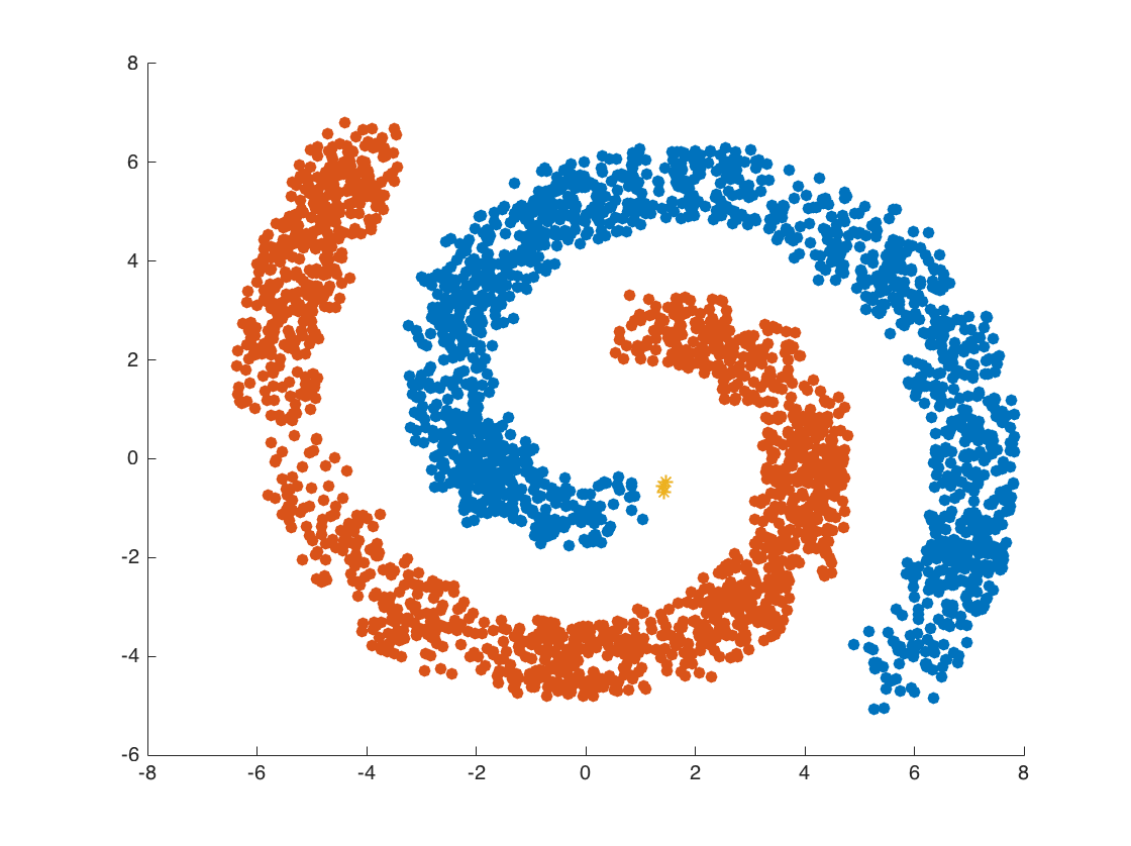


DigitalHouse > Coding School

DATA SCIENCE

Clase 48

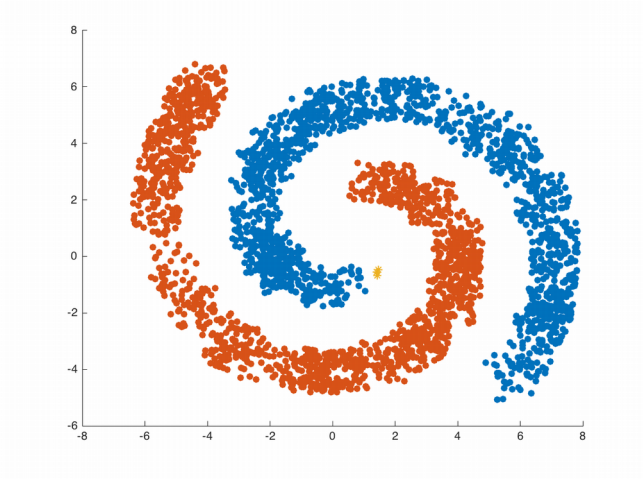
Introducción a DBSCAN

2017

DBSCAN

OBJETIVOS DE LA CLASE

1. **Introducir el concepto de DBSCAN**
2. **Realizar un análisis de clustering usando DBSCAN**
3. **Avanzar en la interpretación de los resultados del análisis**



Density Based Spatial Clustering of Application with Noise (DBSCAN)

- DBSCAN es uno de los algoritmos de clustering más usados:
  - requiere poco tiempo de entrenamiento
  - requiere muy pocos inputs
  - puede encontrar patrones de agrupamiento de (casi) cualquier forma

DBSCAN: ¿Cómo funciona?

Visión general del algoritmo

DBSCAN es un algoritmo basado en la densidad espacial, es decir, busca hallar clusters tratando de encontrar áreas en el dataset que tengan una densidad de puntos mayor al resto del dataset.

Es por eso que puede suceder que, una vez corrido, persistan puntos que no han sido asignados a ningún cluster.

Dos parámetros:

- **epsilon:** es la máxima distancia entre dos puntos para considerarlos como pertenecientes al mismo cluster (también se lo puede pensar como el radio máximo del cluster)

- **min_points:** es decir, el mínimo de puntos necesarios para formar un cluster (la idea de este parámetros es evitar la formación de clusters demasiado pequeños)

Visión general del algoritmo

- Primero definimos los dos parámetros
- Luego, elegimos un punto arbitrario en el dataset y si existen **min_points** dentro de la distancia **epsilon** desde el punto (incluyéndolo) todos ellos forman un cluster.
- Luego, expandimos el cluster chequeando todos los puntos nuevos y evaluando para cada uno de llos si existen **min_points** dentro de la distancia **epsilon** haciendo crecer el cluster recursivamente.
- Eventualmente, se acaban los puntos para agregar al cluster. Elegimos, entonces, un nuevo punto arbitrario del dataset y repetimos el proceso.
- Puede suceder (y es común que ocurra) que alguno de los puntos que son seleccionados tenga menos de **min_points** en su distancia **epsilon** y no forme parte de ningún clluster. En este caso se lo considera punto “ruido” (noise point).

DBSCAN: DIFERENCIAS Y VENTAJAS EN RELACIÓN A OTROS MÉTODOS

- Visualización muy útil para entender DBSCAN y K-means
  - https://www.naftaliharris.com/blog/visualizing-dbscan-clustering/

Diferencias con k-means y jerárquico

DBSCAN: DIFERENCIAS Y VENTAJAS EN RELACIÓN A OTROS MÉTODOS

Para interpretar el dendograma veamos un ejemplo un poco más complejo

- Funciona muy bien con clusters con límites no lineales y de tamaños muy diferentes
- La diferencia fundamental con k-means es que se basa en la “densidad” en lugar de comenzar por la distancia a un punto central (centroide).
- No requiere la definición de un número de clusters a priori
- También es muy útil cuando tenemos datos muy densamente distribuidos:
  - en estos casos, es probable que k-means nos arroje un solo cluster
  - DBSCAN debería poder “romper” este agrupamiento en partes menores
- Puede encontrar clusters con límites de formas “irregulares”
- Incorpora en el proceso el concepto de “ruido”, lo que lo hace robusto a outliers
- Es determinístico (una vez fijados los parámetros iniciales)

DBSCAN: ALGUNAS DESVENTAJAS

- Una de las mayores dificultades se presenta en datasets con zonas con densidades muy diferentes en el espacio de puntos: en estos puede no funcionar correctamente

- En algunos casos, si no se tiene clara la escala de los datos los parámetros pueden ser difíciles de encontrar.

Práctica Guiada: DBSCAN con dataset iris

CONCLUSIÓN

- Presentamos DBSCAN
- No requiere definir previamente un número de clusters de antemano
- Es determinístico - no arroja un resultado diferente cada vez que se ejecuta
- Puede encontrar “bordes” de clusters con formas irregulares
- Es robusto a outliers