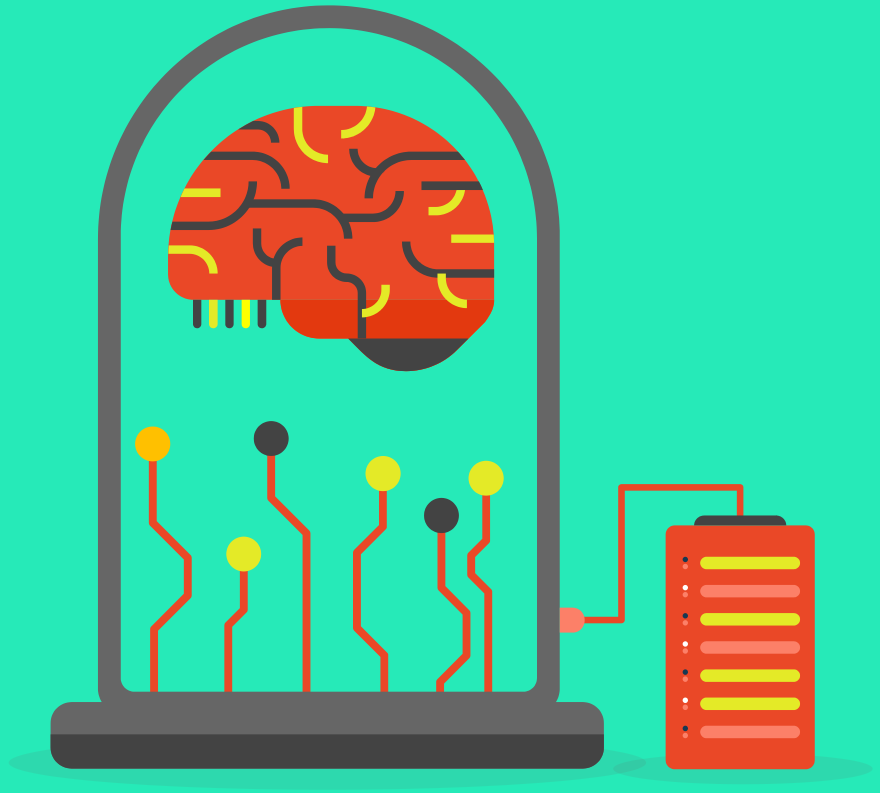
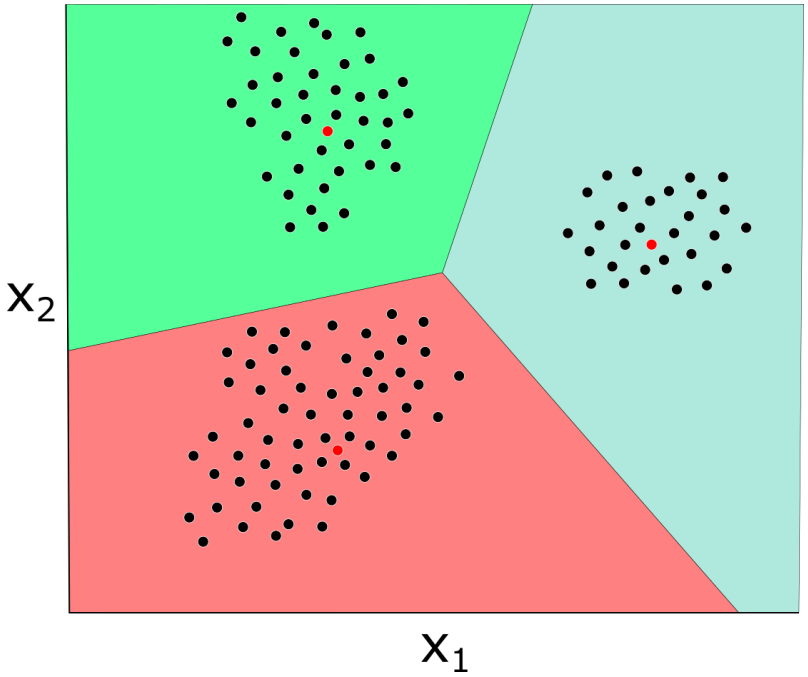
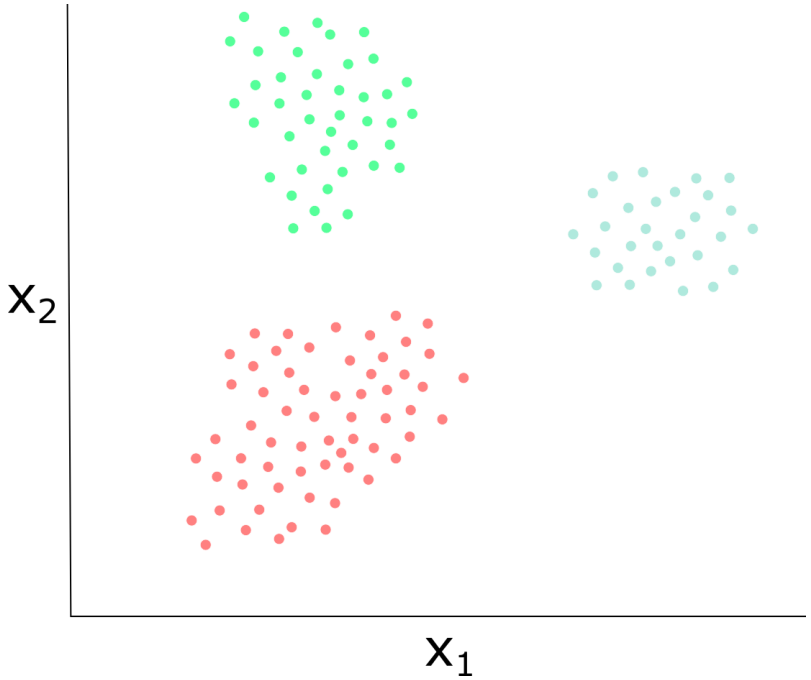
17. Algoritmo DBSCAN



Planteamiento del Problema

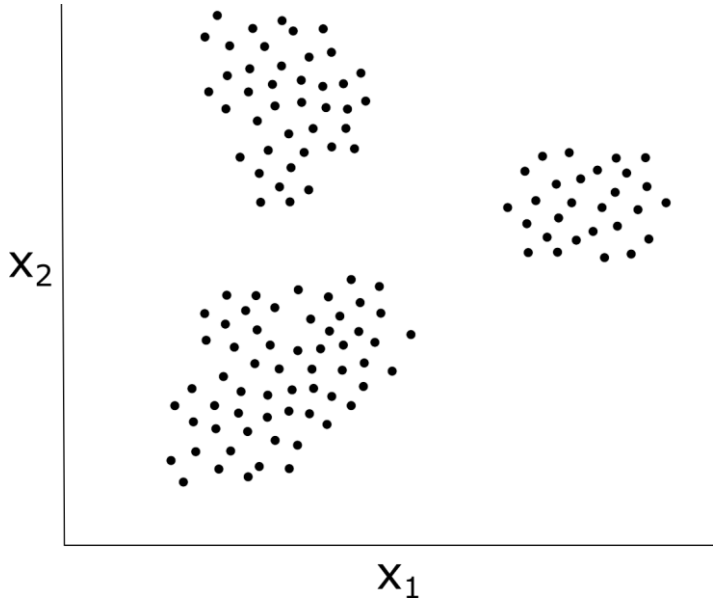
En el tema anterior introdujimos el algoritmo de K Medios, cuyo funcionamiento se basa en encontrar agrupamientos de datos a partir de centroides iniciales que cambiaban su posición en función de los datos cercanos.





Planteamiento del Problema

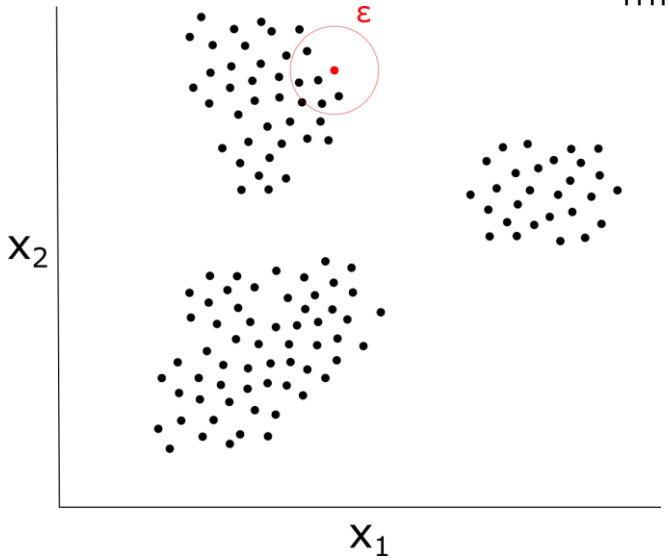
Sin embargo, algo que podemos encontrar en cualquier tipo de dato con características de agrupamientos, es la presencia de “densidad” en dichos datos. Es decir, lugares en donde datos de clases distintas tienen mayor densidad que en otros lugares.



Algoritmo DBSCAN

(Density Based Spatial Clustering of Applications with Noise)

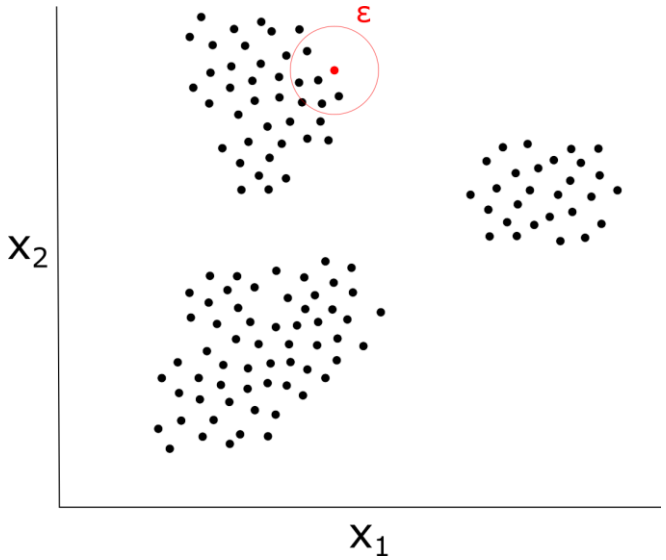
En este sentido, otro tipo de algoritmos, llamados algoritmos basados en “Densidades”, se aprovechan de esta característica para llevar a cabo el proceso de agrupamiento. Específicamente, el algoritmo de DBSCAN funciona de la siguiente manera: para cada dato, se selecciona un radio definido por una medida de distancia **ε** y una cantidad mínima de puntos **p**.

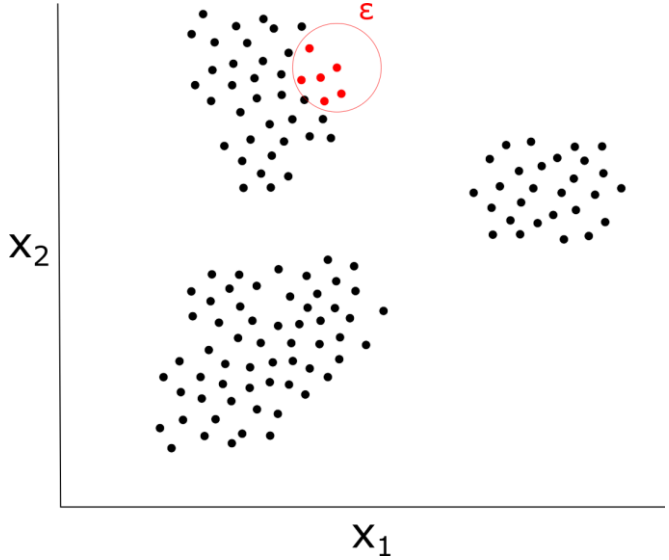


Algoritmo DBSCAN

(Density Based Spatial Clustering of Applications with Noise)

Si dentro de dicho radio hay al menos la cantidad mínima de puntos definida, se considera esa región como perteneciente a una clase. Los puntos fuera se consideran otra clase. Luego, este procedimiento se va repitiendo para cada uno de los puntos del conjunto de datos.

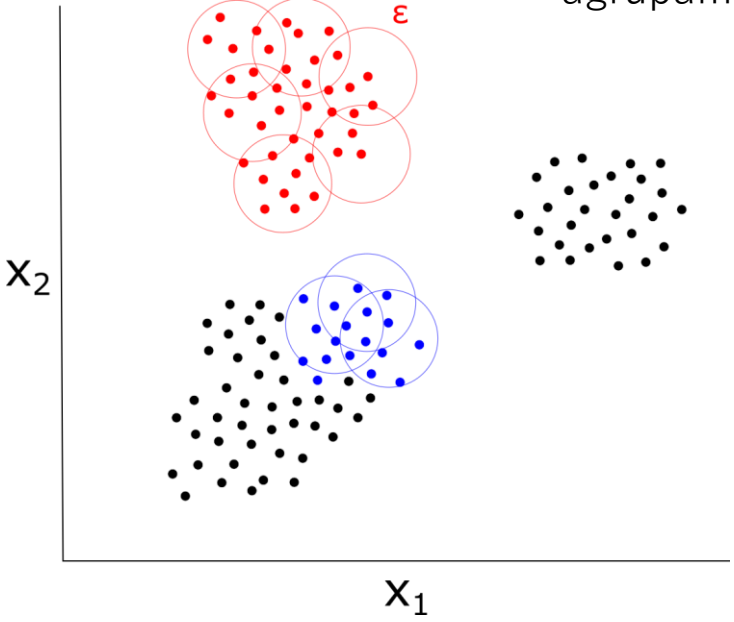


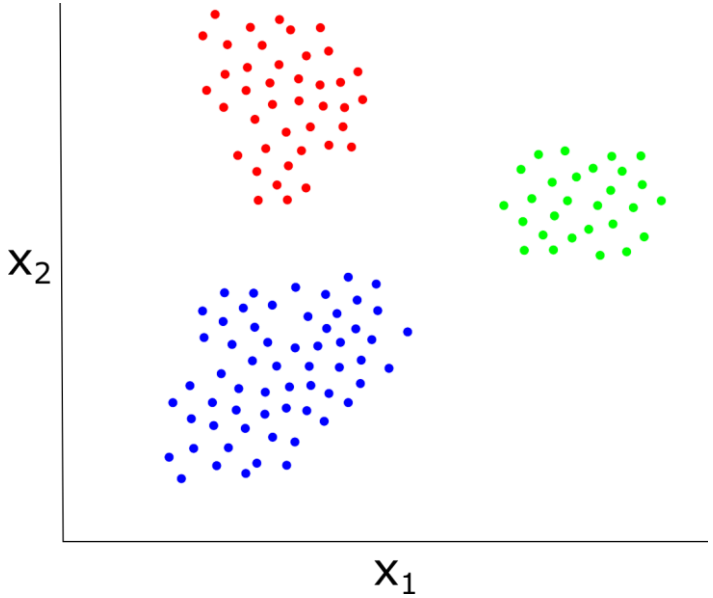


Algoritmo DBSCAN

(Density Based Spatial Clustering of Applications with Noise)

Lo que va a ocurrir es que aquellos puntos que se encuentren en zonas densas, se agruparan en clases similares, estas a su vez separadas por zonas poco densas. Al recorrer todos los puntos, y en base a los parámetros del algoritmo, se conseguirán los agrupamientos detectados.





Algoritmo DBSCAN

(Density Based Spatial Clustering of Applications with Noise)

Una de las ventajas del algoritmo de DBSCAN frente al de K Medios, es que la cantidad de clases se obtiene de manera automática en el caso del DBSCAN, pero dicha cantidad dependerá a su vez del valor de los parámetros de distancia **ε** y puntos mínimos **p**.

Sin embargo, una desventaja frente a K Medios, es que DBSCAN requiere más cuidado para determinar los mejores valores para dichos parámetros, además de que suele no ofrecer buenos resultados de agrupamiento cuando la cantidad de datos es poca o, en efecto, no son densos en el espacio en donde estos viven.

Algoritmo DBSCAN

(Density Based Spatial Clustering of Applications with Noise)

Otra desventaja, como lo veremos en la implementación práctica, es que DBSCAN puede generar agrupamientos para todos los datos presentes en un conjunto, pero no puede por sí mismo predecir a qué clase pertenece un dato nuevo no usando para entrenamiento. Para ello, hay que aplicar clasificación con otros algoritmos para realizar esa tarea.

Décimo Séptimo Notebook Práctico

Realicemos entonces la implementación práctica del algoritmo de DBSCAN a fin de compararlo con el de K Medios y entender sus ventajas y desventajas.

Décimo Séptimo Notebook Práctico