UNIVERSIDAD
NACIONAL
DE COLOMBIA

Clasificación y Reconocimiento de Patrones

John W. Branch
Carlos Madrigal

Departamento de Ciencias de la Computación y de la Decisión

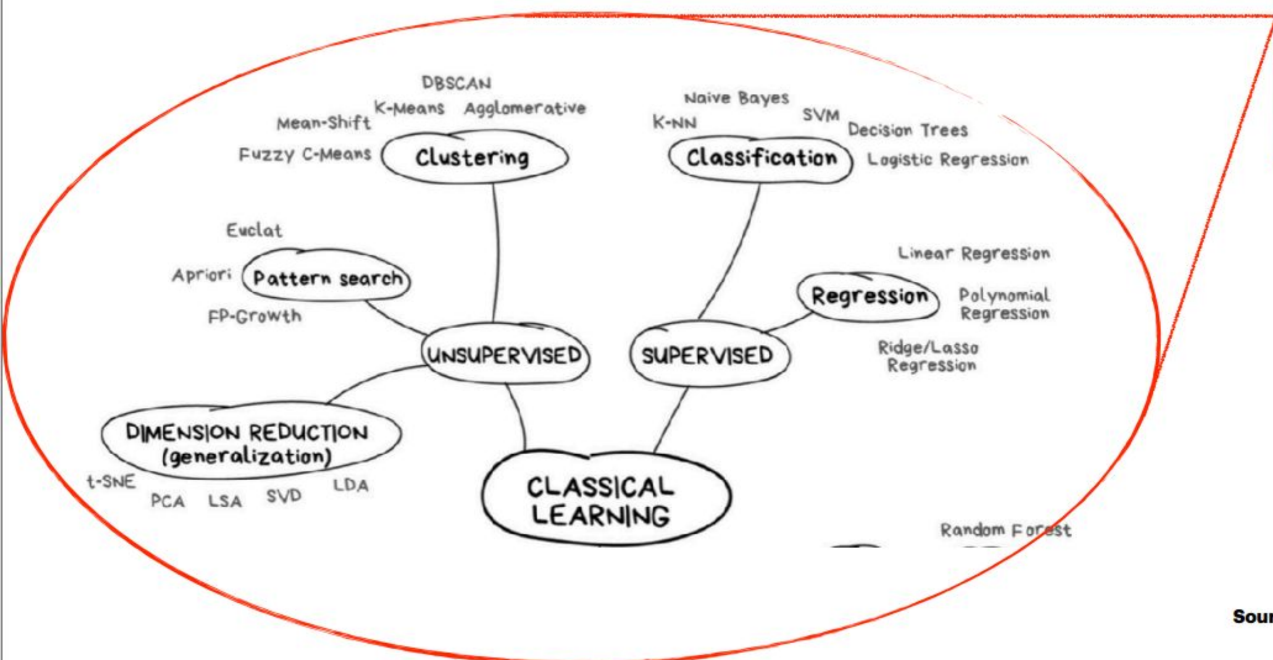
Nota: Este material se ha adaptado con base a diferentes fuentes de información académica

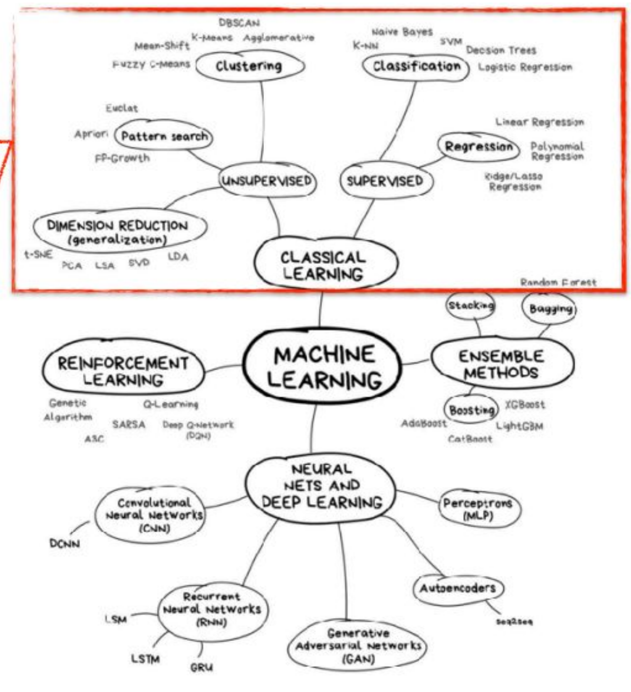
AGENDA

Sesion 1: Image Segmentation

1. ML Landscape.
   - Data Labelling Service
2. Unsupervised Learning Fundamentals
   - Definition
   - Taxonomy
   - Applications
3. Clustering
   - Partitioning-based clustering
     - K-Means Clustering.
     - Determining the Optimal K for K-Means
   - Density-based clustering
     - Mean-Shift.
     - DBScan
4. Python practice: Image Segmentation.
5. Conclusions.

ML Landscape





Source: https://medium.com/better-programming/from-machine-learning-to-reinforcement-learning-mastery-47f33d9f6b41

DATA LABELLING SERVICE

Humans in the Loop

Human workforce for continuously better computer vision

Google

AI Platform

LIONBRIDGE

Data Labeling Services

Power machine learning models with Lionbridge's end-to-end data labeling platform.

LABEL YOUR DATA

SECURE DATA ANNOTATION FOR AI

Desirable Features of Clustering
Kleinberg's axioms

1. **Scale Invariance:**
   This simple axiom indicates that a clustering algorithm should not modify its results when all distances between points are scaled by the factor determined by a constant α.

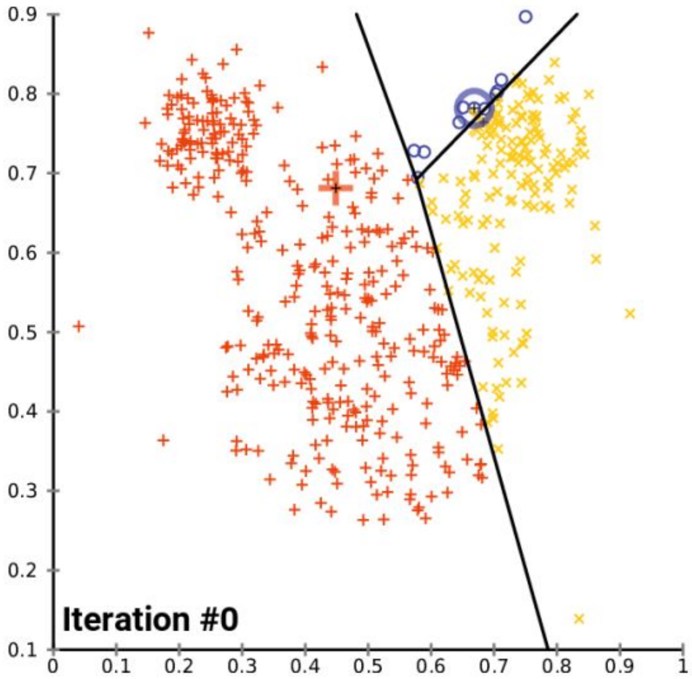
2. **Richness:**
   This means that the the clustering function must be flexible enough to produce any arbitrary partition/clustering of the input data set.

3. **Consistency:**
   A clustering process is "consistent" when the clustering results do not change if the distances within clusters decrease and/or the distances between clusters increase.

(Palacio and Berzal, 2019). Evaluation Metrics for Unsupervised Learning Algorithms.

K-Means Clustering

1. Input Data $X = x_1, x_2, \ldots x_N$ and number of clusters K
2. Centroids $c_1, c_2, \ldots c_K$ = random K points of X
3. foreach data point x_i
   - Compute distance $d_{ij} = d(x_i, c_j)$
     $i = \{1,\ldots,N\}, j = \{1,\ldots,K\}$
   - Assign x_i to the nearest centroid: $y_i = argmin_j(d_{ij})$
4. Compute the new centroids of each cluster
   $c_j^* = mean(x_i)$ for $y_i = j$
5. if $c_j^* \neq c_j$ then $c_j = c_j^*$ goto step 3
6. Output: $c_1^*, c_2^*, \ldots c_K^*$ and y_i for $i = \{1,\ldots,N\}$

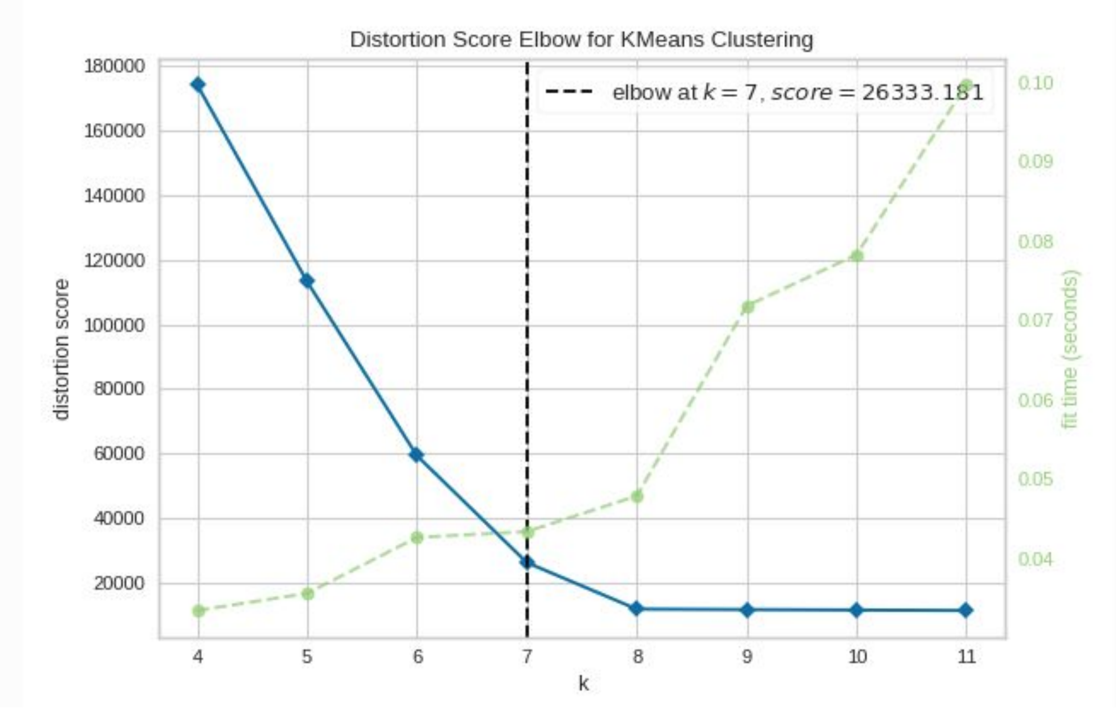


Source: https://en.wikipedia.org/wiki/K-means_clustering

Visualizing K-Means Clustering

Determining the Optimal K for K-Means

The Elbow Method

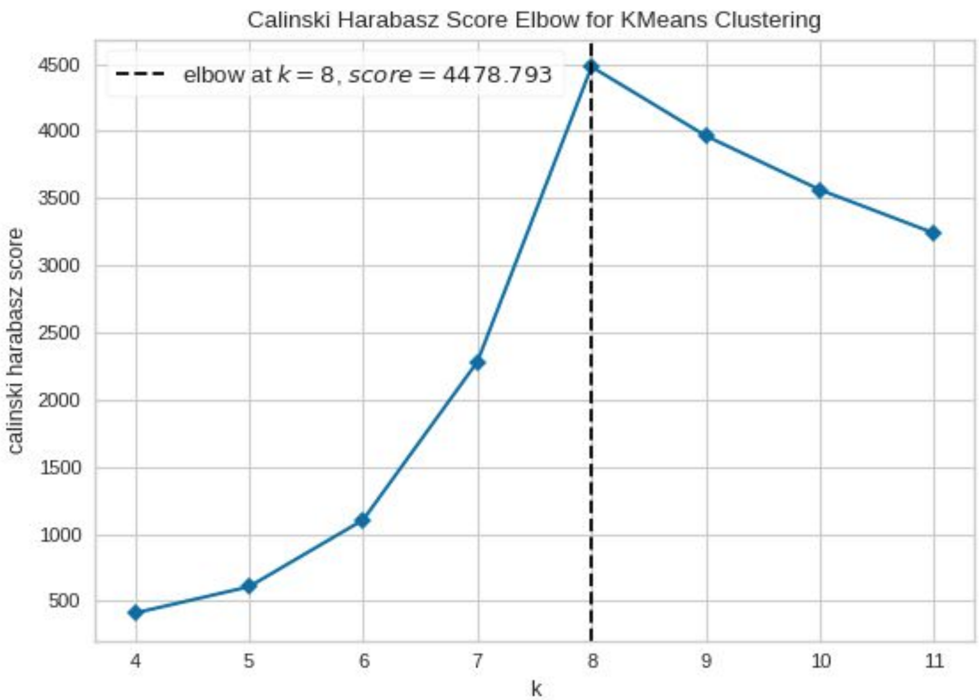


computes the sum of squared distances from each point to its assigned center for different values of k, and choose the k for which SSE becomes first starts to diminish.

https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

Determining the Optimal K for K-Means

Calinski_harabas Score

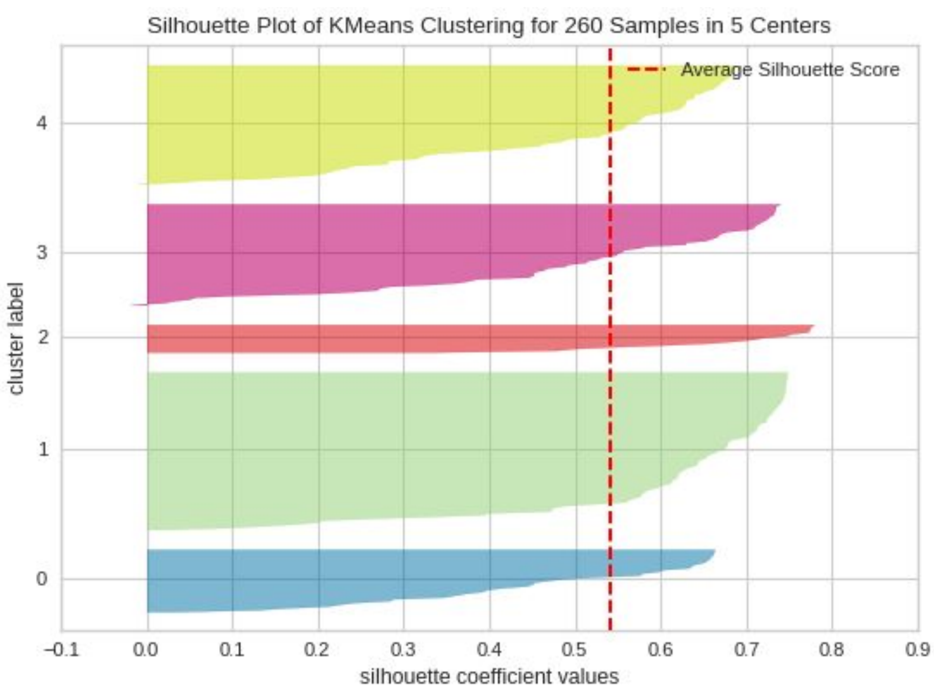


Computes the ratio of dispersion between and within clusters

https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

Determining the Optimal K for K-Means

The Silhouette Method

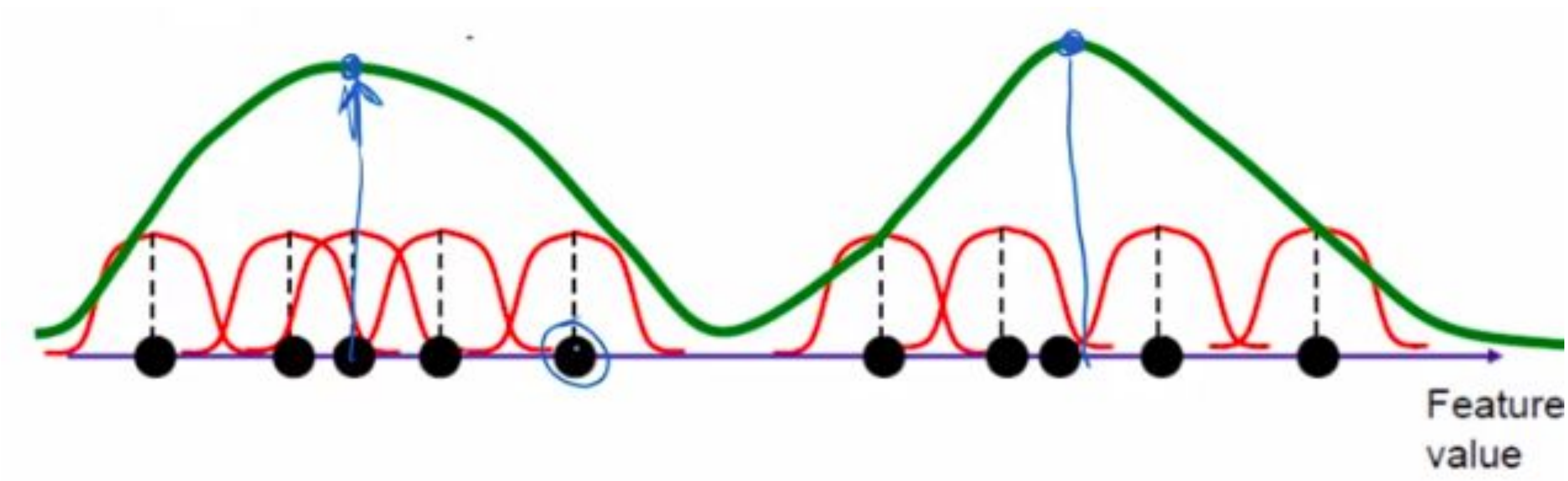


https://www.scikit-yb.org/en/latest/api/cluster/elbow.html

The silhouette value measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation). The range of the Silhouette value is between +1 and -1. A high value is desirable and indicates that the point is placed in the correct cluster. If many points have a negative Silhouette value, it may indicate that we have created too many or too few clusters.

Mean Shift

1. Convert the set of points into a continuous function using a gaussian kernel (**pdf**)
2. The mode in the probability density function correspond to the cluster in the data



https://elvex.ugr.es/idbis/dm/slides/43%20Clustering%20-%20Density.pdf

http://user.ceng.metu.edu.tr/~akifakkus/courses/ceng574/mean-shift/

DBSCAN: Density Based Spatial Clustering of Applications with Noise

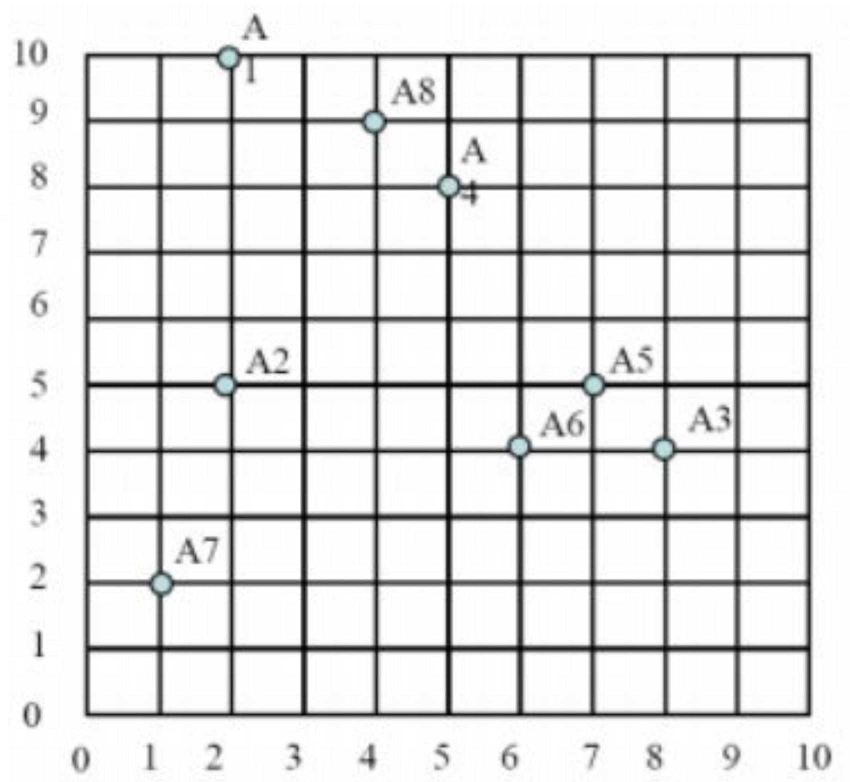
Agrupar los 8 puntos de la figura utilizando el algoritmo DBSCAN.

Número mínimo de puntos en el "vecindario":

MinPts = 2

Radio del "vecindario":

Epsilon $\sqrt{2}$ › $\sqrt{10}$



https://elvex.ugr.es/idbis/dm/slides/43%20Clustering%20-%20Density.pdf

DBSCAN: Density Based Spatial Clustering of Applications with Noise

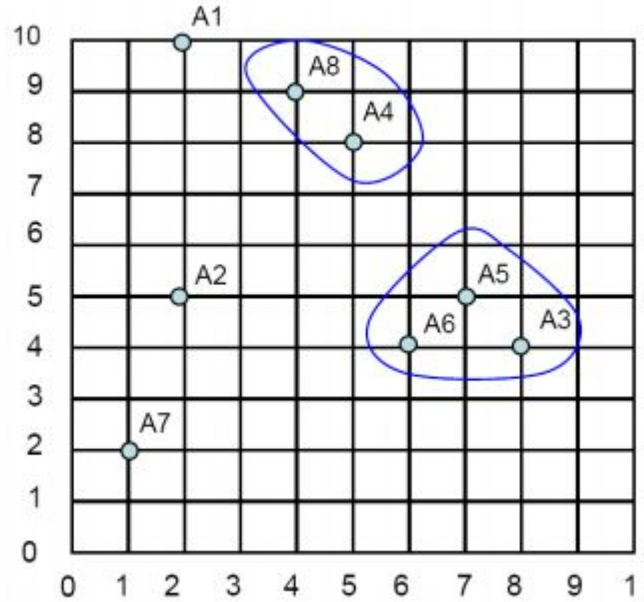
Distancia euclídea

	A1	A2	A3	A4	A5	A6	A7	A8
A1	0	$\sqrt{25}$	$\sqrt{36}$	$\sqrt{13}$	$\sqrt{50}$	$\sqrt{52}$	$\sqrt{65}$	$\sqrt{5}$
A2		0	$\sqrt{37}$	$\sqrt{18}$	$\sqrt{25}$	$\sqrt{17}$	$\sqrt{10}$	$\sqrt{20}$
A3			0	$\sqrt{25}$	$\sqrt{2}$	$\sqrt{2}$	$\sqrt{53}$	$\sqrt{41}$
A4				0	$\sqrt{13}$	$\sqrt{17}$	$\sqrt{52}$	$\sqrt{2}$
A5					0	$\sqrt{2}$	$\sqrt{45}$	$\sqrt{25}$
A6						0	$\sqrt{29}$	$\sqrt{29}$
A7							0	$\sqrt{58}$
A8								0

https://elvex.ugr.es/idbis/dm/slides/43%20Clustering%20-%20Density.pdf

DBSCAN: Density Based Spatial Clustering of Applications with Noise

A1, A2 y A7 no tienen vecinos en su vecindario,
por lo que se consideran "outliers" (no están en zonas densas):

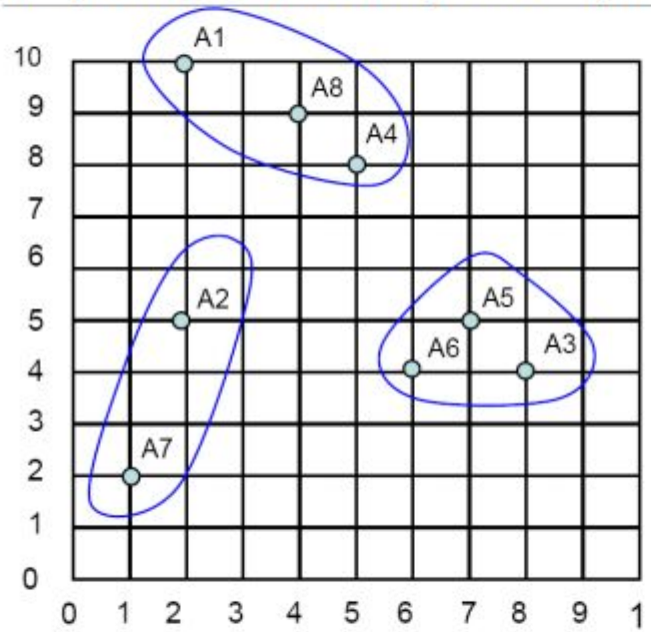


https://elvex.ugr.es/idbis/dm/slides/43%20Clustering%20-%20Density.pdf

DBSCAN: Density Based Spatial Clustering of Applications with Noise

Al aumentar el valor del parámetro Epsilon,
el vecindario de los puntos aumenta y todos quedan agrupados:

Epsilon $= \sqrt{10}$



Epsilon $= \sqrt{10}$

https://elvex.ugr.es/idbis/dm/slides/43%20Clustering%20-%20Density.pdf

Ejemplos

Session 1 - Determining the Optimal K for K-Means

Clustering-SKLearn.ipynb

Image Segmentation.ipynb

Session 2 -Unsupervised Image Classification

PREGUNTAS

UNIVERSIDAD
NACIONAL
DE COLOMBIA