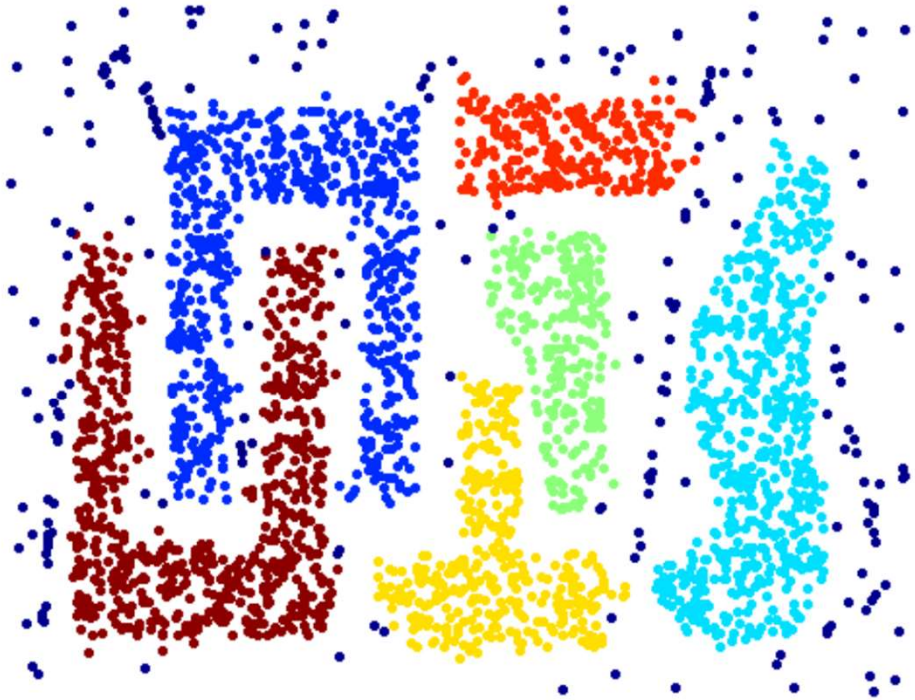
Machine Learning
Clustering
Métodos de densidad: **DBSCAN**

Contenidos

- Introducción
- **Clustering**
  - Introducción
  - Métodos de partición
  - Métodos de densidad
  - Métodos jerárquicos
  - Métodos difusos
  - Evaluación

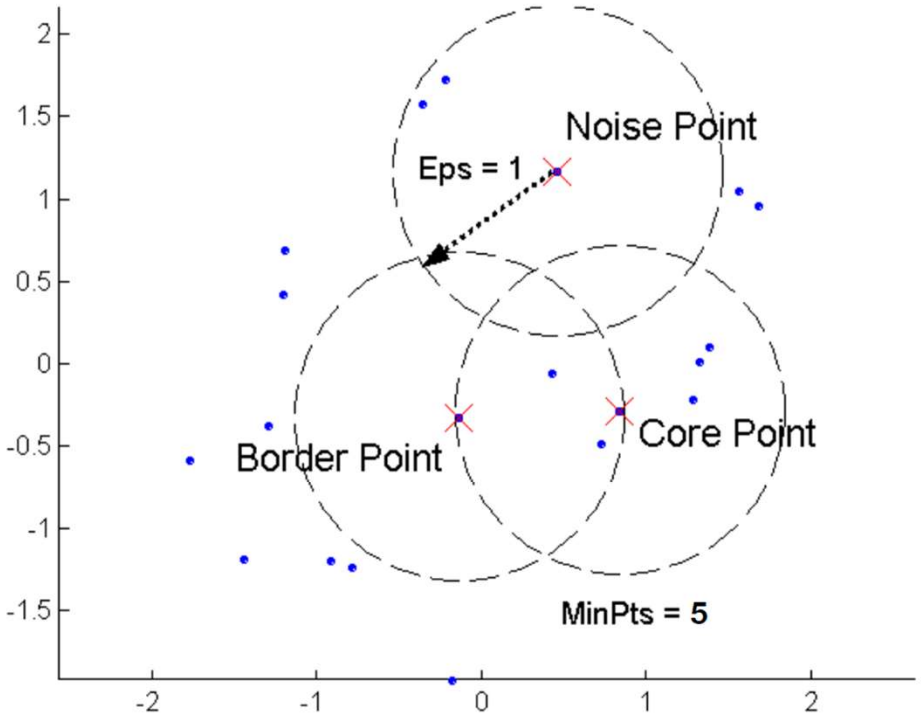
Métodos de densidad, DBSCAN, introducción

- DBSCAN es un método de clustering basado en densidad, donde dado un set de datos en el espacio, agrupa los datos que están cercanos entre ellos.



Métodos de densidad, DBSCAN, definiciones

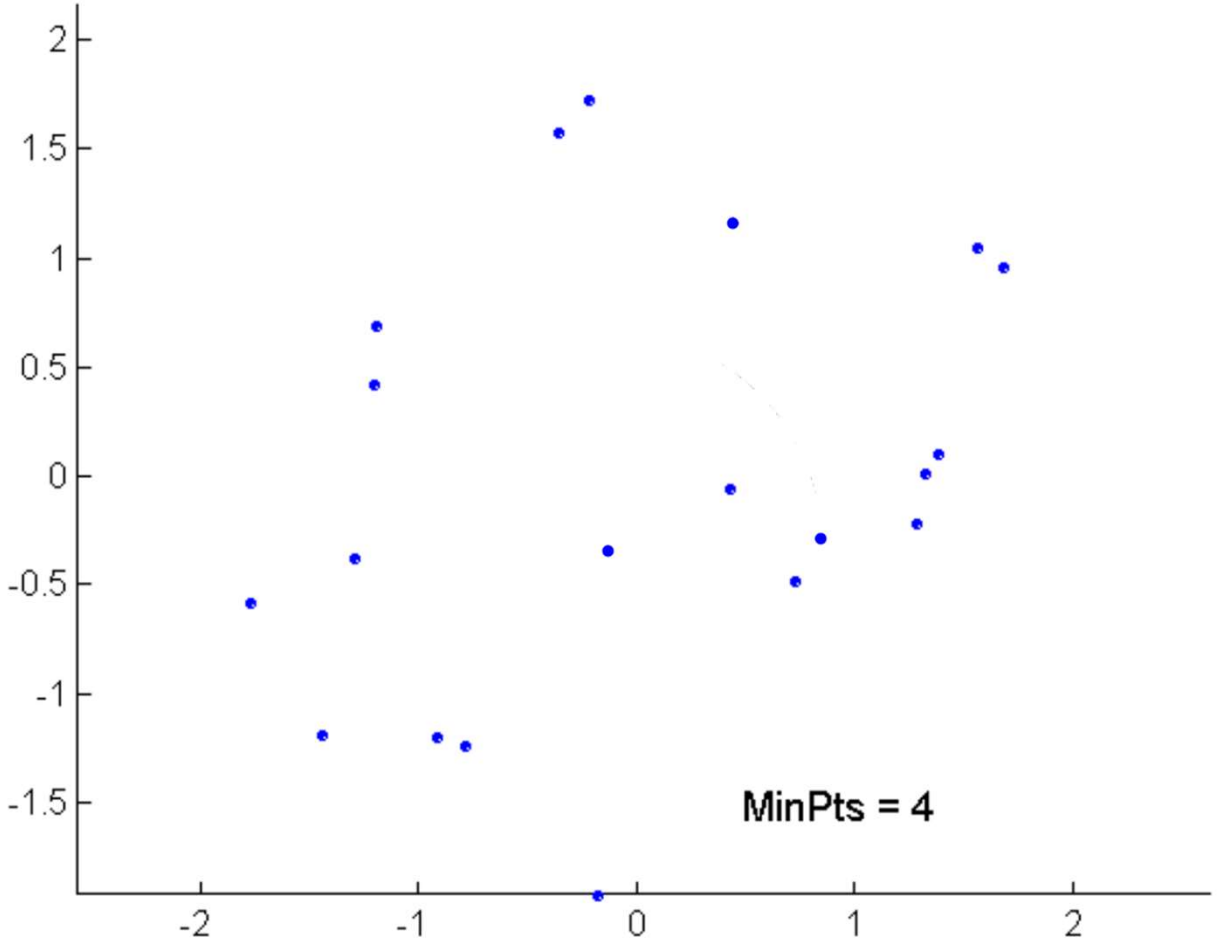
- **Densidad** es el número de puntos dentro de un **radio** específico denominado **Eps.**
- Un **punto central/core** es aquel que tiene al menos **MinPts** puntos dentro de la esfera definida por Eps (se incluye el mismo).
- Un **punto de borde** tiene menos puntos que MinPts del EPS, pero está dentro de la esfera de un punto central.
- Un **punto de ruido** es todo aquel que no es punto central ni de borde.



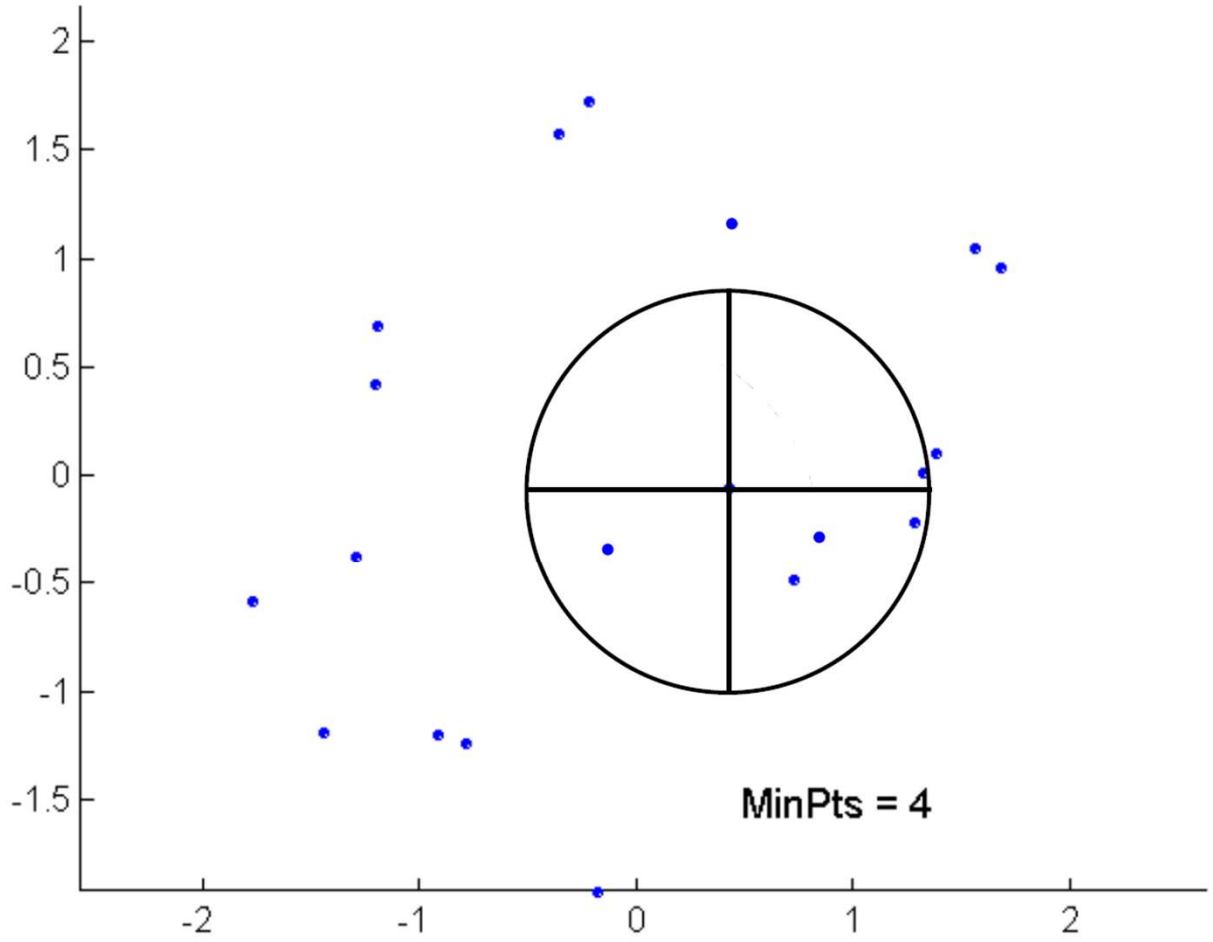
Métodos de densidad, DBSCAN, aprendizaje

- Hay que definir **eps** y **MinPts**
- Se determinan los puntos **centrales**, **borde**, y **ruido**
- Elimina los puntos de ruido

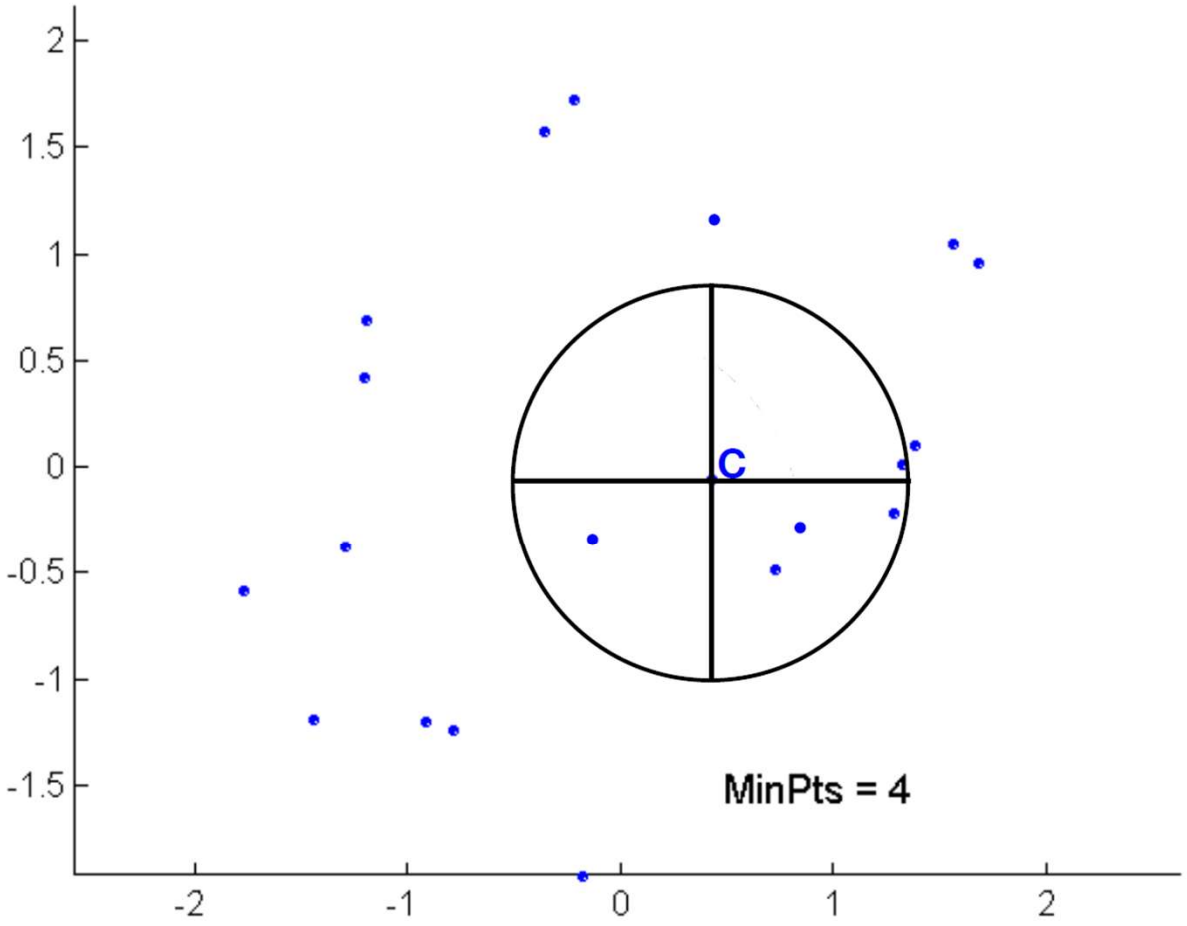
Métodos de densidad, DBSCAN, ejemplo



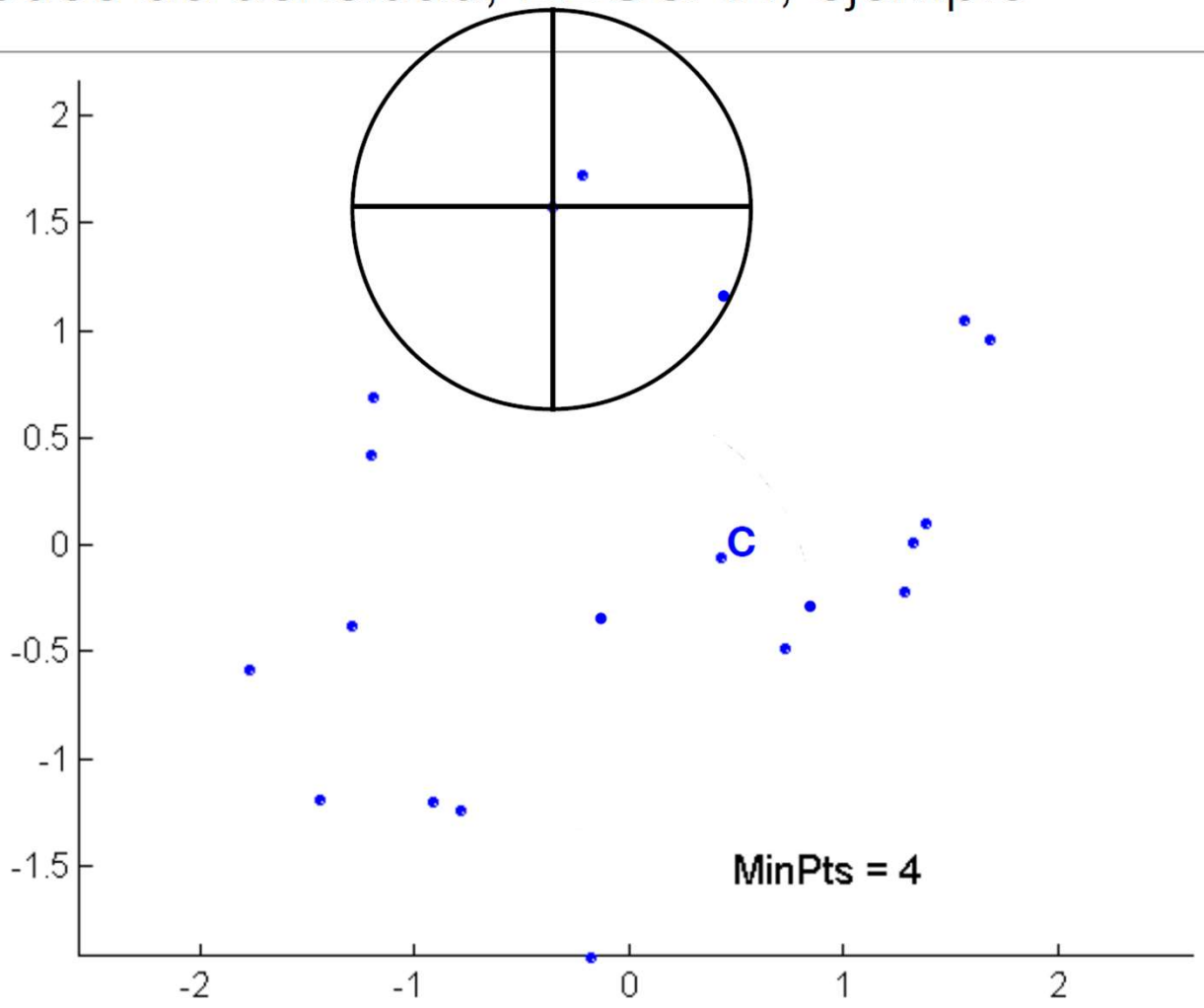
Métodos de densidad, DBSCAN, ejemplo



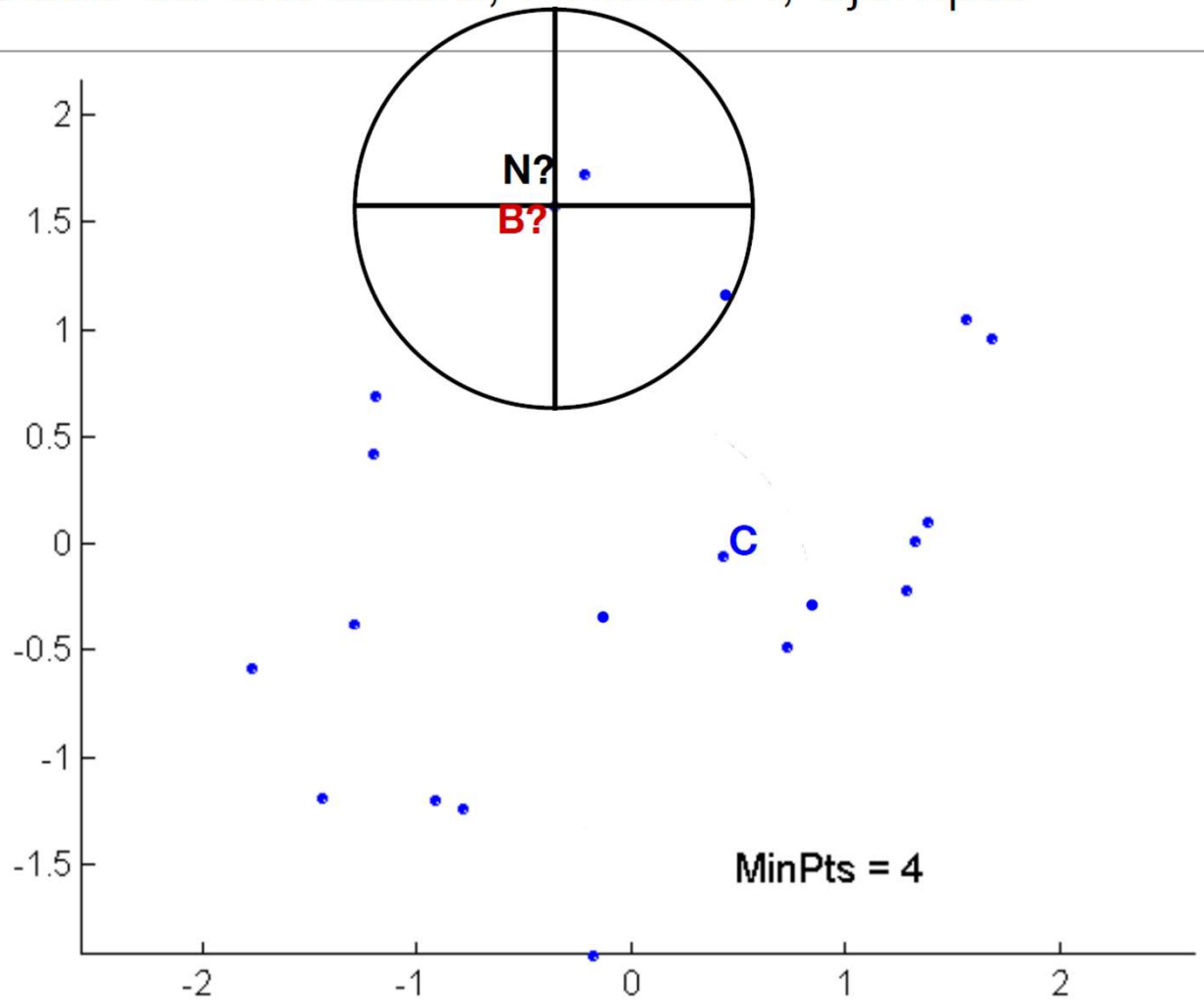
Métodos de densidad, DBSCAN, ejemplo



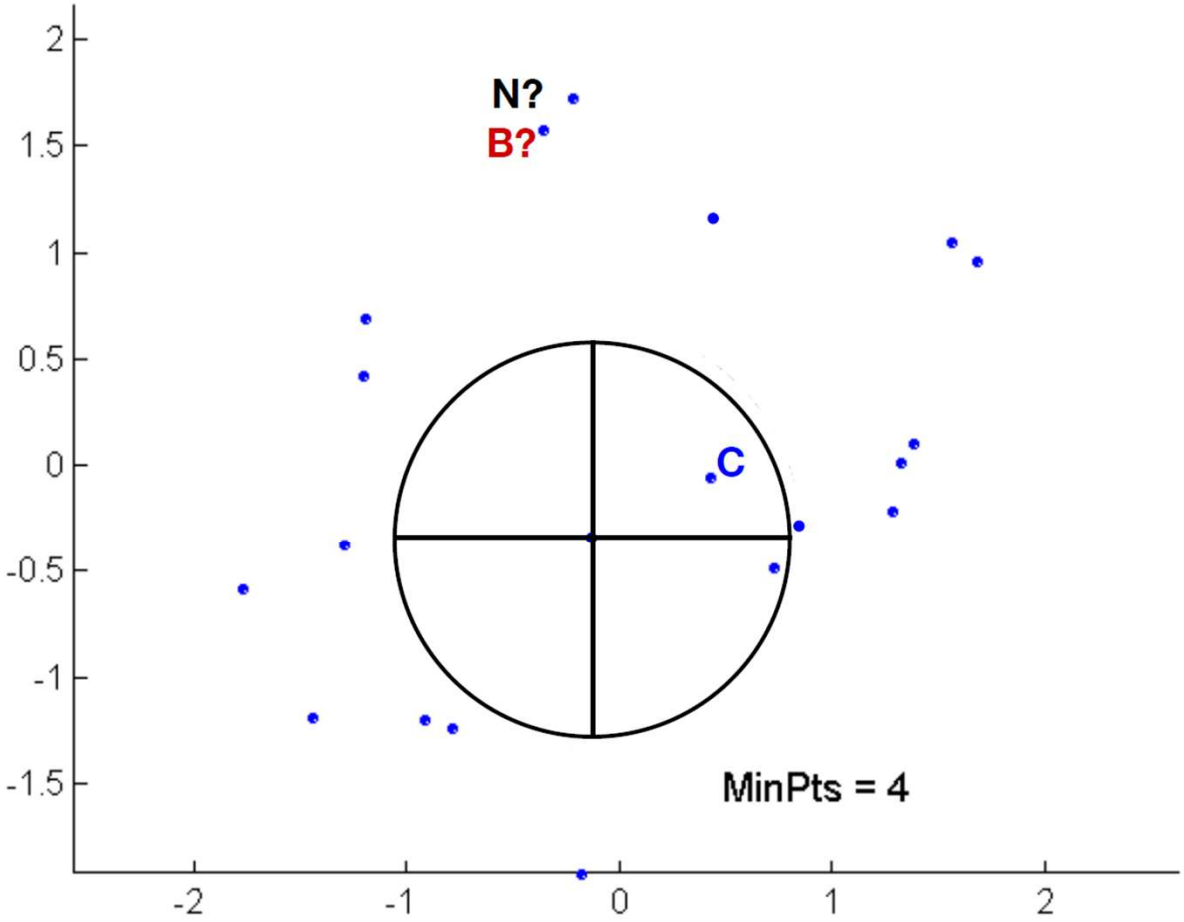
Métodos de densidad, DBSCAN, ejemplo



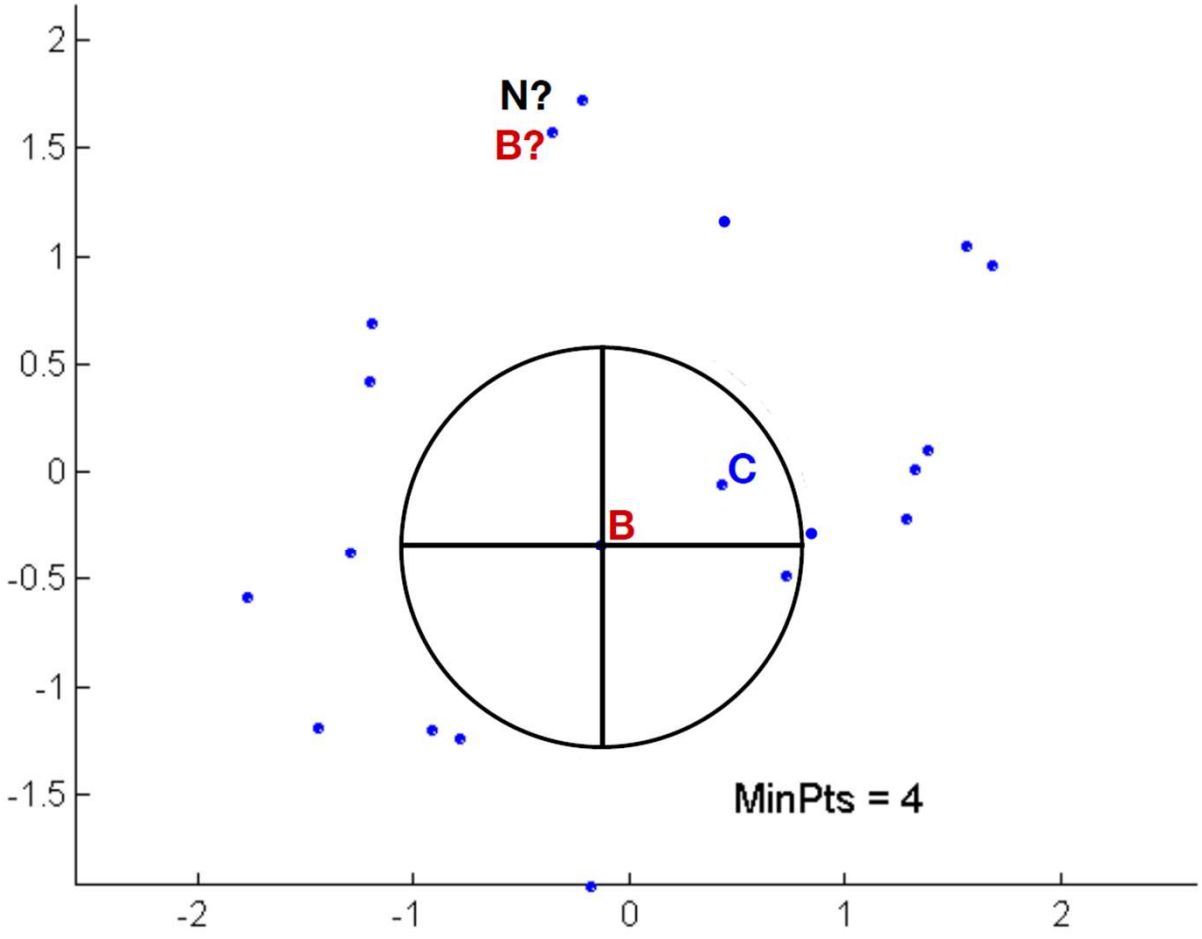
Métodos de densidad, DBSCAN, ejemplo



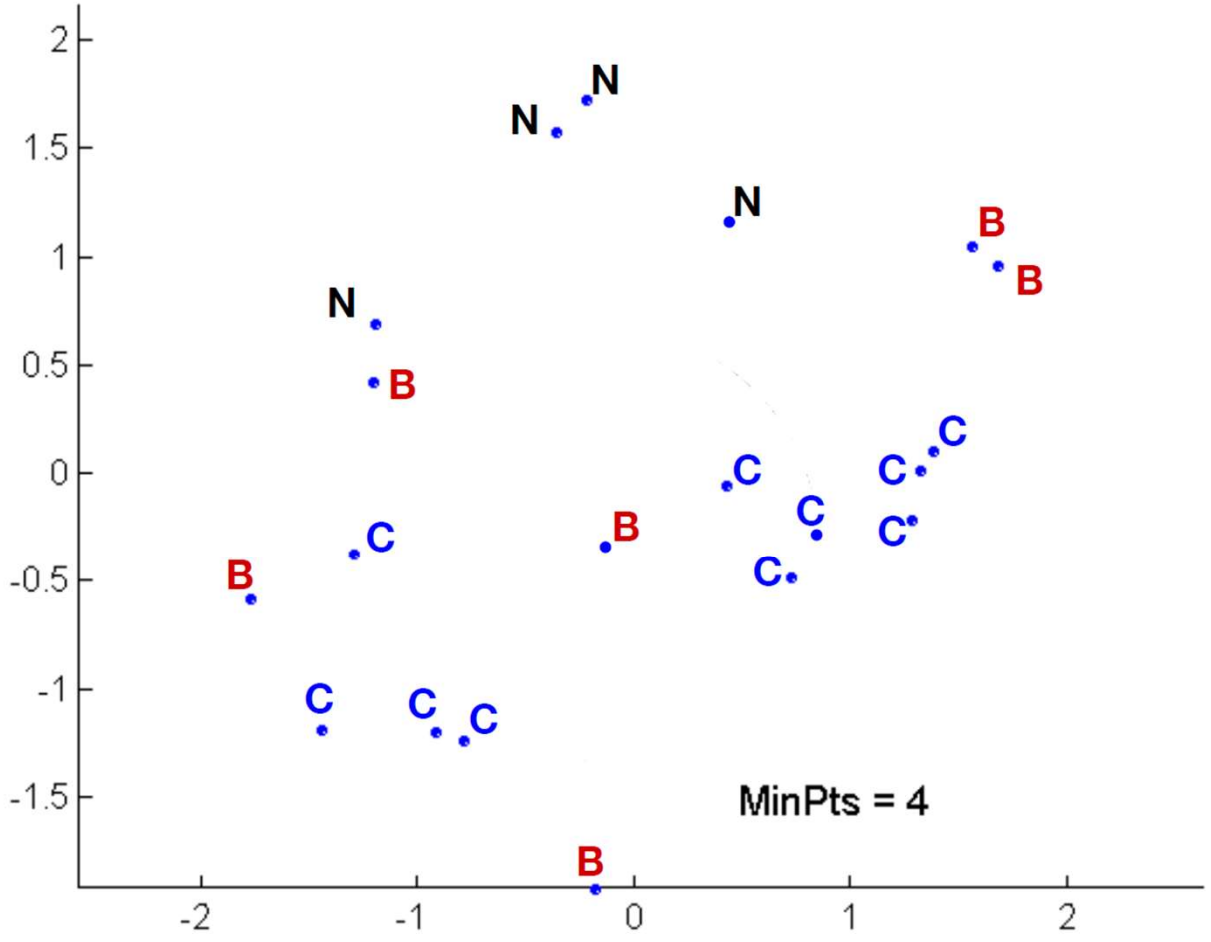
Métodos de densidad, DBSCAN, ejemplo



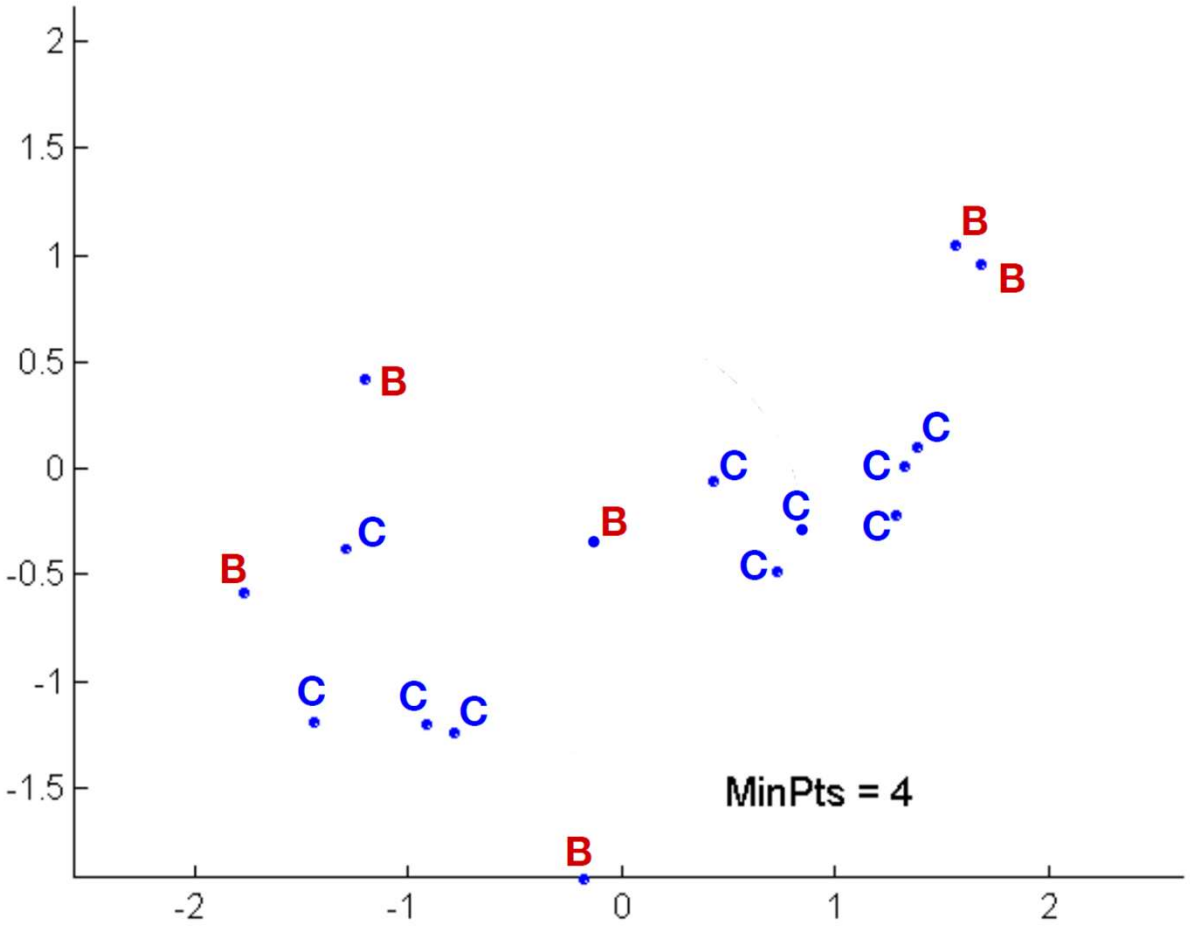
Métodos de densidad, DBSCAN, ejemplo



Métodos de densidad, DBSCAN, ejemplo



Métodos de densidad, DBSCAN, ejemplo



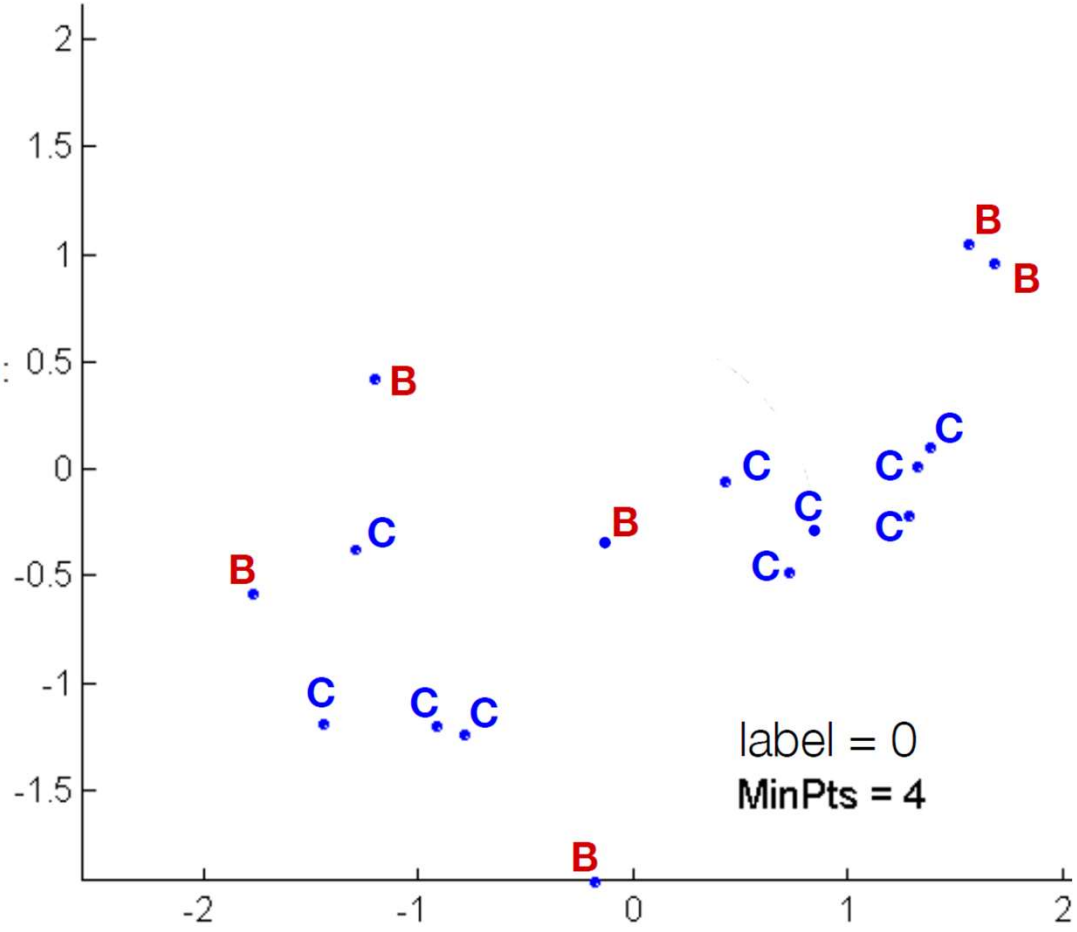
Métodos de densidad, DBSCAN, aprendizaje

- Hay que definir **eps** y **MinPts**
- Se determinan los puntos **centrales**, **borde**, y **ruido**
- Elimina los puntos de ruido
- Aplicar el siguiente algoritmo de clustering:

```
labelCluster = 0
FOR todos los puntos centrales:
    IF punto central no tiene label:
        labelCluster = labelCluster + 1
        Asignar el label actual al punto central
    FOR todos los puntos dentro de la esfera definida por eps:
        IF el punto no tiene label:
            Asignar el label actual al punto central
```

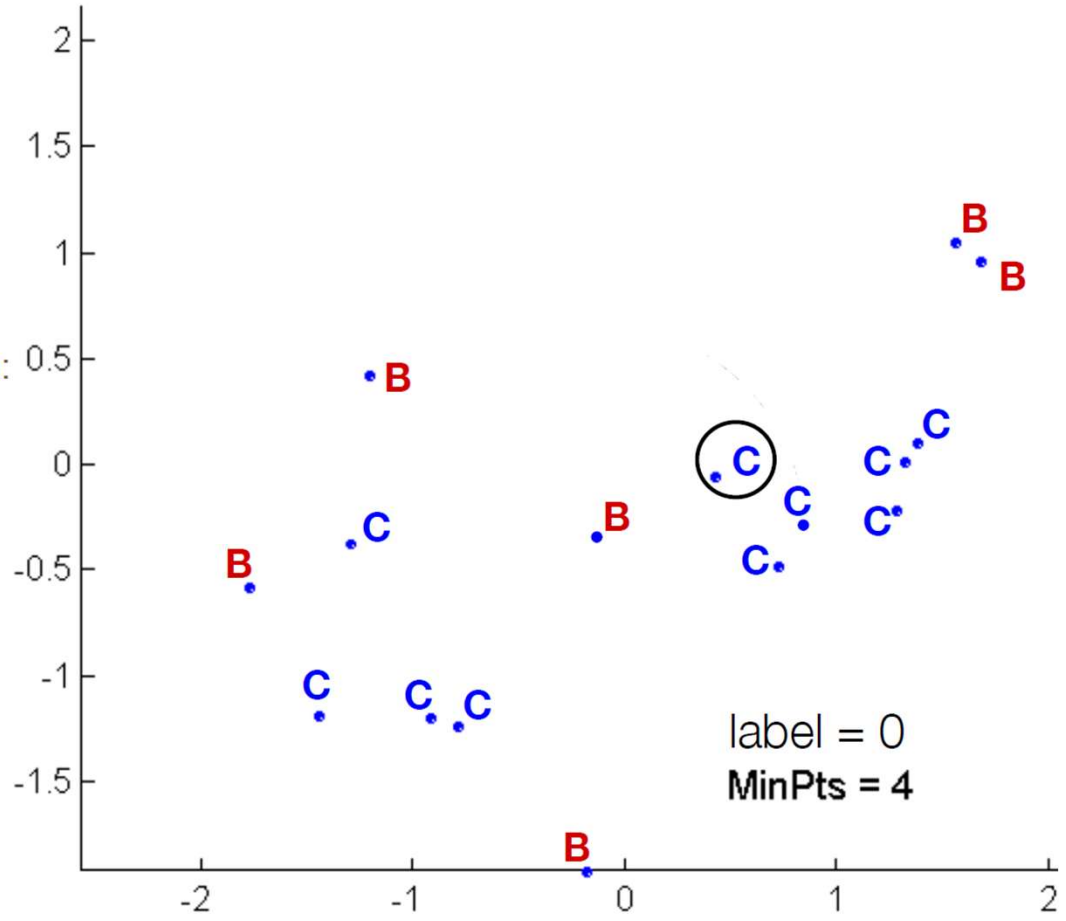
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



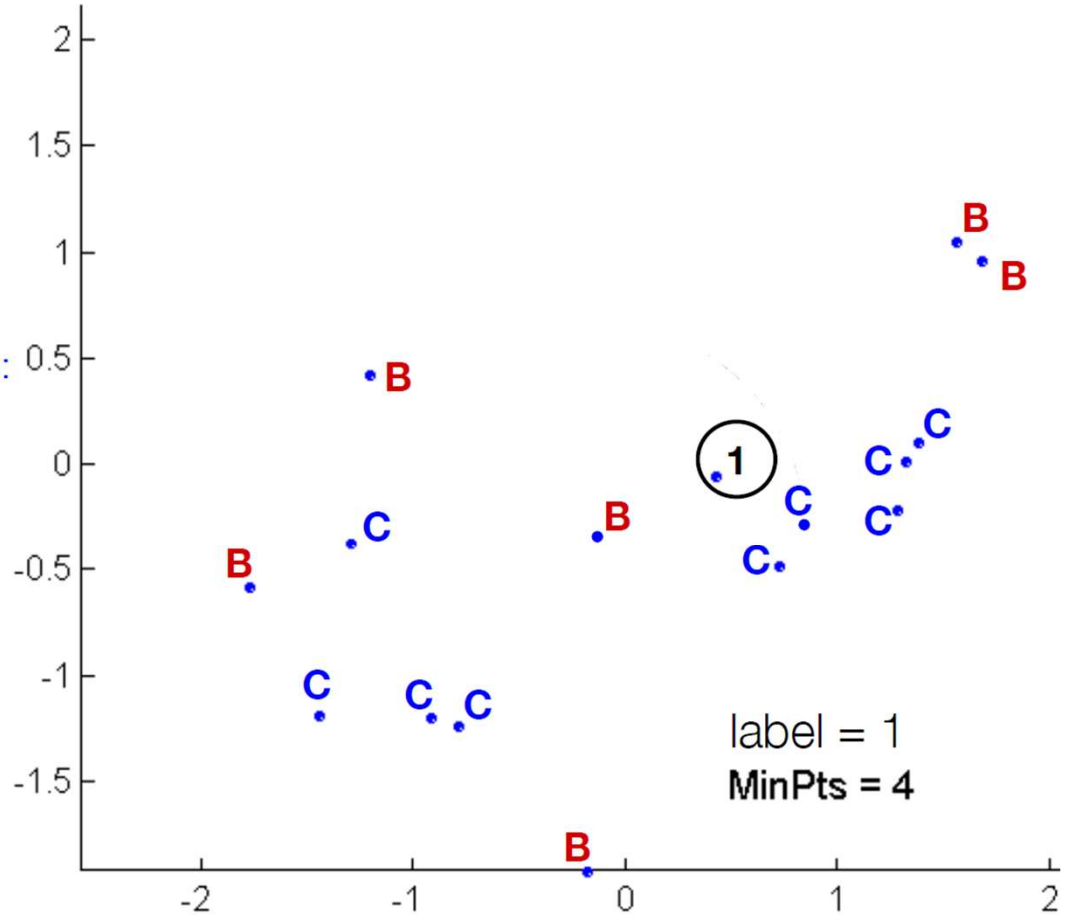
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



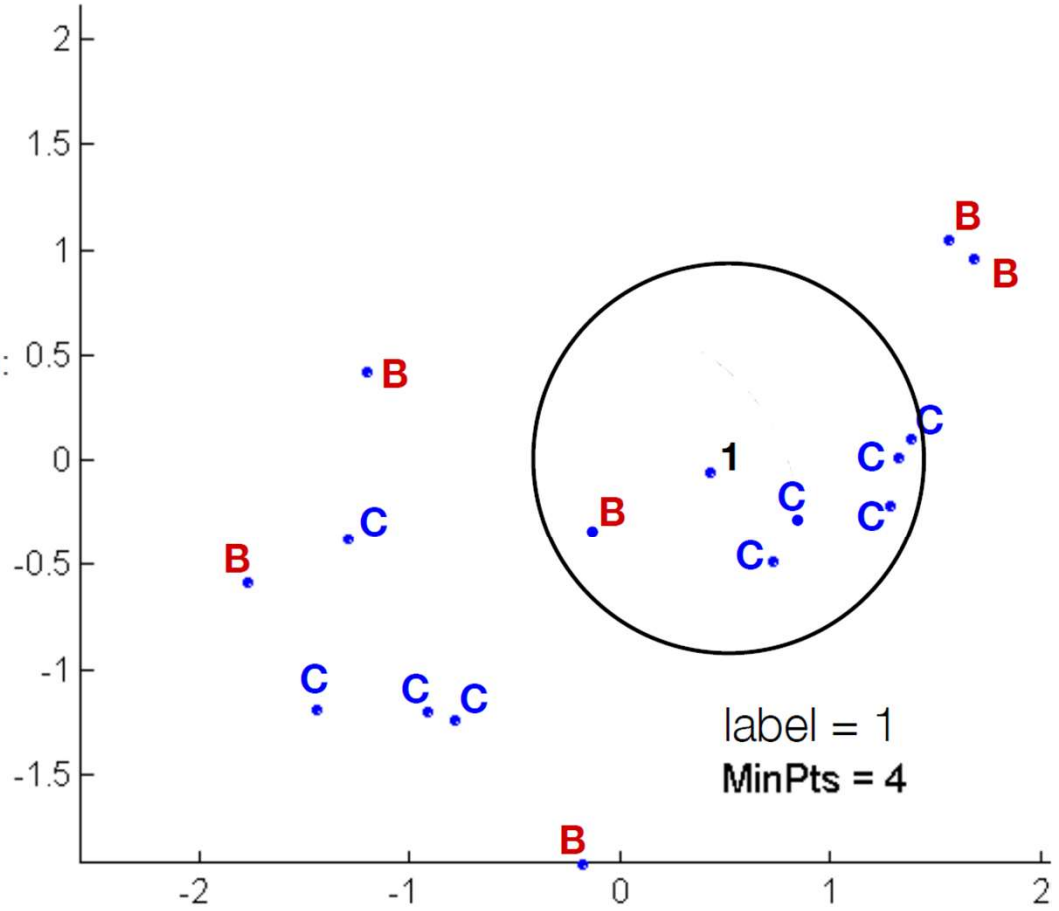
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



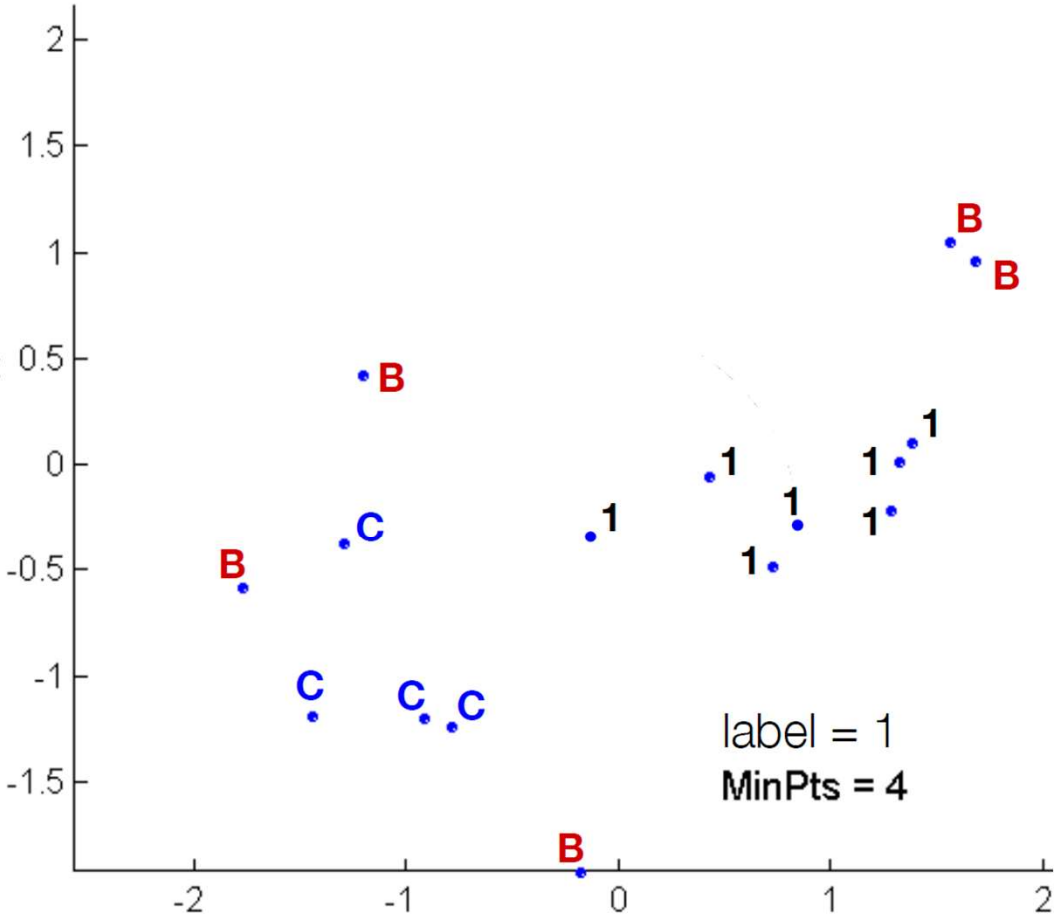
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



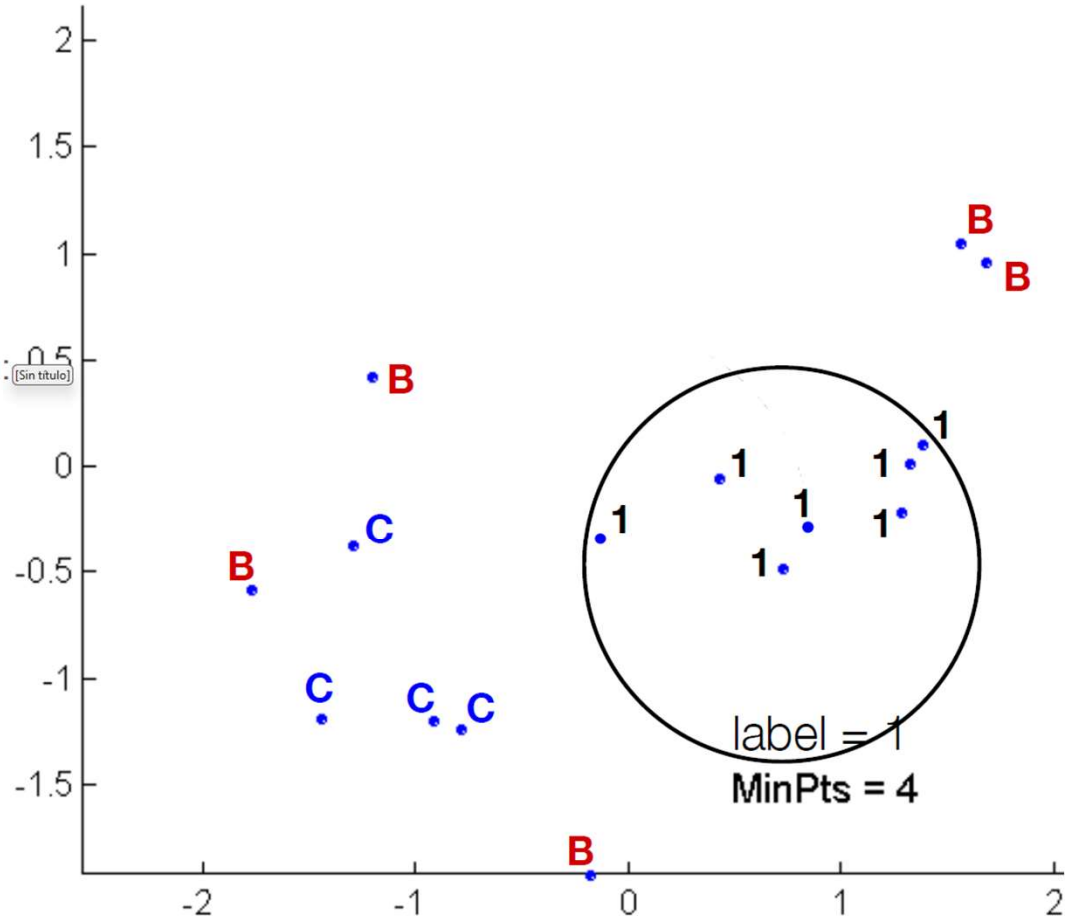
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



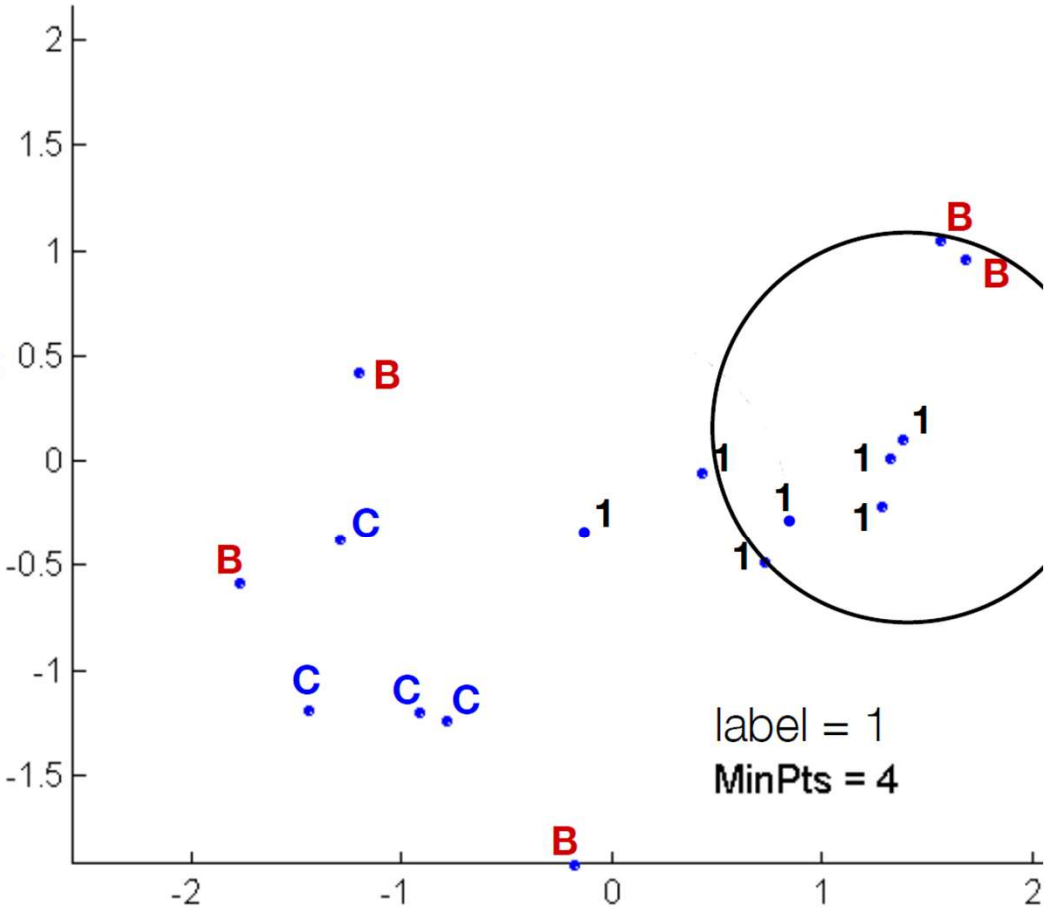
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



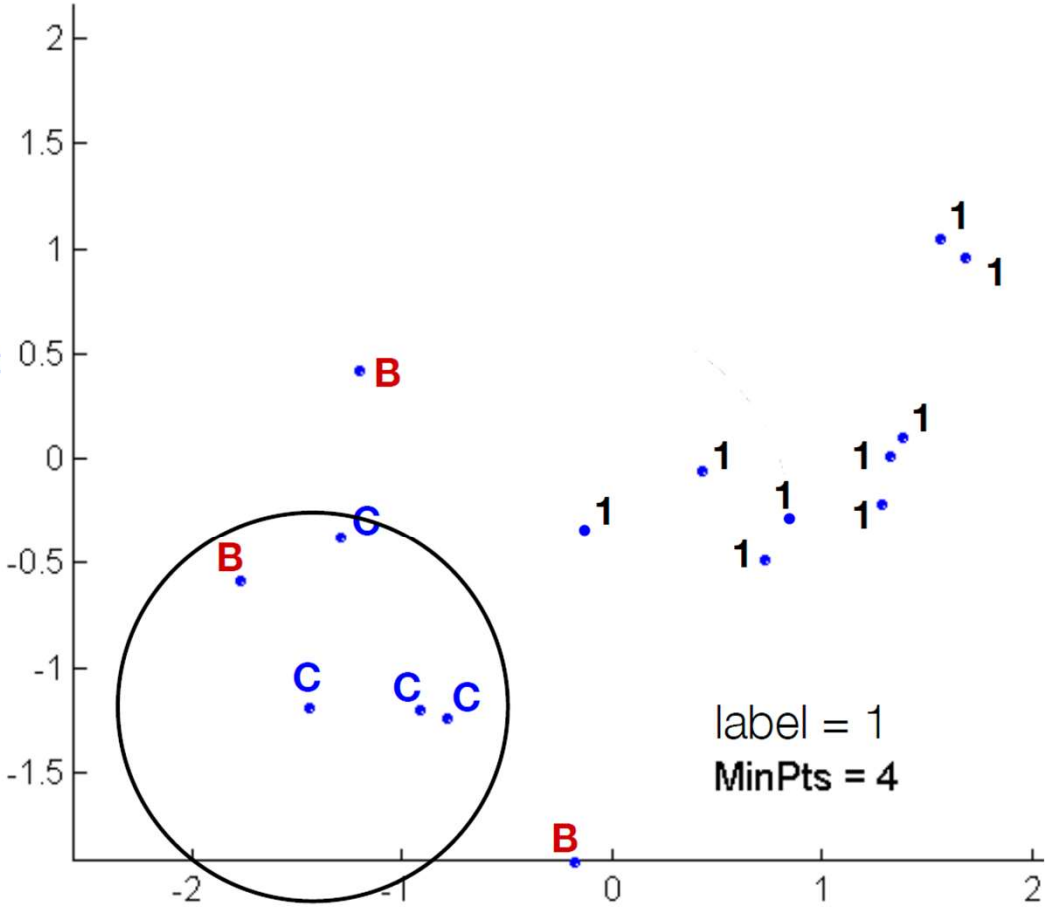
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



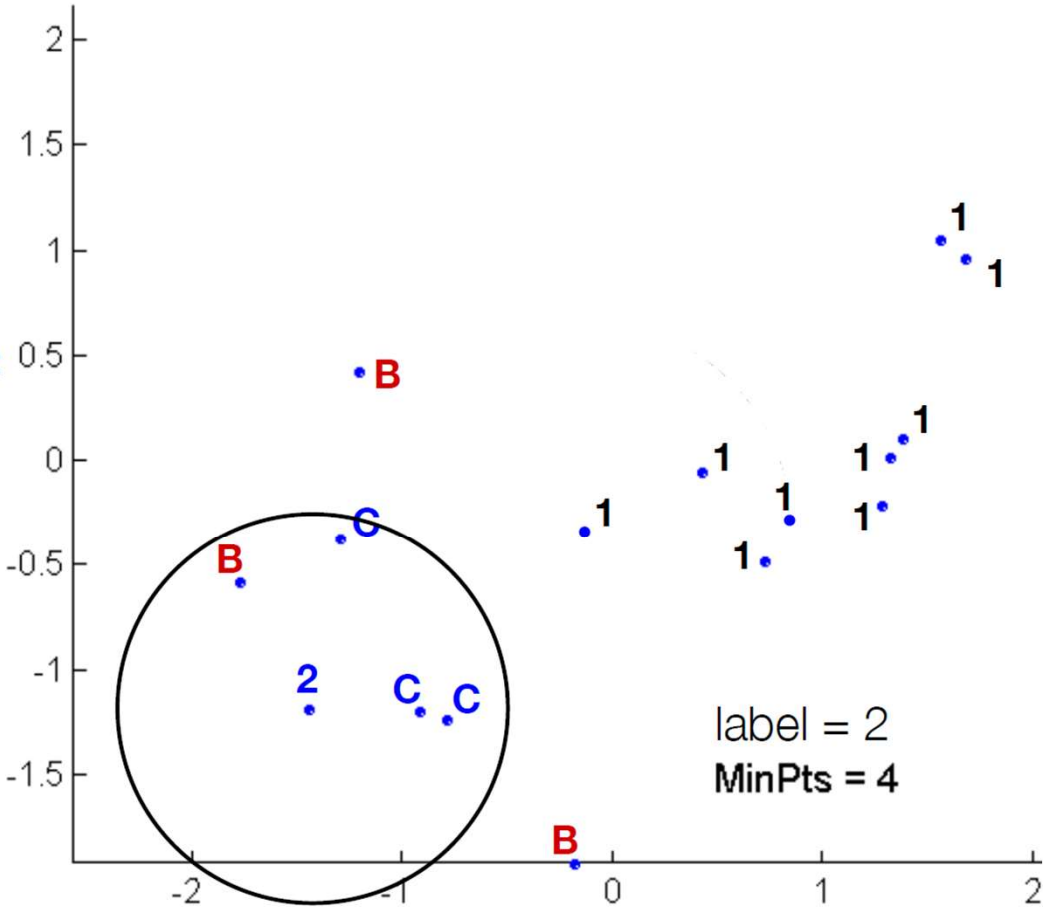
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



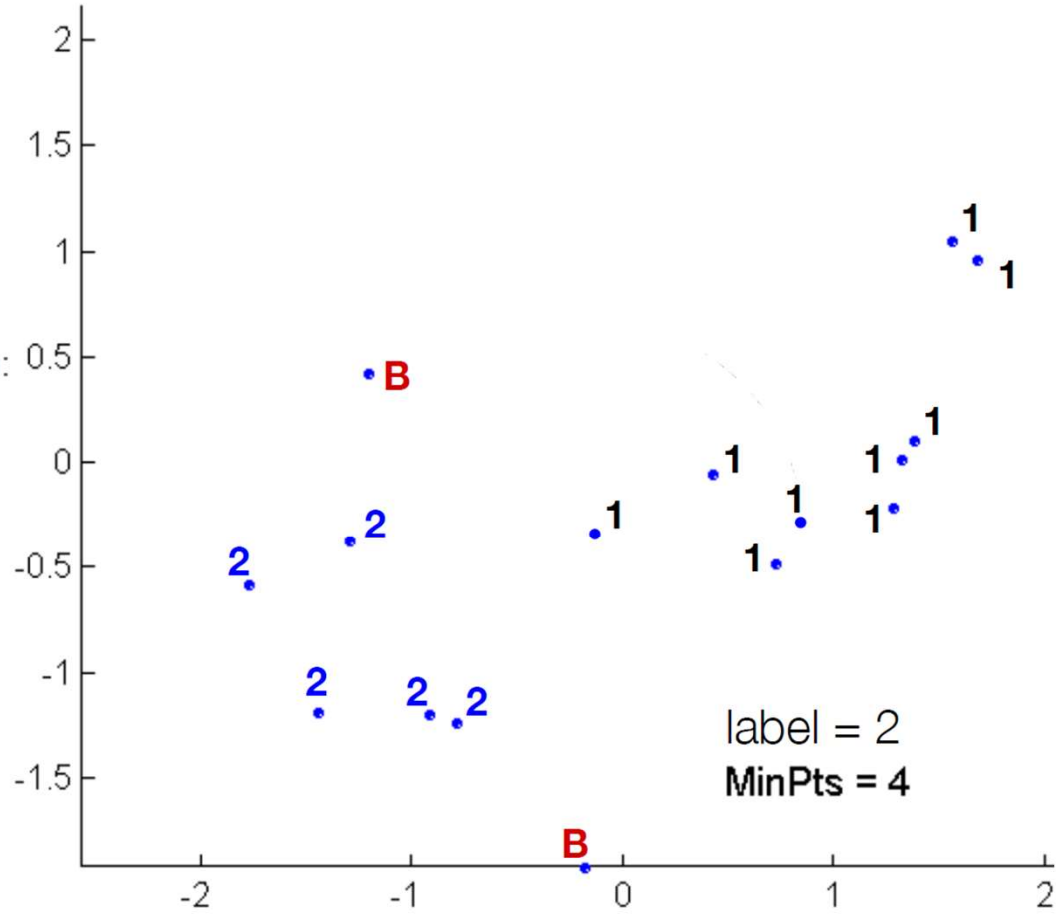
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



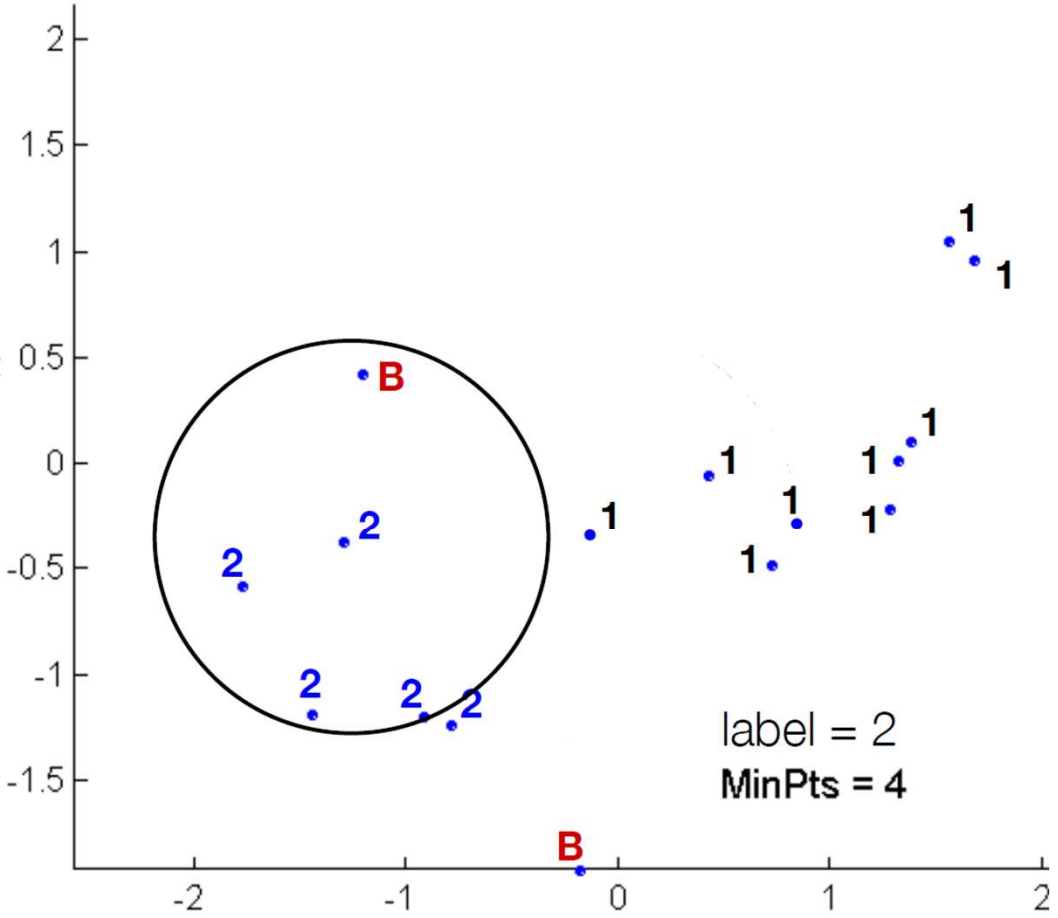
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



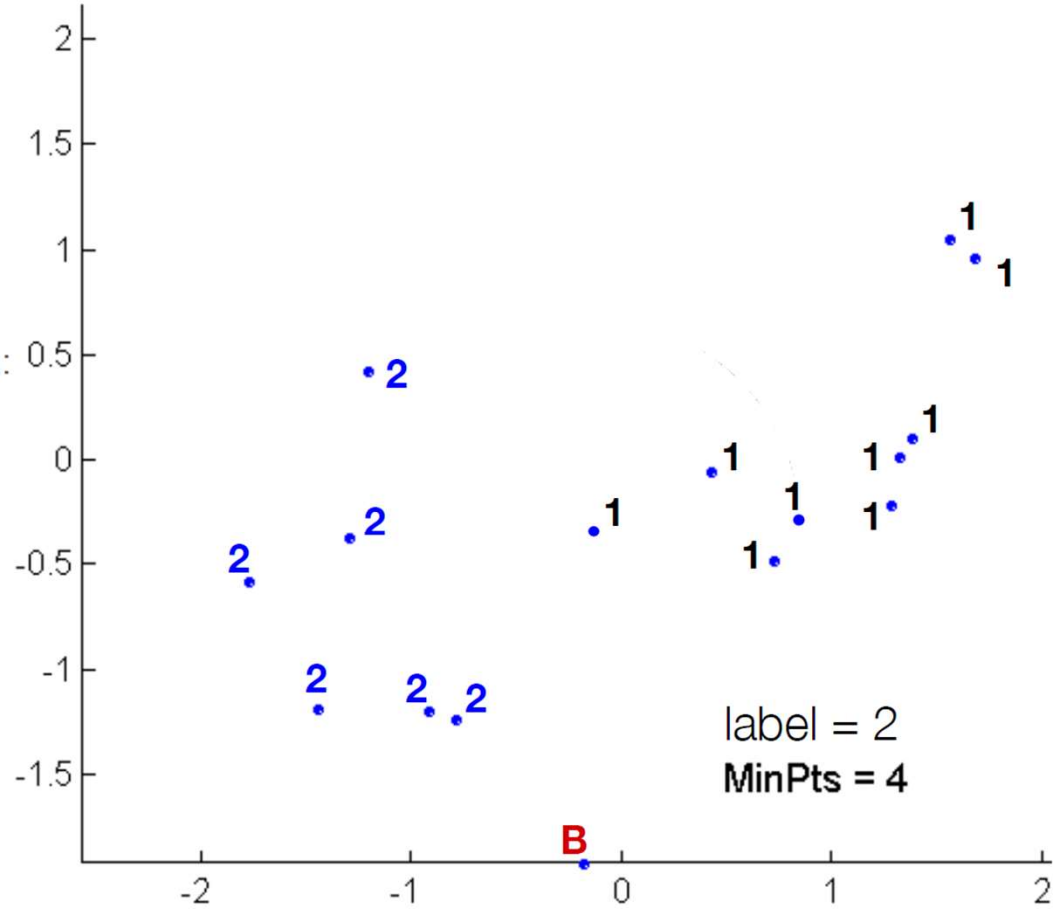
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



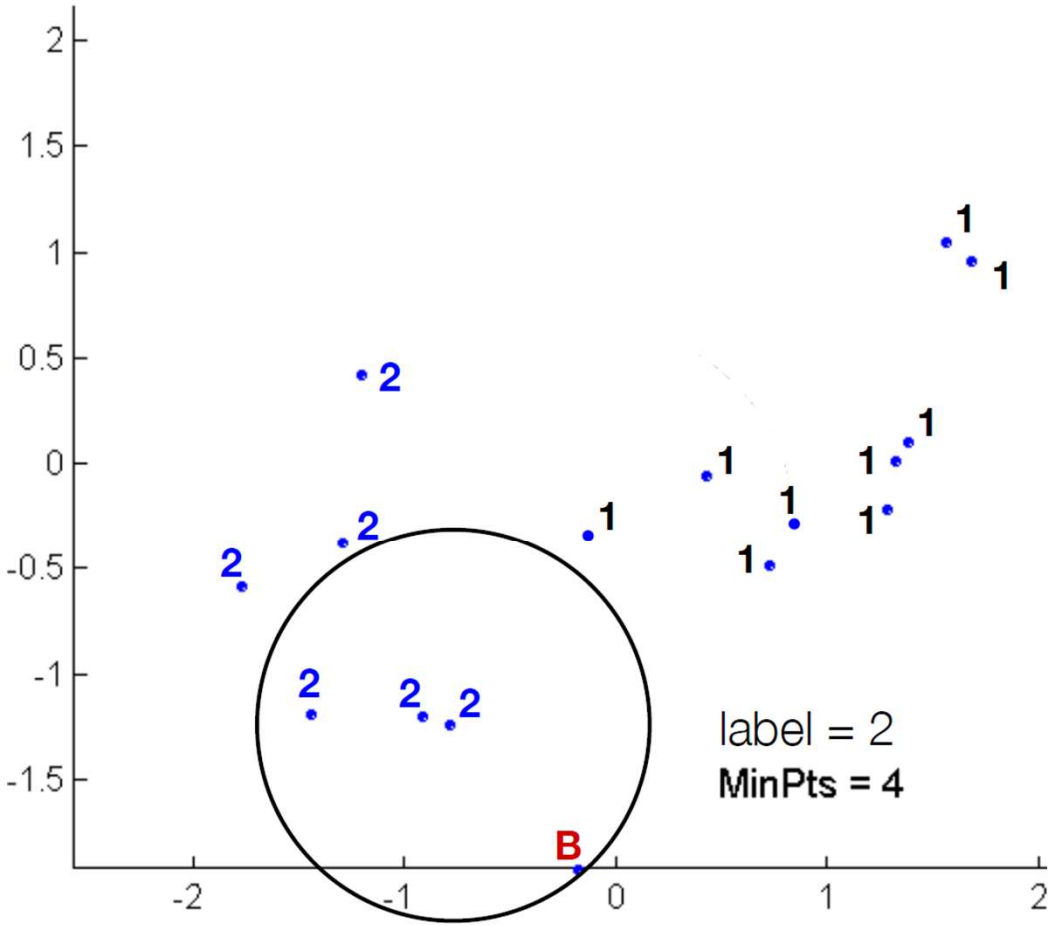
DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```

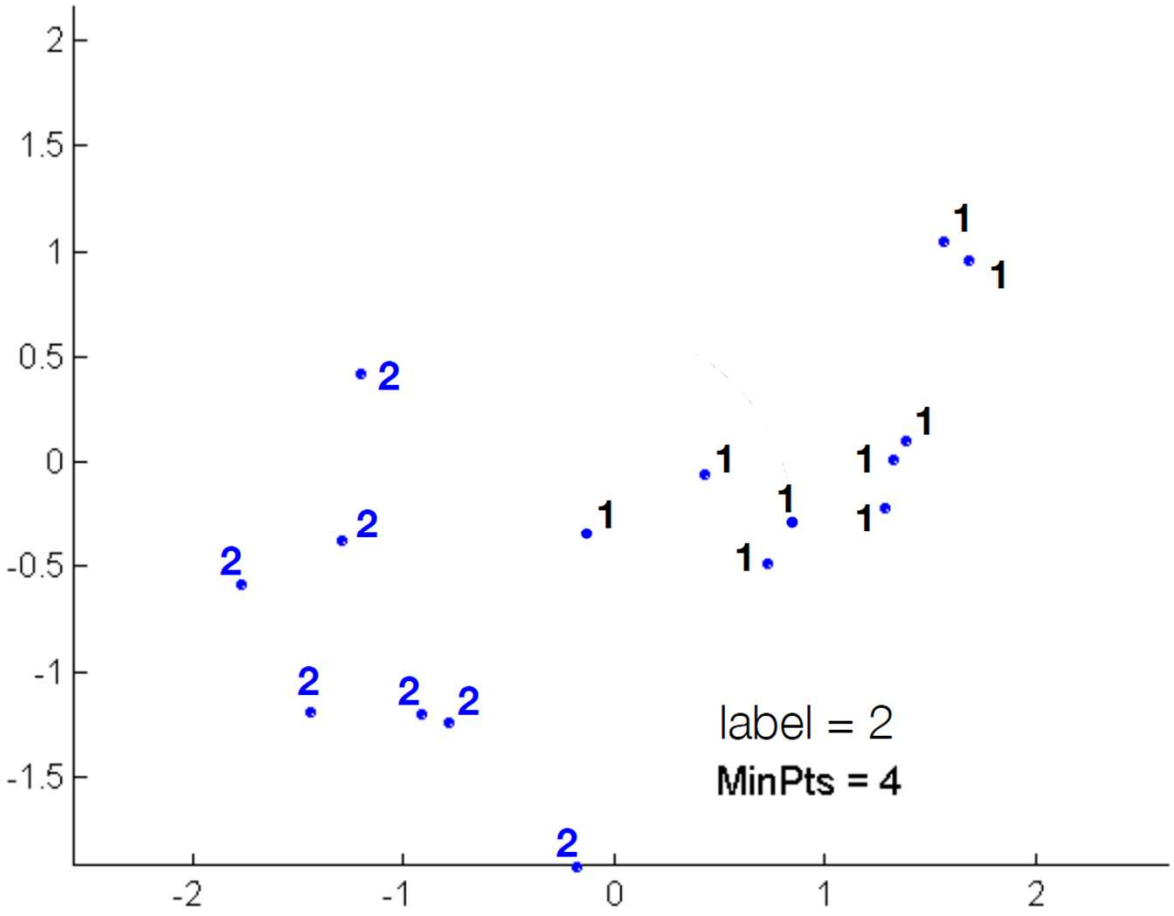


DM, clustering, DBM, DBSCAN, example

```
label=0
i=0
FOR cada puntoCentral:
  IF puntoCentral[i] sin etiqueta:
    label = label + 1
    puntoCentral[i] = label
  j=0
  FOR cada punto en esfera:
    IF punto[j] sin etiqueta:
      punto[j] = label
```



DM, clustering, DBM, DBSCAN, example

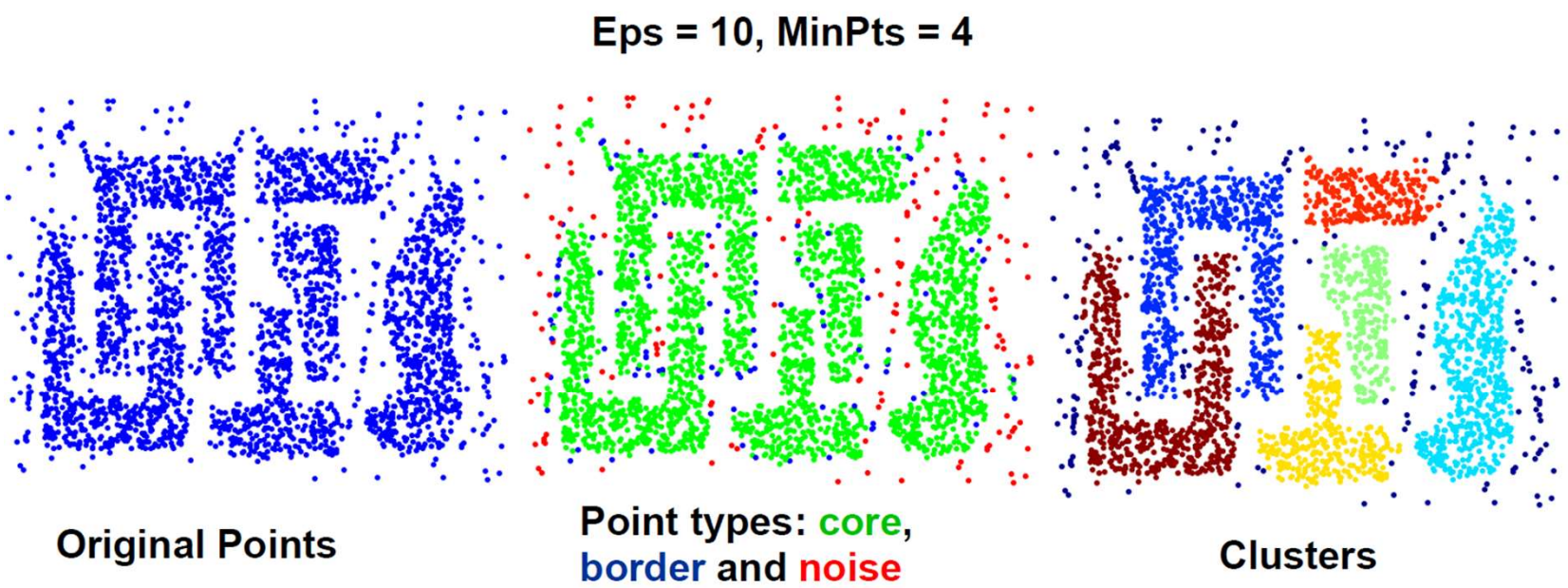


Métodos de densidad, DBSCAN, ejemplo

- ¿Qué pasaría si utilizamos K-medias con K=6 para este tipo de datos? Como K-medias genera clusters circulares, no podría determinar los clusters en forma correcta.

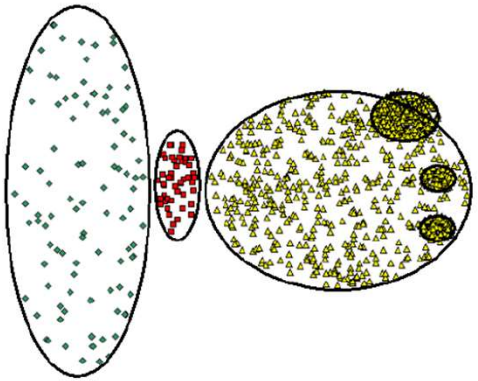
Original Points

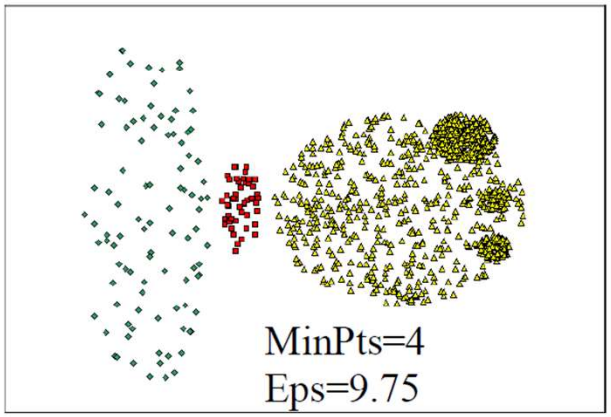
Métodos de densidad, DBSCAN, ejemplo

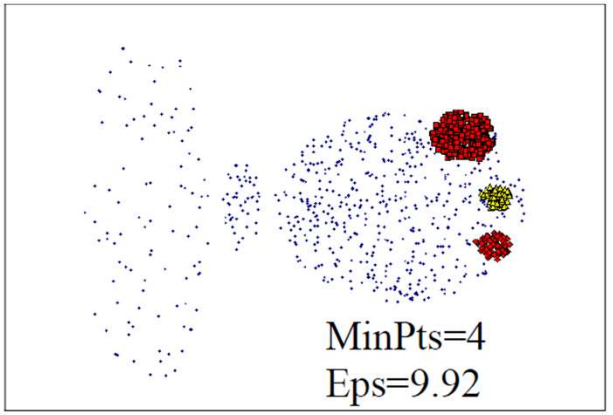


Métodos de densidad, DBSCAN, características

- El algoritmo es computacionalmente caro debido al cálculo de distancia entre todos los puntos.
- **Fortalezas:**
  Resistente a datos atípicos
  Generar clusters de distintos tamaños y formas
- **Debilidad:**
  Afectado por la densidad de los datos
  Datos con un número alto de dimensiones

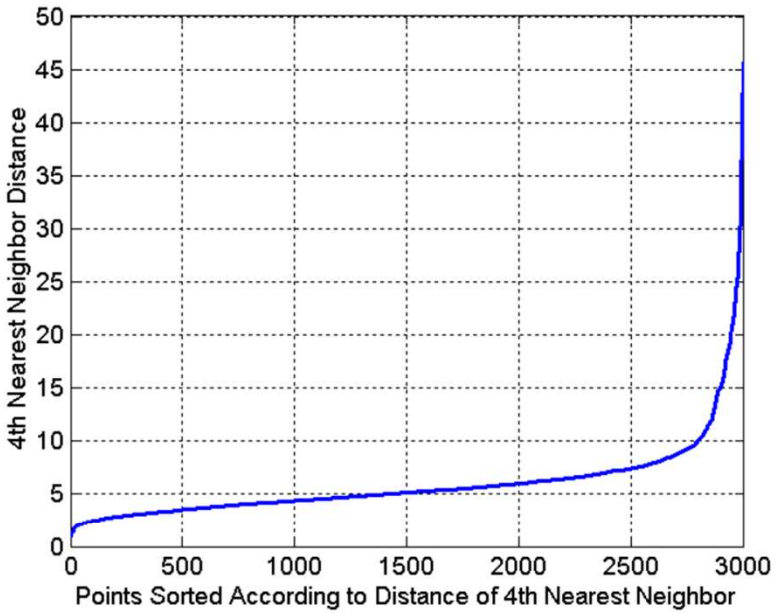






Métodos de densidad, DBSCAN, selección de parámetros

- ¿Como seleccionar **EPS** y **MinPts**?
- Para datos con multiples dimensiones (dim), la regla básica es **minPts ≥ dim+1**
- La idea es que la distancia de los puntos dentro de un cluster a su k^{mo} vecino, sean similares.
- Puntos atípicos tienen a su k^{mo} vecino a una distancia mayor.
- Se calcula la distancia de cada apunto a su k^{mo} vecino, se ordenan de menor a mayor y se grafican, luego se selecciona el **EPS** cercano al crecimiento exponencial.



Métodos de densidad, DBSCAN, código en Python