Laboratorio de Datos
Clustering: DBSCAN

Primer Cuatrimestre 2024
Turnos tarde y noche

Facultad de Ciencias Exactas y Naturales, UBA

Repaso

Aprendizaje no supervisado

- El objetivo es encontrar patrones o estructuras ocultas en los datos.
- No conocemos o no hay a priori una respuesta correcta.
- Ejemplos: agrupación (clustering), reducción de dimensionalidad.

Clustering

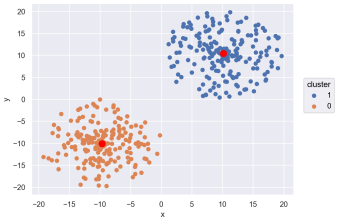
- El objetivo es agrupar datos similares en conjuntos llamados clústeres.
- Aplicaciones: segmentación de mercado, análisis de redes sociales, imagenes médica, etc.

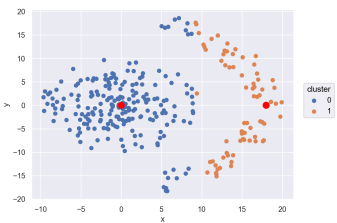
Repaso: k-medias

- K-medias es un algoritmo de agrupación que particiona los datos en k clústeres.
- Se definen k centros para los clusters.
- Se supone que cada cluster es un conjunto de datos que están razonablemente bien aproximados por el centro del cluster (el promedio de los valores del cluster).

Requerimiento importante: los clusters deben ser esféricos e isotrópicos (el mismo radio en todas las direcciones).

K-medias no funciona bien si los clusters tienen formas irregulares.

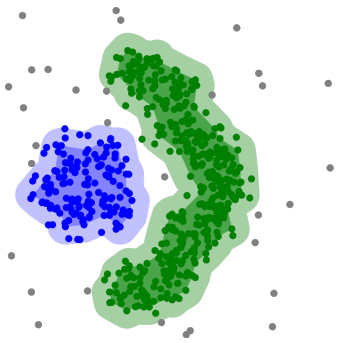




DBSCAN

DBSCAN es “Density-based spatial clustering of applications with noise” (lo importante es “agrupamiento espacial basado en densidad“)

- Dado un conjunto de puntos en algún espacio, se agrupan puntos que están muy juntos (puntos con muchos puntos vecinos).
- Se marcan como valores atípicos (outliers) puntos que se encuentran solos en regiones de baja densidad.



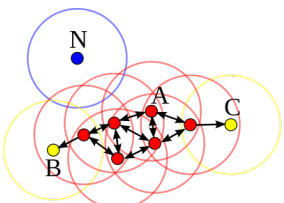
DBSCAN

DBSCAN depende de dos parámetros:

- Un radio ε (epsilon)
- Un valor $minPts$ que indica cuandos puntos se espera encontrar en la vecindad de un punto de un cluster.

Utilizando esos parámetros, los datos se clasifican en

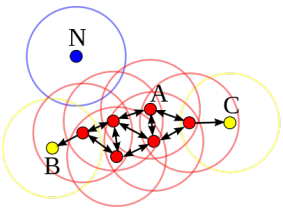
1. Puntos *centrales* p.
2. Puntos *directamente alcanzables* q desde un punto central p.
3. Punto q *alcanzables* desde un punto central p.
4. Puntos *atípicos*, todos los puntos que no son alcanzables desde ningún punto central.



DBSCAN - Definiciones

Concretamente, a partir de ε y $minPts$ definimos:

1. Punto *central* p, si al menos $minPts$ puntos están a una distancia ε de p (incluido p).
2. Punto *directamente alcanzable* q desde un punto central p, si q está a distancia menor o igual que ε de un punto central p.
3. Punto q *alcanzable* desde p si hay un camino $p_1, \ldots, p_n$ con $p_1 = p$ y $p_n = q$, donde cada p_{i+1} es directamente alcanzable desde p_i.
4. Puntos *atípicos*, todos los puntos que no son alcanzables desde ningún punto central.

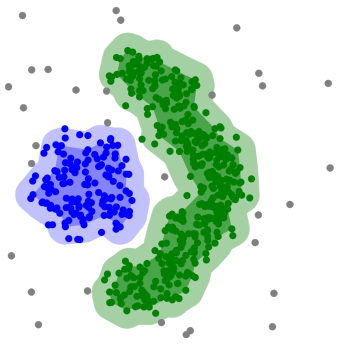


Puntos centrales y puntos bordes

Cada cluster queda compuesto por

- puntos centrales
- puntos alcanzables desde un punto central que no son centrales, estos puntos se denominan puntos frontera.

En el gráfico vemos la región central (puntos del plano que tienen al menos $minPts$ vecinos) y la región borde (puntos del plano con menos de $minPts$ vecinos alcanzables desde un punto central).



Algoritmo para construir un cluster

El cluster asociado a un punto central p está constituido por todos los puntos alcanzables desde p.

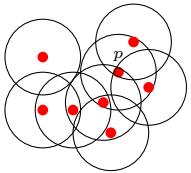
Obtenemos el siguiente algoritmo para construir un cluster a partir de un punto central:

1. Identificamos un punto central p (que no pertenezca a ningún cluster ya construido).
2. Agregamos al cluster a todos los puntos q directamente alcanzables desde p.
3. Si agregamos puntos centrales nuevos al cluster, agregamos al cluster a todos los puntos alcanzables desde los nuevos puntos centrales.
4. Repetimos el Paso 3 hasta que no hayamos agregado ningún nuevo punto central al cluster.

Ejemplo: construir el cluster asociado al punto central p

Tomamos $\varepsilon = 1$ (radios utilizados en el gráfico) y minPts $= 4$.

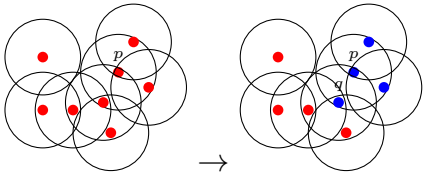
1. p es un punto central



Ejemplo: construir el cluster asociado al punto central p

Tomamos $\varepsilon = 1$ (radios utilizados en el gráfico) y minPts $= 4$.

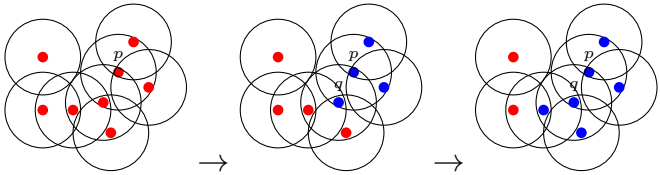
1. p es un punto central
2. Añadimos al cluster los 3 puntos directamente alcanzables desde p.
3. De estos 3 puntos, solo q es un punto central.



Ejemplo: construir el cluster asociado al punto central p

Tomamos $\varepsilon = 1$ (radios utilizados en el gráfico) y minPts $= 4$.

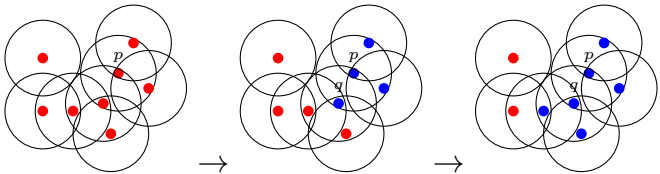
1. p es un punto central
2. Añadimos al cluster los 3 puntos directamente alcanzables desde p.
3. De estos 3 puntos, solo q es un punto central.
4. Añadimos al cluster los 2 puntos alcanzables desde ese nuevo punto central.



Ejemplo: construir el cluster asociado al punto central p

Tomamos $\varepsilon = 1$ (radios utilizados en el gráfico) y minPts $= 4$.

1. p es un punto central
2. Añadimos al cluster los 3 puntos directamente alcanzables desde p.
3. De estos 3 puntos, solo q es un punto central.
4. Añadimos al cluster los 2 puntos alcanzables desde ese nuevo punto central.
5. No añadimos ningún nuevo punto central, por lo tanto finalizamos.



Algoritmo DBSCAN

Comenzamos con un conjunto de puntos sin etiquetar.

1. Seleccionamos un punto central P no etiquetado y le asignamos un nuevo cluster.
2. Construimos el cluster asociado a P como vimos antes y les asignamos a todos los puntos la etiqueta correspondiente al cluster.
3. Repetimos los puntos 1 y 2 hasta que no haya ningún punto central no etiquetado.
4. Todos los puntos que quedaron sin etiquetar, los etiquetamos como valores atípicos.

Nota: los puntos frontera pueden quedar asignados a clusters distintos según el orden en que etiquetamos, pero los puntos centrales en cada cluster son siempre los mismos.

Comparación de K-means y DBSCAN

K-means

Basado en particiones

Supuestos:

- Los datos son esféricos o isotrópicos
- Los clusters tienen tamaños similares

Fortalezas:

- Simple y rápido
- Funciona bien con grandes conjuntos de datos

Debilidades:

- Requiere pre-especificar K
- Sensible a la inicialización
- Desempeño pobre en presencia de outliers

DBSCAN

Basado en densidad

Supuestos:

- Los clusters son regiones densas en el espacio de datos

Fortalezas:

- Puede encontrar clusters de formas arbitrarias
- No es necesario especificar el número de clusters
- Robusto a valores atípicos (outliers)

Debilidades:

- Requiere establecer ε y minPts
- Desempeño pobre con densidad variable