

Clustering

Cristian López Del Alamo
clopezd@utec.edu.pe
IPRODAM3D - Research group

Mayo, 2021



Programa



1. Kmenas
2. Mean Shift
3. DBSCAN

1

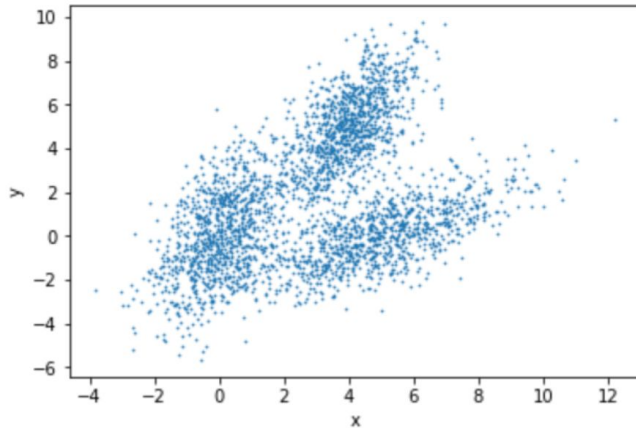
Clustering



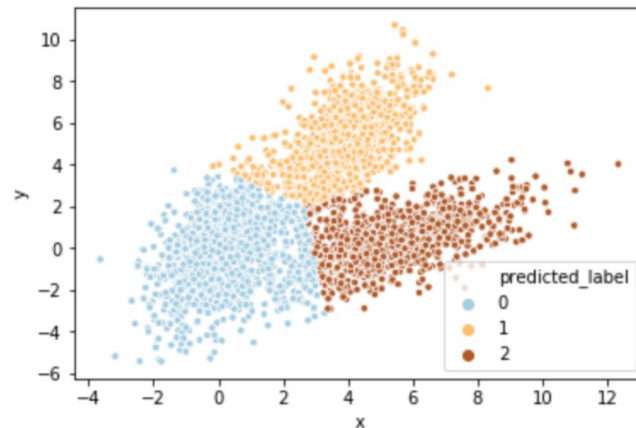
Clustering

- Técnica de **machine learning** que permite agrupar, de manera **no supervisada** un conjunto de datos de acuerdo a su estructura o **características similares**.

Before



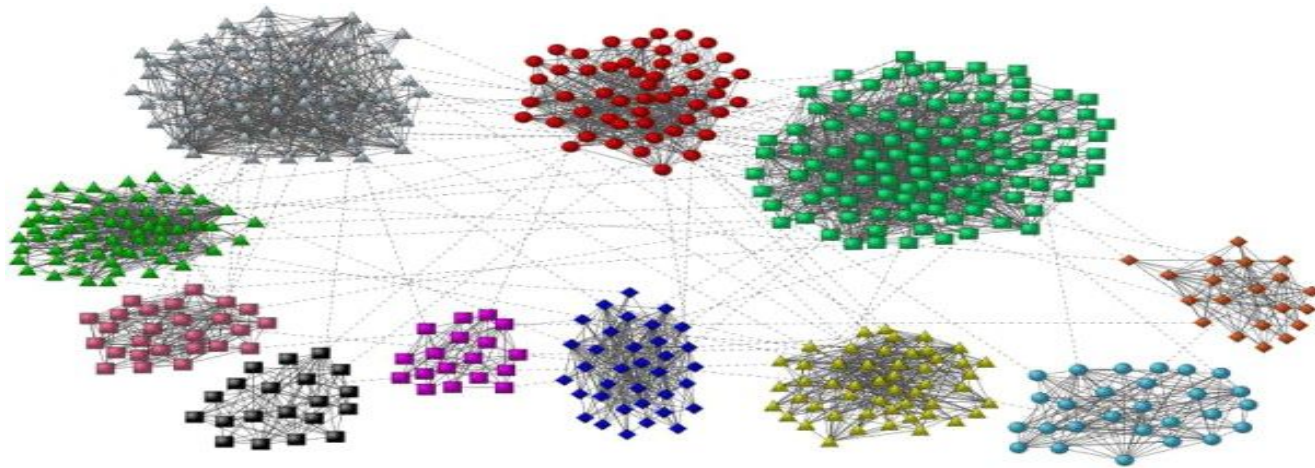
After



[Fuente: Click](#)

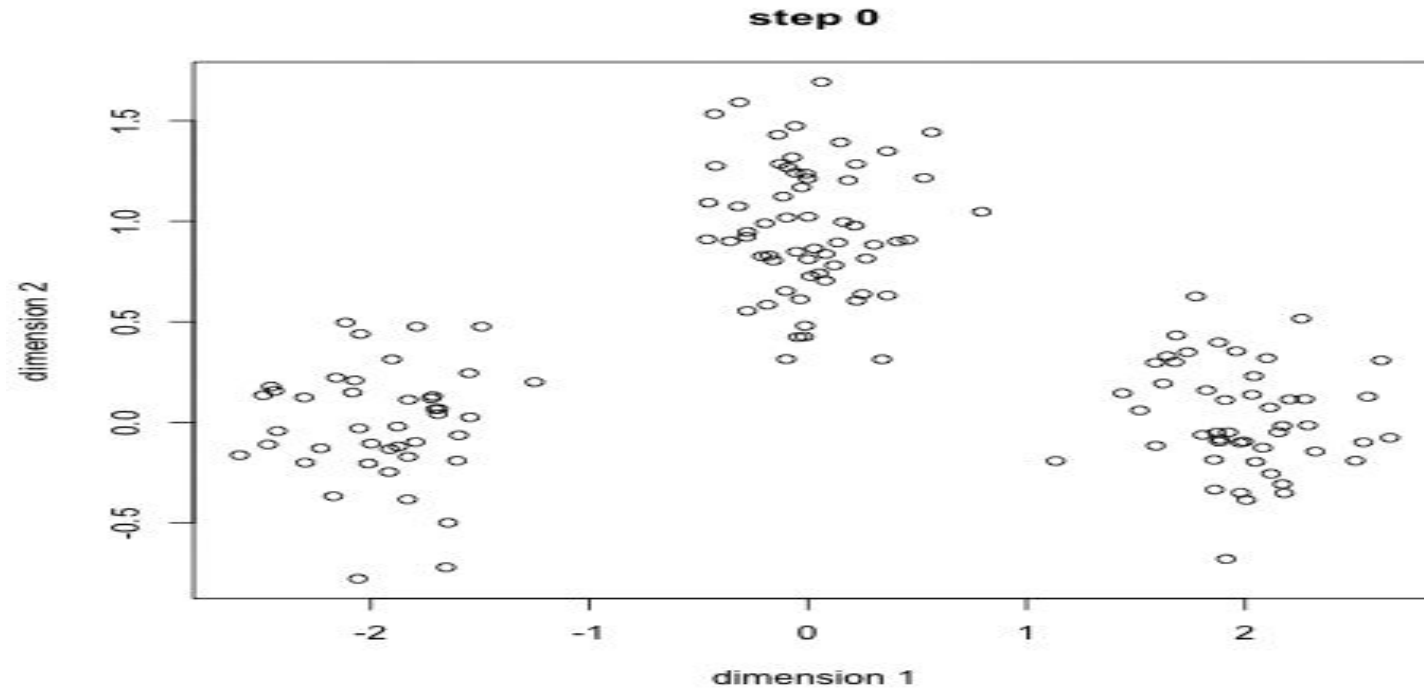
Clustering

- El conjunto de datos que pertenecen a **un mismo grupo** deben tener características **propiedades similares**, y a la vez, características **muy disímiles** respecto a elementos de **otros grupos**



[Fuente: Click](#)

K-Means



[Fuente: Click](#)

Algorithm: K-Means

Input —Dataset
 —number of clusters

Output —K clusters

Step-1: —Initialize K centers of the cluster

Step-2: —Repeat

 —Calculate the mean of all the objects belonging to that cluster

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q$$

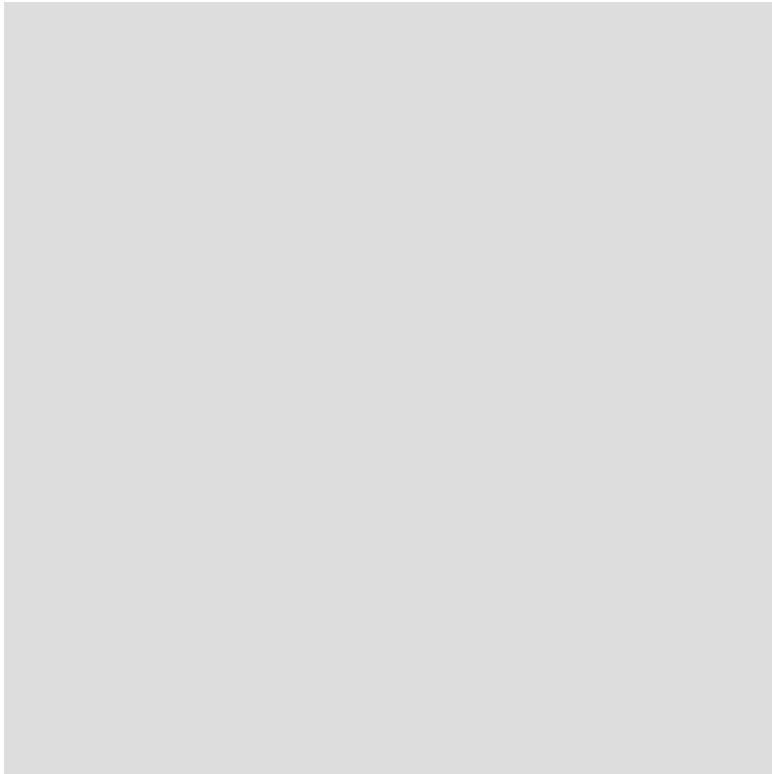
 where μ_k is the mean of cluster k and N_k is the number of p
 belonging to that cluster

 —Assign objects to the closest cluster centroid

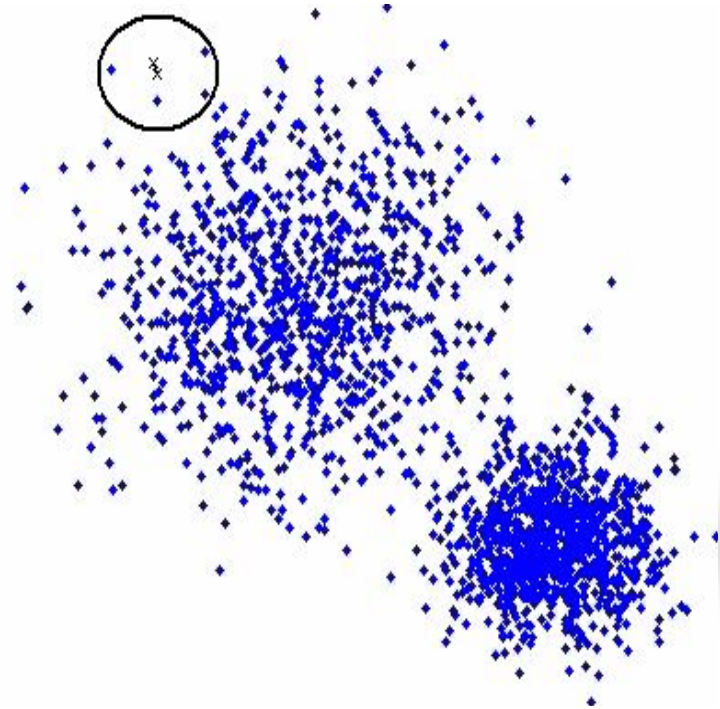
 —Update cluster centroids based on the assignment

 —**Until** centroids do not change

Mean-Shift

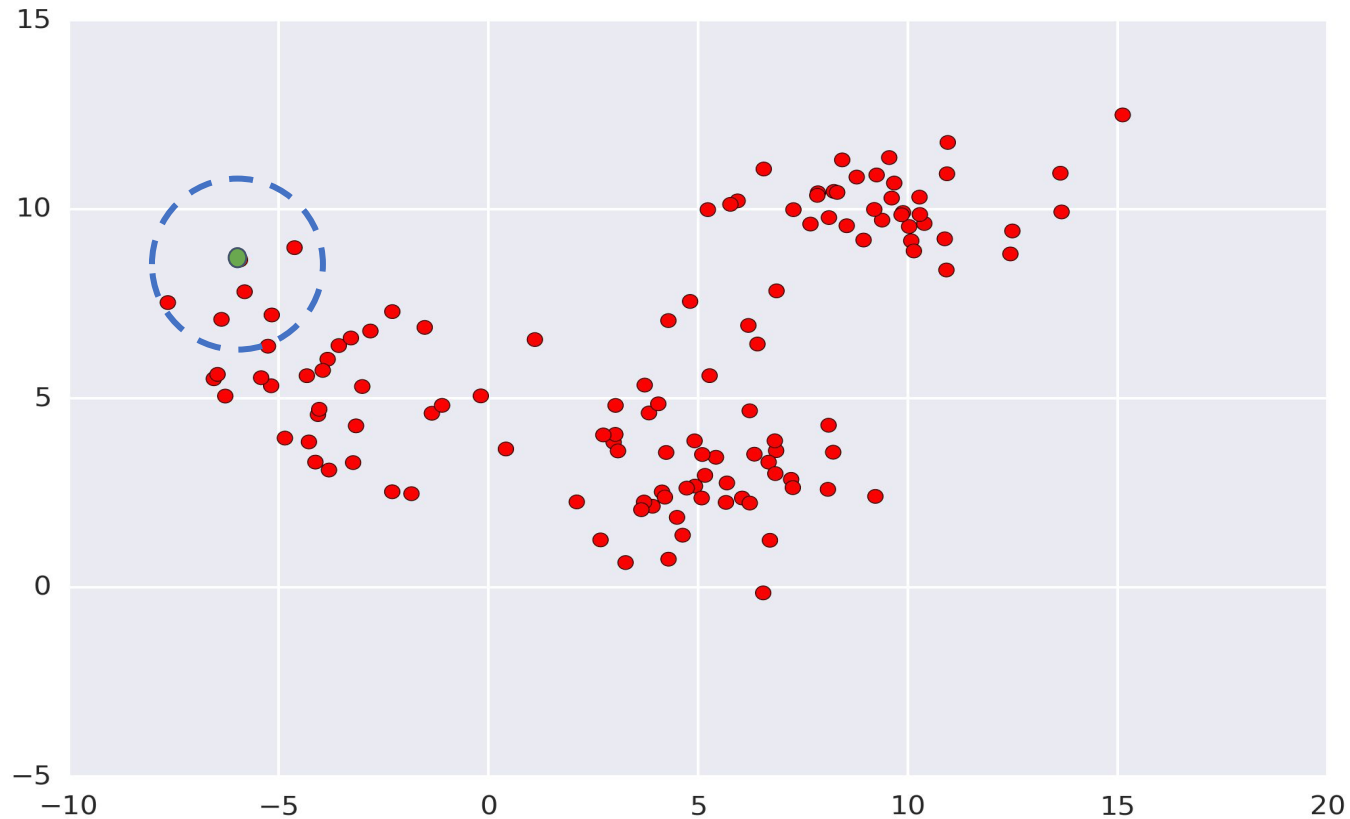


[Fuente: Click](#)



[Fuente: Click](#)

Mean-Shift



Mean-Shift

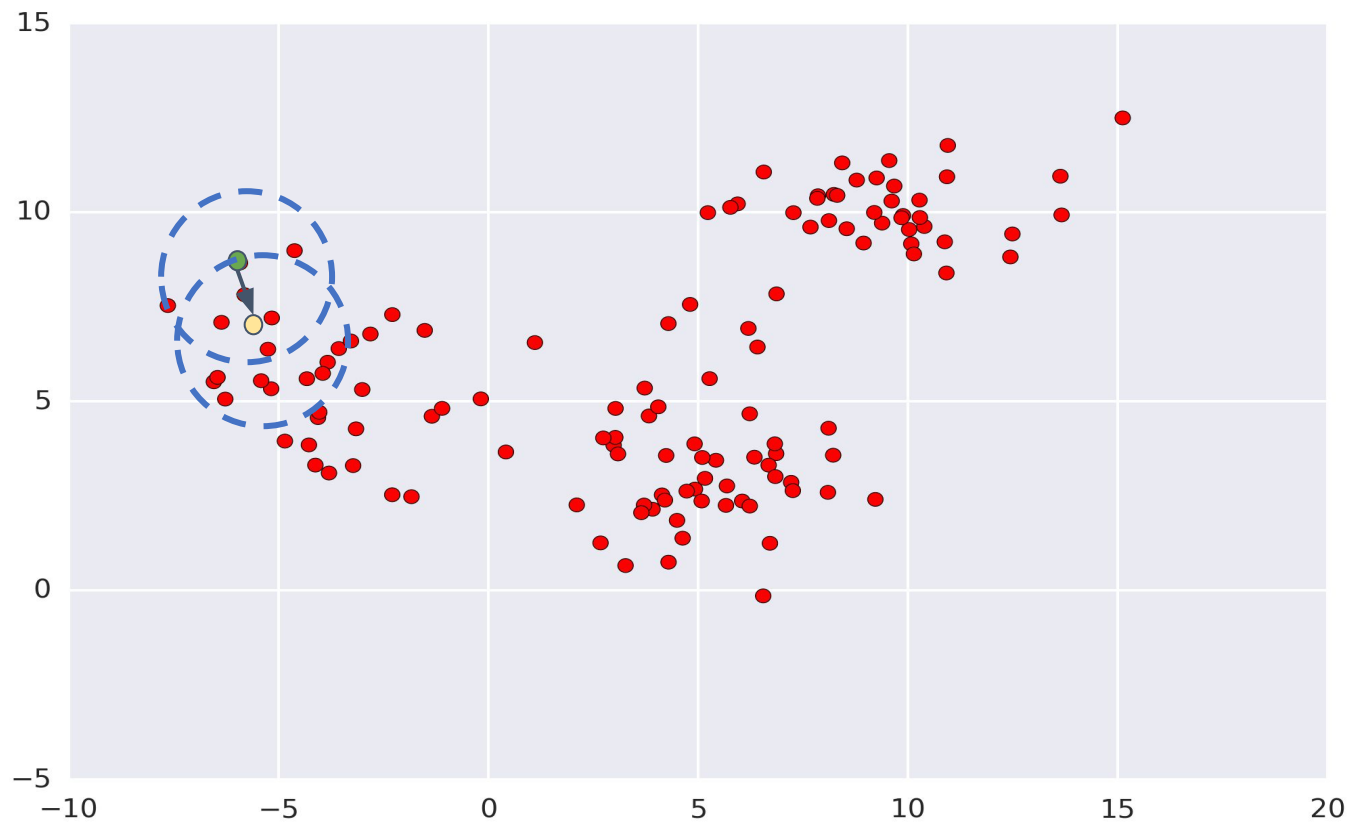
- Smoothing kernel

$$f(x) = \sum_m K(x - x_m) = \sum_m k\left(\frac{x - x_m}{h}\right)$$

- Gaussian Kernel

$$k(x) = (2\pi)^{-d/2} \exp\left(-\frac{1}{2} ||x||^2\right)$$

Mean-Shift

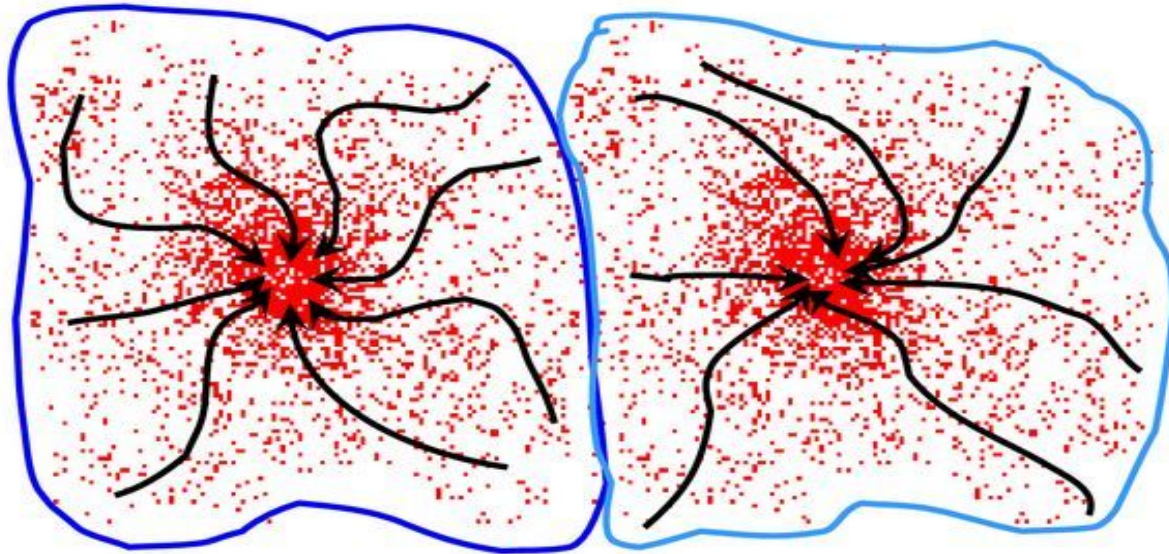


Mean-Shift

- Calculamos el mean shift vector
- Movemos el kernel windows

$$\hat{x} = \frac{\sum_m x_m \exp\left(-\frac{1}{2} \left\| \frac{x - x_m}{h} \right\|^2\right)}{\sum_m \exp\left(-\frac{1}{2} \left\| \frac{x - x_m}{h} \right\|^2\right)} - x$$

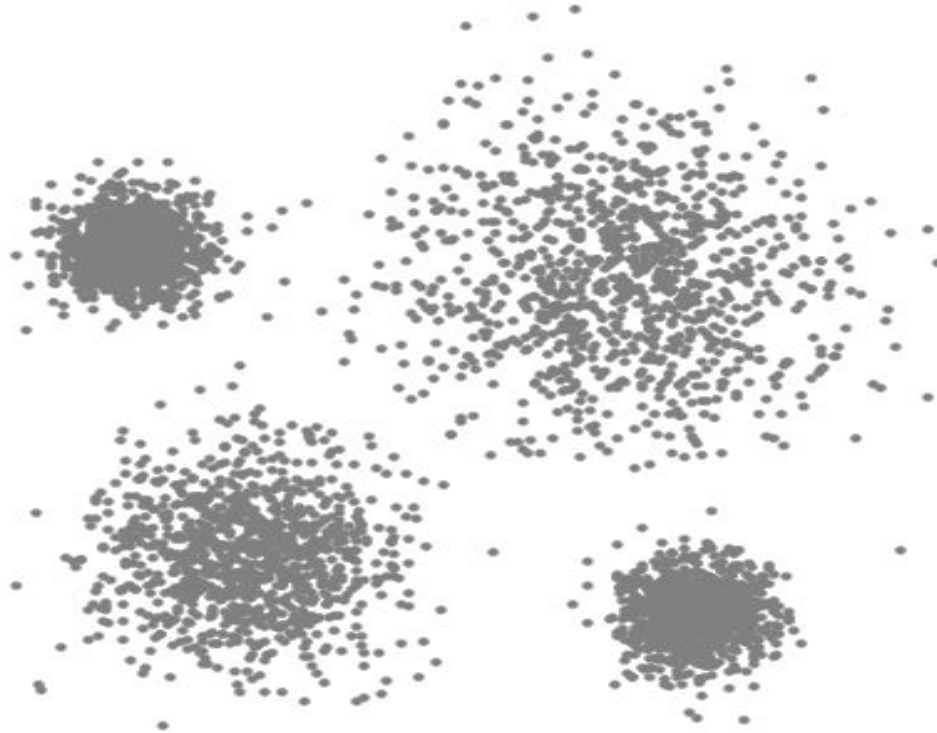
Mean-Shift



Mean-Shift

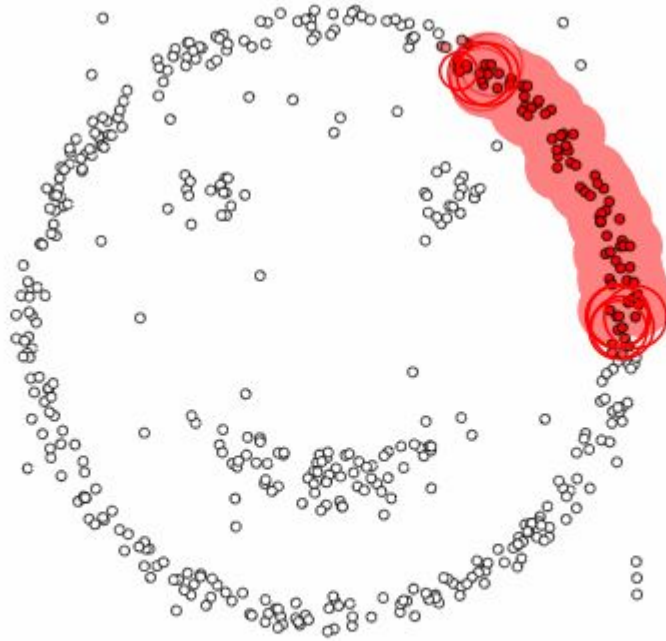
1. Elegir un kernel y un radio
2. Por cada punto p :
 - a. Centrar una ventana en el punto p
 - b. Computar la media de los datos en la ventana de búsqueda
 - c. Centrar la ventana en la nueva localización
 - d. Repetir (b,c) hasta que converga
3. Asignar el cluster el centroide a todos los nodos que le dieron origen.

Mean-Shift



[Fuente: Click](#)

Density-Based Spatial Clustering of Applications with Noise (DBSCAN)



epsilon = 1.00
minPoints = 4

[Fuente: Click](#)

ALGORITHM 1: Pseudocode of Original Sequential DBSCAN Algorithm

Input: *DB*: Database
Input: ϵ : Radius
Input: *minPts*: Density threshold
Input: *dist*: Distance function
Data: *label*: Point labels, initially *undefined*

```

1 foreach point p in database DB do                                // Iterate over every point
2   if label(p)  $\neq$  undefined then continue                        // Skip processed points
3   Neighbors N  $\leftarrow$  RANGEQUERY(DB, dist, p,  $\epsilon$ )           // Find initial neighbors
4   if |N| < minPts then                                           // Non-core points are noise
5     label(p)  $\leftarrow$  Noise
6     continue
7   c  $\leftarrow$  next cluster label                                    // Start a new cluster
8   label(p)  $\leftarrow$  c
9   Seed set S  $\leftarrow$  N \ {p}                                    // Expand neighborhood
10  foreach q in S do
11    if label(q) = Noise then label(q)  $\leftarrow$  c
12    if label(q)  $\neq$  undefined then continue
13    Neighbors N  $\leftarrow$  RANGEQUERY(DB, dist, q,  $\epsilon$ )
14    label(q)  $\leftarrow$  c
15    if |N| < minPts then continue                                // Core-point check
16    S  $\leftarrow$  S  $\cup$  N
  
```

Gracias

Inserta texto adicional aquí

