering(n_clusters=NK)

clusterH.fit_predict(X)
labelsH = clusterH.labels_
nk      = clusterH.n_clusters_

dend = shc.dendrogram(shc.linkage(X, method='ward'))
```

- 2.3| Techniques – **Subtractive Clustering**
  - ra
  - rb



- 2.4| Techniques – **DBSAN**
  - Dataset=5
  - eps=4
  - minPts=5




```
from sklearn.cluster import DBSCAN

clusterD  = DBSCAN(eps=4, min_samples=5)
clusterD.fit_predict(X)
labelsD = clusterD.labels_
```

Contents

- 4| Conclusions

  - **1| Clustering**
    - Implement: Subtractive (density-based method)
    - Scikit-Learning kmeans, agglomerative, DBSCAN
    - Study: Clustering techniques and respective evaluation

  - **2| Select a clustering method for each dataset !**

  - **3| Other improvements ?**
    - How to select the number of clusters?
    - Any other clustering method ..
    - ...