Clustering

Densidades

Christian Oliva Moya

Dpto. de Ingeniería Informática, Escuela Politécnica Superior

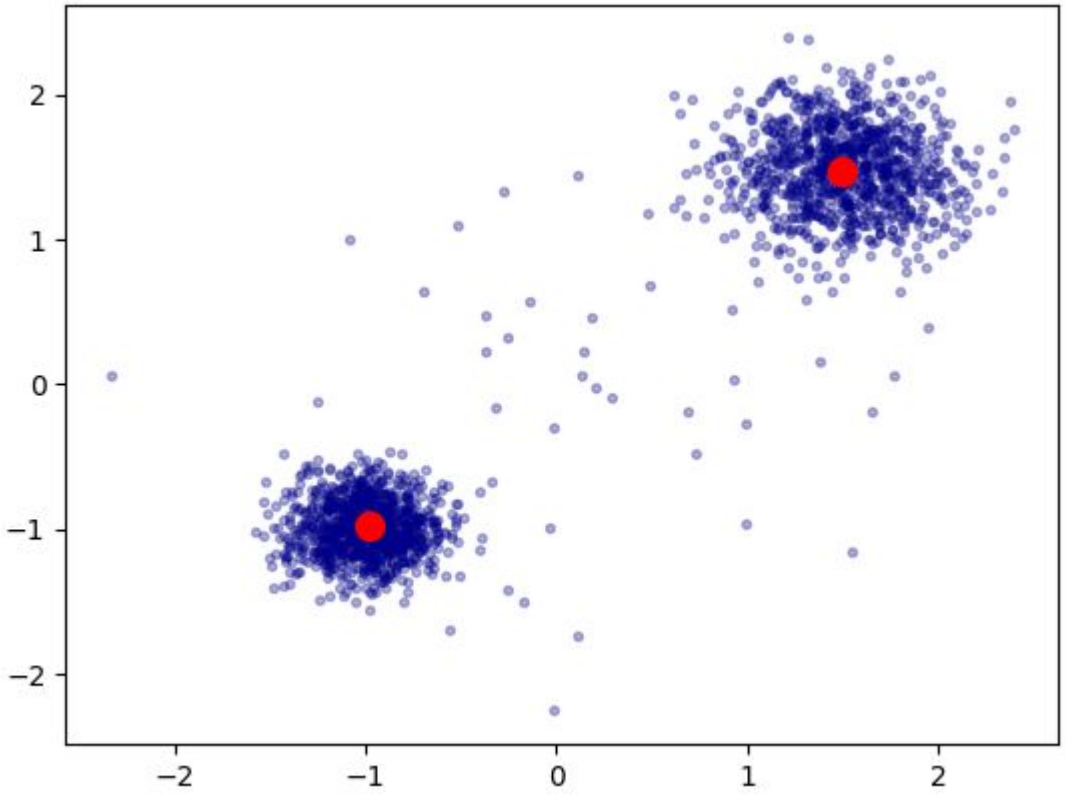
Universidad Autónoma de Madrid

28049 Madrid, Spain

v1 - 28/02/2024

¿Qué es clustering? (1)

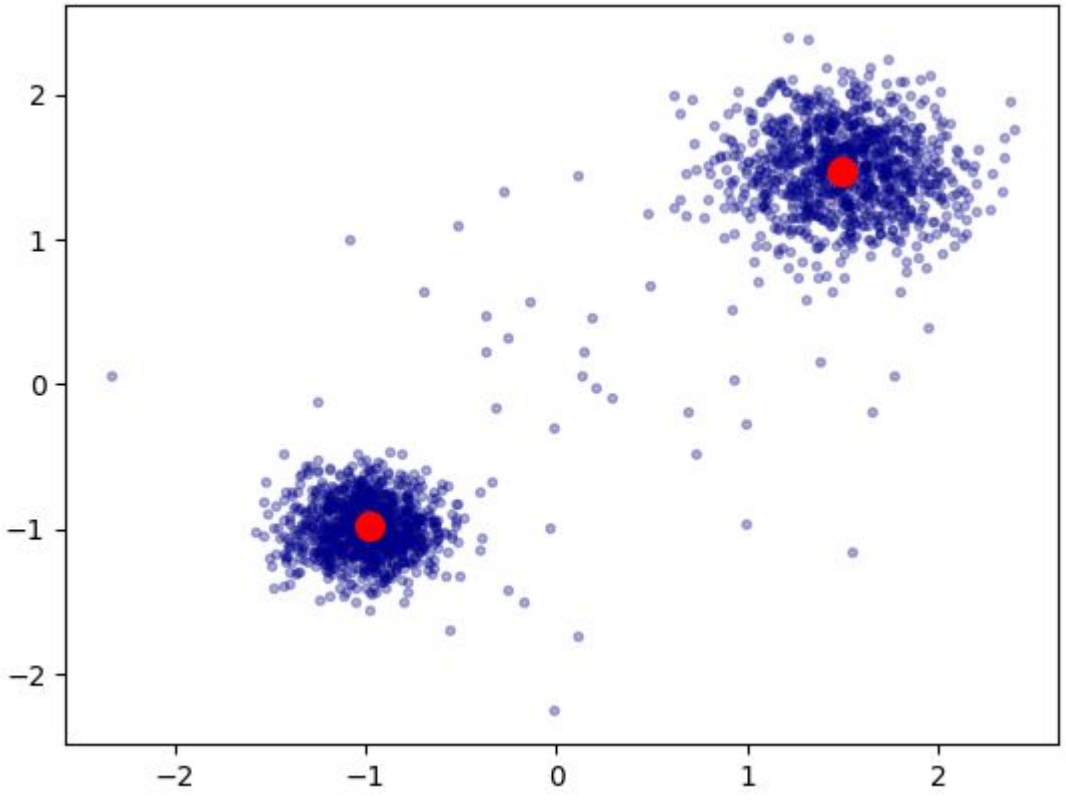
- ¿Esto es clustering?



¿Qué es clustering? (2)

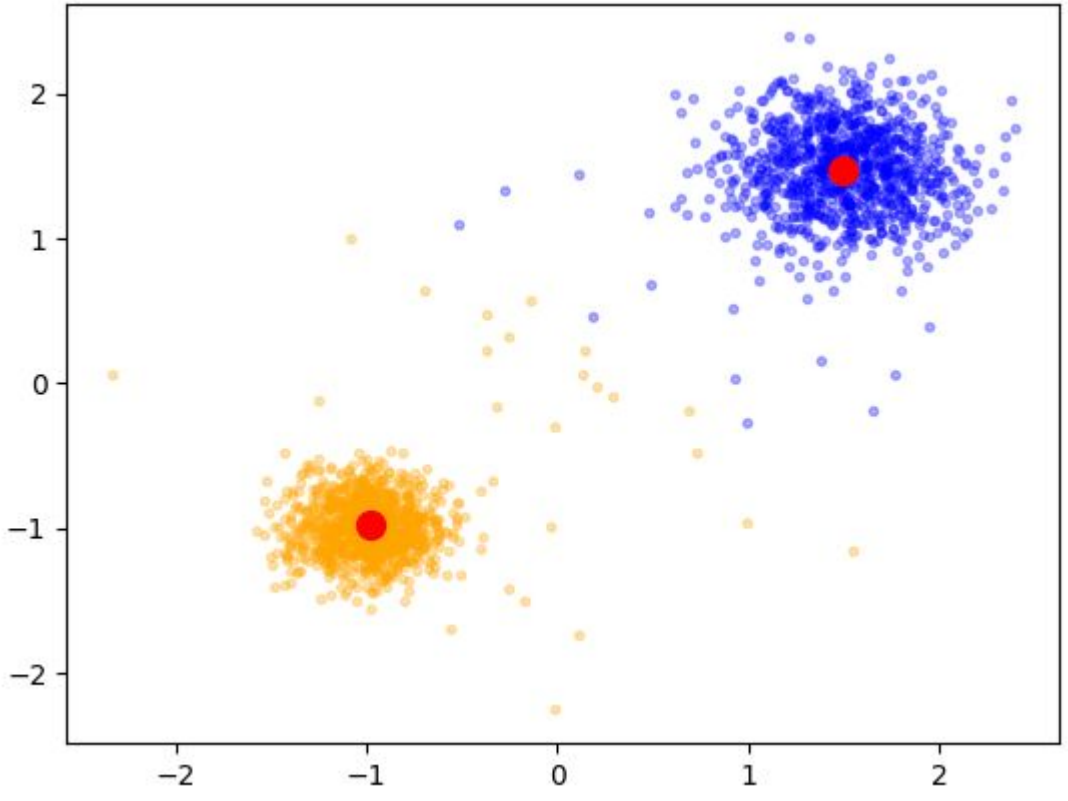
- ¿Esto es clustering?

  Sí, ya que encuentro dos grupos



¿Qué es clustering? (3)

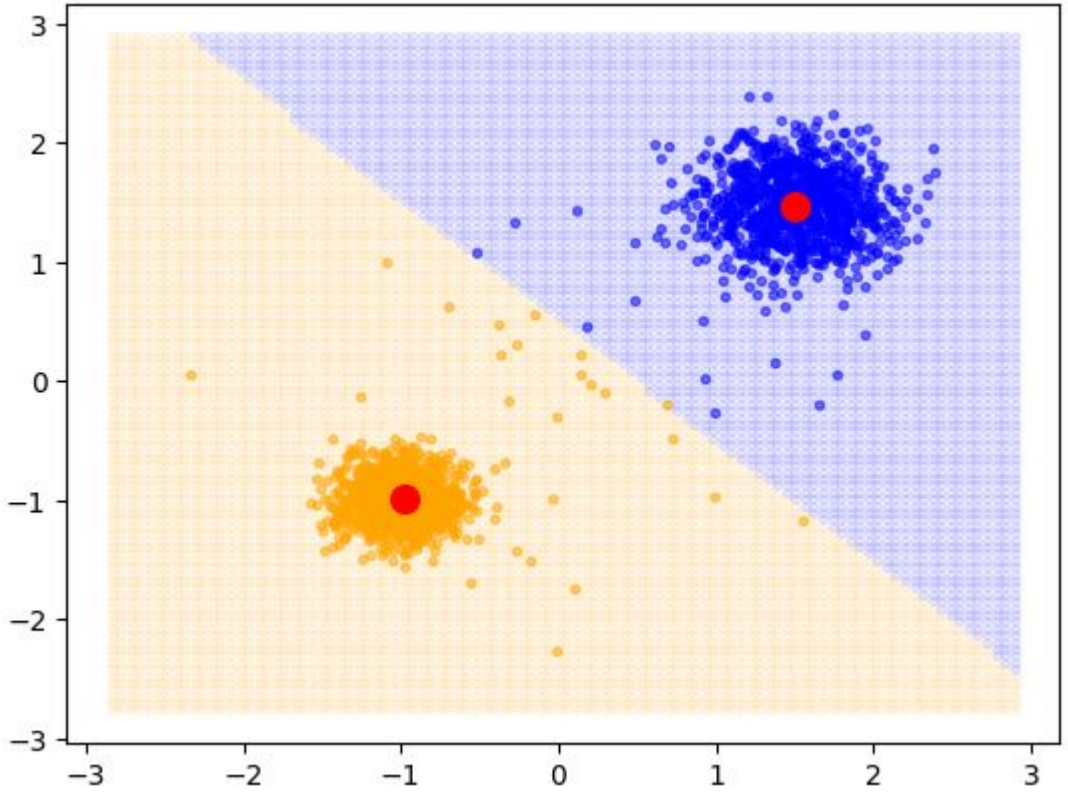
- Aplicando K-Means (K=2)
- ¿Estoy haciendo clustering?



¿Qué es clustering? (4)

- Aplicando K-Means (K=2)
- ¿Estoy haciendo clustering?

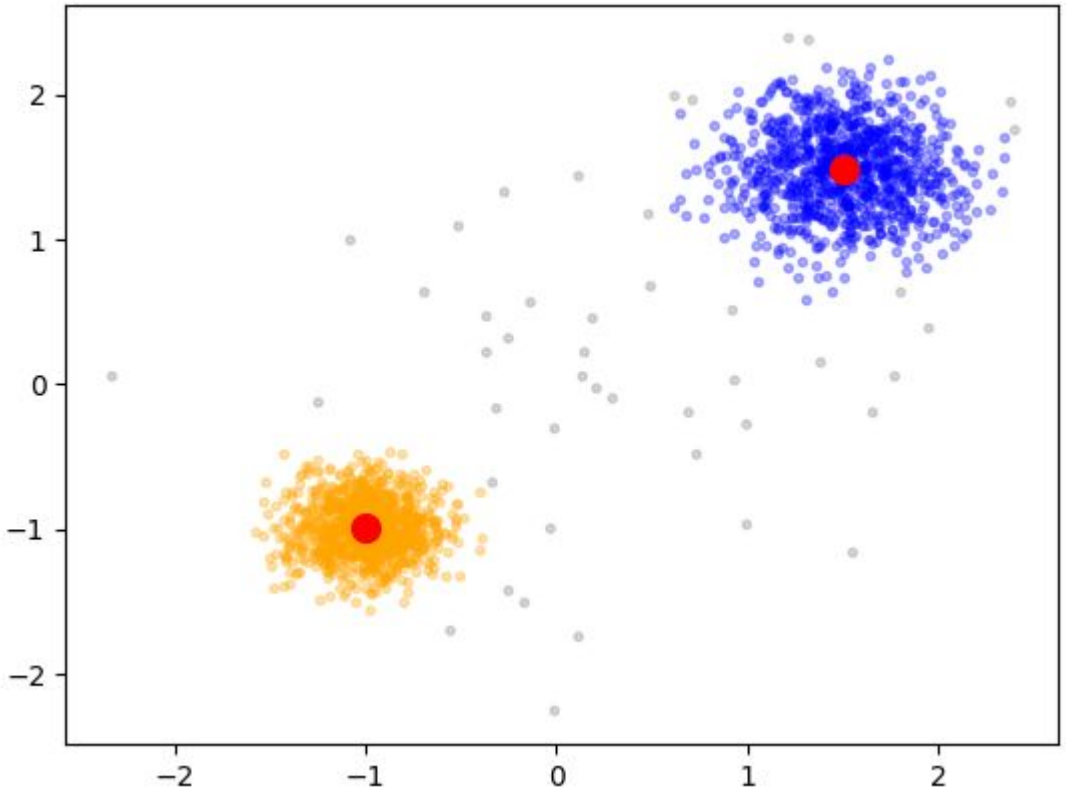
  Depende



¿Qué es clustering? (5)

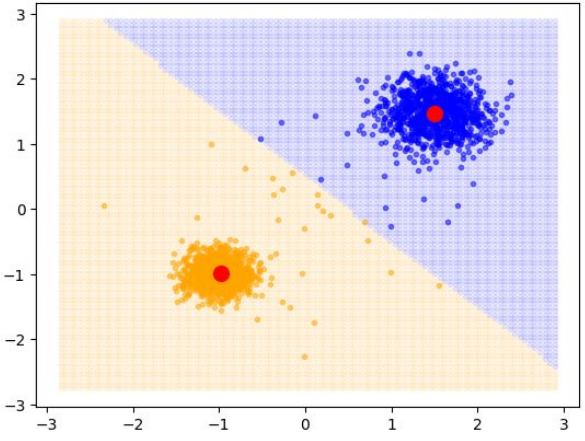
- ¿Y ahora?
- ¿Estoy haciendo clustering?

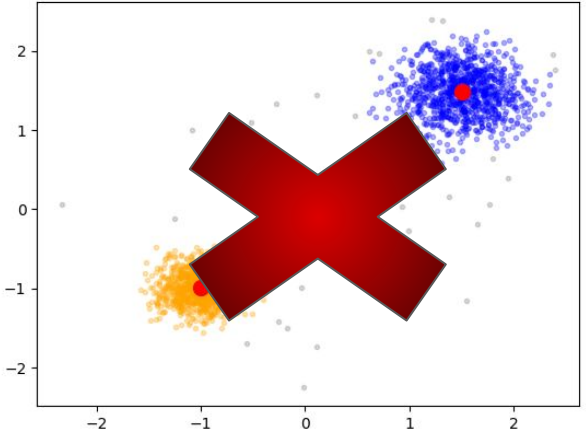
  Depende



¿Qué es clustering? (6)

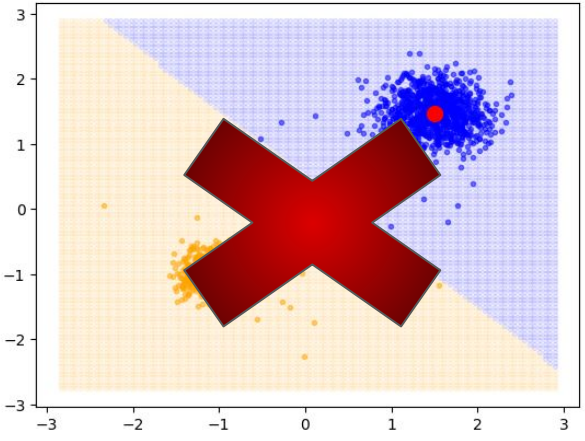
- Si es una colección de patrones basados en una similitud
- Si es una región en el espacio con una alta densidad de puntos

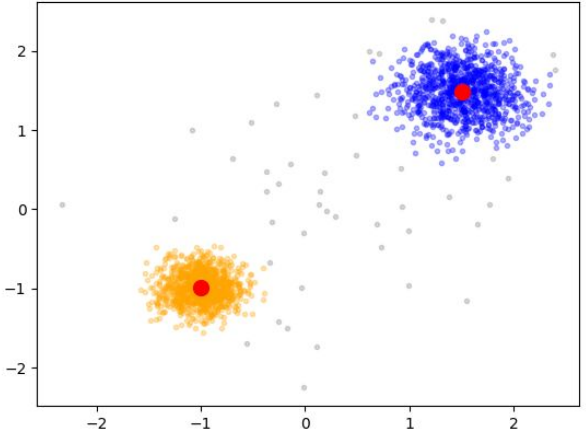




¿Qué es clustering? (7)

- Si es una colección de patrones basados en una similitud
- Si es una región en el espacio con una alta densidad de puntos





Particional vs Densidad

- Los algoritmos particionales (K-Means, EM) dividen el espacio en regiones separadas
    - Definiendo K clusters
    - Definiendo una métrica de similitud

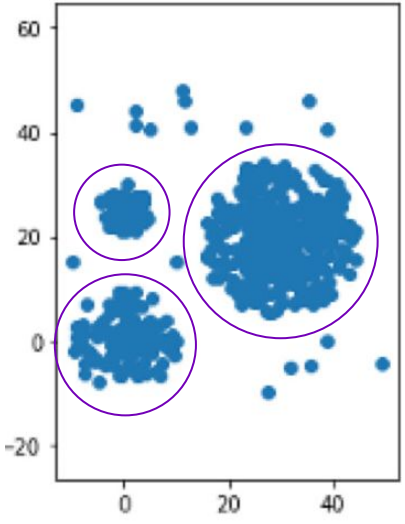
  Asumen que todos los puntos pertenecen al cluster más cercano

- Los algoritmos basados en densidades (DBSCAN) identifica zonas muy pobladas
    - Definiendo una métrica de similitud

  Puede haber puntos que no pertenecen a ningún cluster

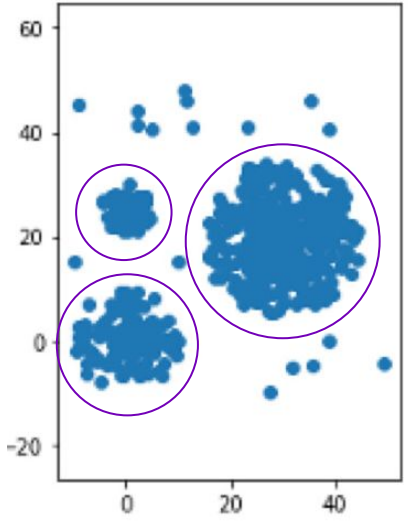
DBSCAN (1)

- DBSCAN surge de un método intuitivo de agrupamiento humano
- Al mirar la figura, cualquiera identifica fácilmente tres grupos
- ¿Qué estamos observando?
  - La densidad de puntos, es decir, cuán agrupados están los puntos
  - Parece una aproximación más realista del concepto de clustering



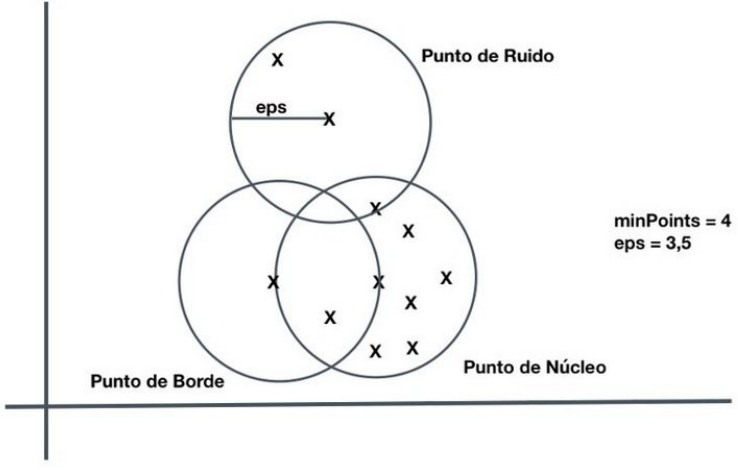
DBSCAN (2)

- DBSCAN no necesita definir K, en contraposición necesitamos definir:
  - Épsilon (eps): especifica lo cerca que deben estar los puntos entre sí para ser considerados un único cluster. Si la distancia entre dos puntos es menor que eps, esos dos puntos son vecinos.
  - Puntos mínimos (minPts): el número de puntos necesarios para formar una región densa.



DBSCAN (3)

- Con los valores de eps y minPts, tenemos tres tipologías de puntos:
  - Puntos de núcleo: Tiene más de un número especificado de puntos minPts en su radio (eps).
  - Puntos de borde: Tiene menos de un número especificado de puntos minPts pero es vecino de un punto de núcleo.
  - Puntos de ruido: Todos los demás.

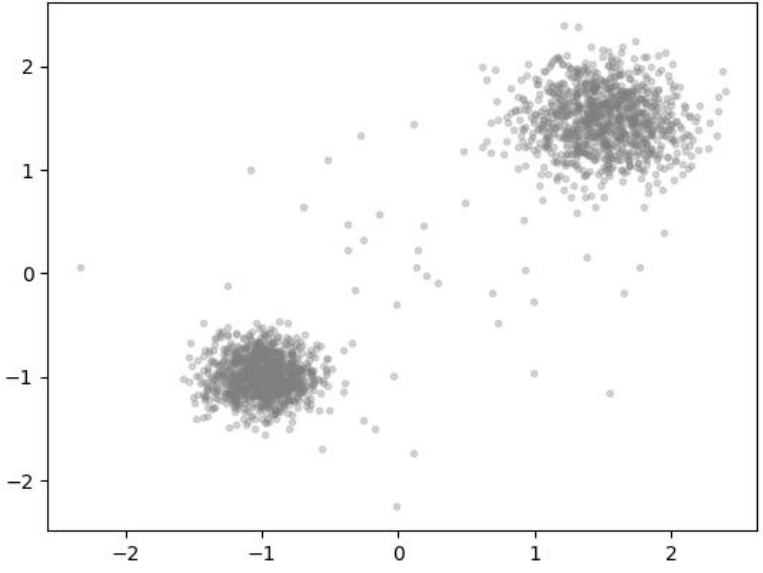


DBSCAN (4)

- Algoritmo DBSCAN:
  - Selecciona un punto arbitrario que no haya sido previamente visitado. Se calculan las distancias a todos y la información de su vecindario se recupera según el valor de eps.
  - Si es punto de núcleo (contiene minPts o más puntos en el vecindario):
    - Todos los puntos del vecindario forman un cluster, además de todos los vecinos de aquellos vecinos que son también puntos de núcleo.
  - Si no:
    - Se etiqueta como punto de ruido. Este punto podrá ser modificado más adelante si se encuentra en el vecindario de otro punto de núcleo (era punto de borde).

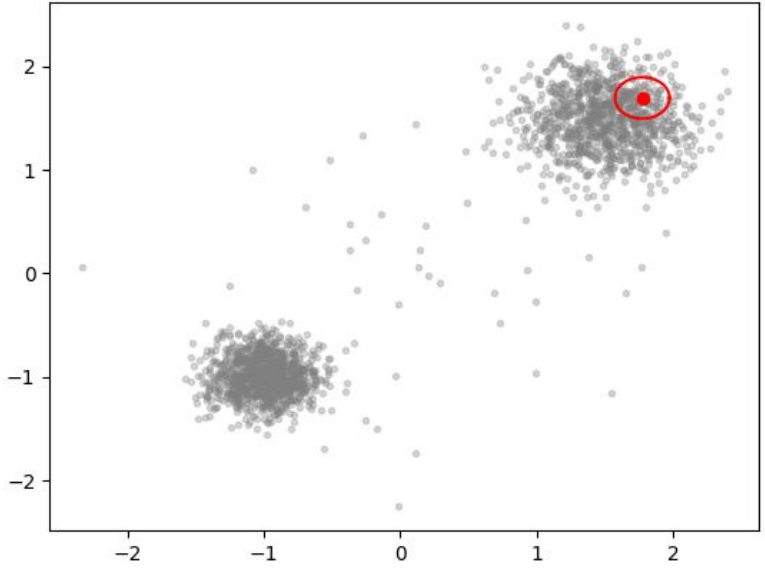
DBSCAN (5) - Ejemplo de ejecución

- Ejemplo de ejecución:
  - eps = 0.2
  - minPts = 10



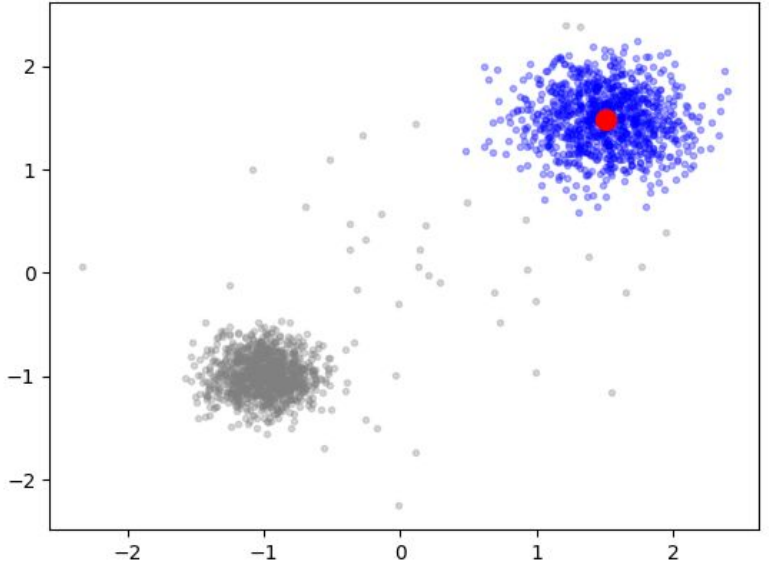
DBSCAN (5) - Ejemplo de ejecución

- Selecciona un punto al azar y traza el círculo con radio eps:
  - eps = 0.2
  - minPts = 10



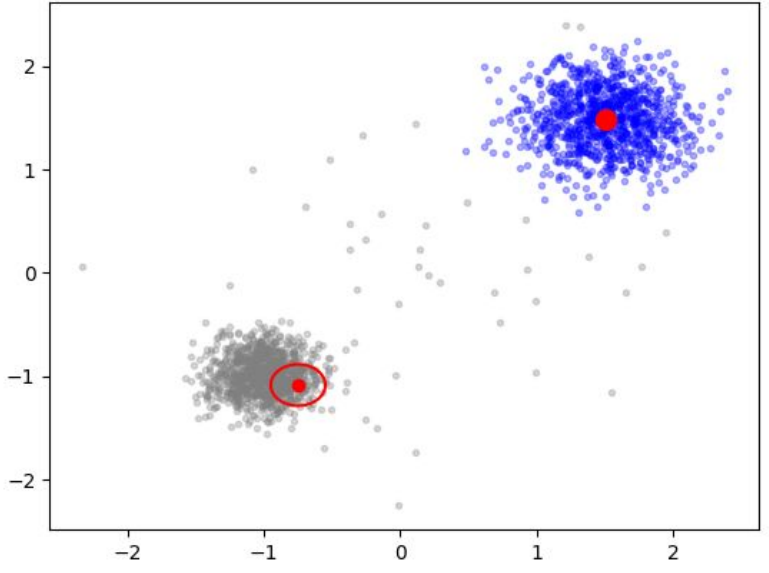
DBSCAN (5) - Ejemplo de ejecución

- Como es un punto de núcleo, es decir, hay más de minPts dentro del radio → De forma recursiva, todos los vecinos forman un cluster
  - eps = 0.2
  - minPts = 10



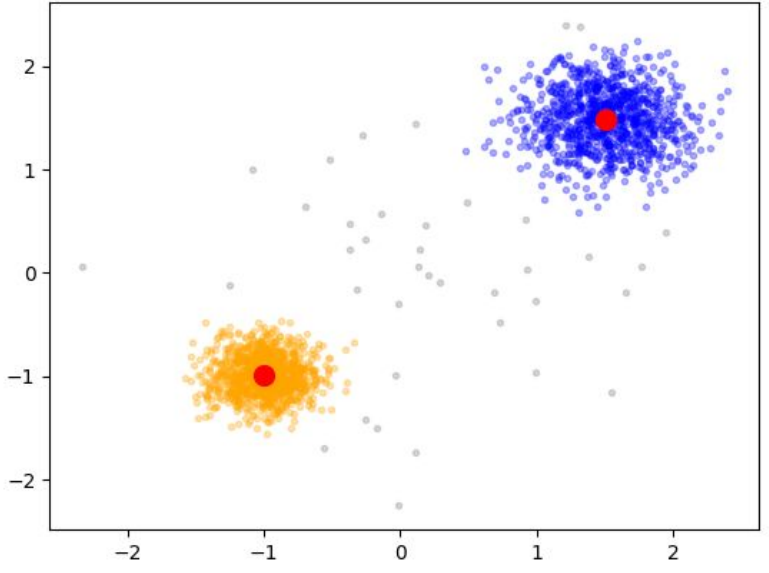
DBSCAN (5) - Ejemplo de ejecución

- Selecciona un punto al azar y traza el círculo con radio eps:
  - eps = 0.2
  - minPts = 10



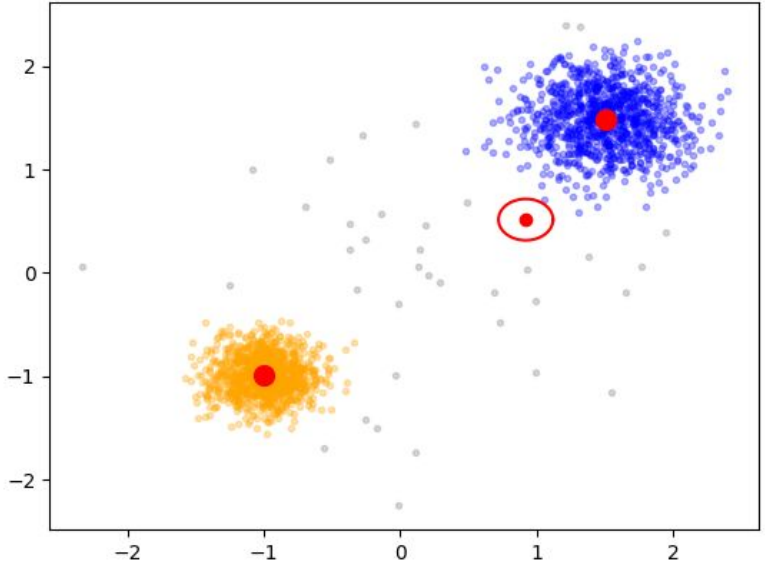
DBSCAN (5) - Ejemplo de ejecución

- Como es un punto de núcleo, es decir, hay más de minPts dentro del radio → De forma recursiva, todos los vecinos forman un cluster
  - eps = 0.2
  - minPts = 10



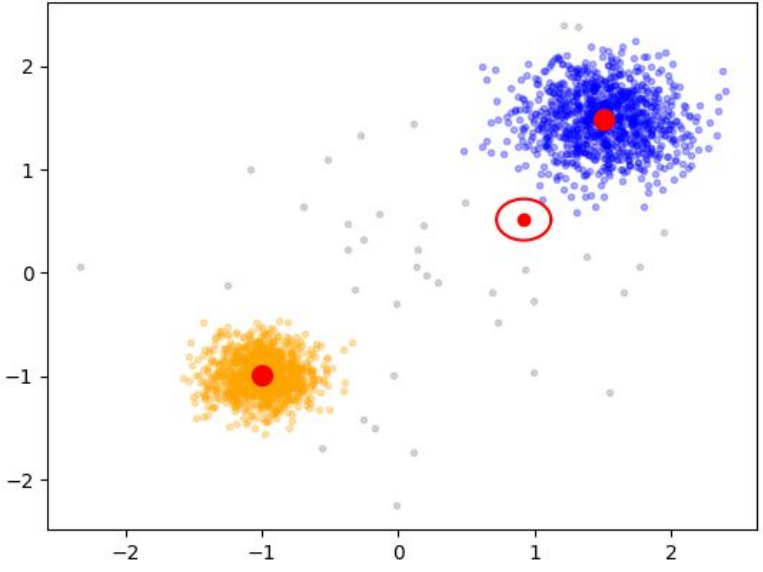
DBSCAN (5) - Ejemplo de ejecución

- Selecciona un punto al azar y traza el círculo con radio eps:
  - eps = 0.2
  - minPts = 10



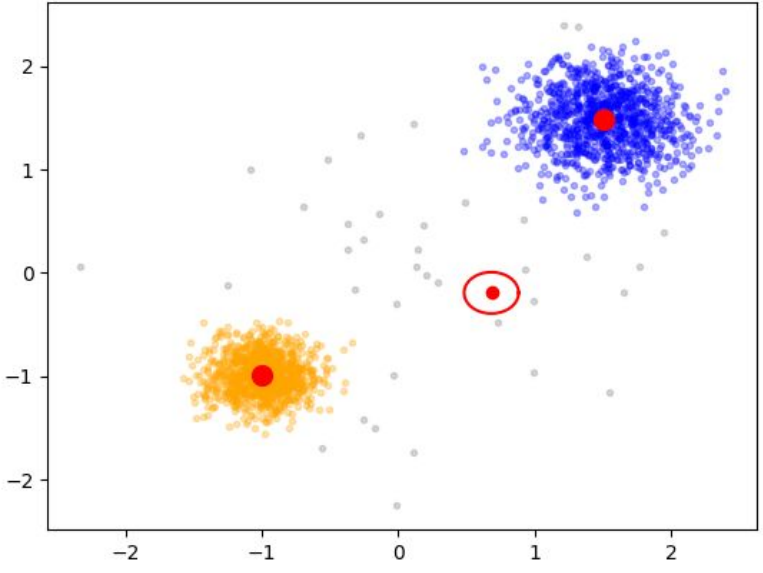
DBSCAN (5) - Ejemplo de ejecución

- Como es un punto de ruido, es decir, hay menos de minPts dentro del radio, no hago nada
  - eps = 0.2
  - minPts = 10



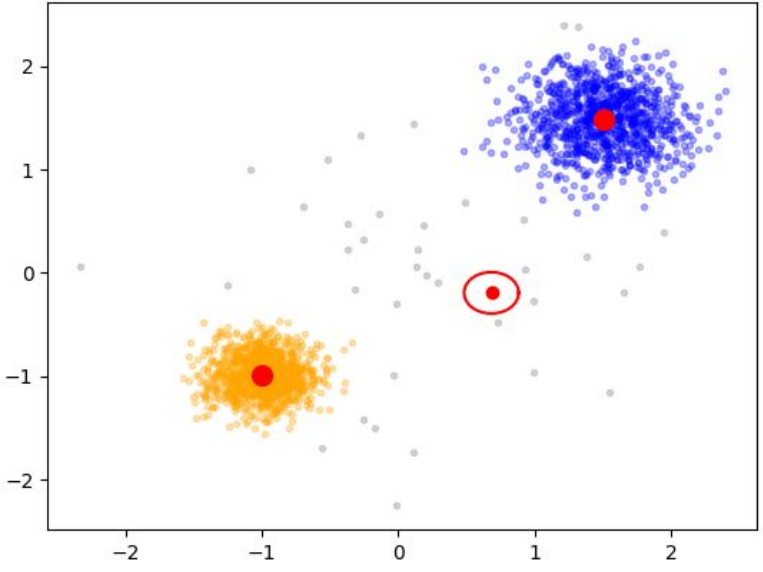
DBSCAN (5) - Ejemplo de ejecución

- Selecciona un punto al azar y traza el círculo con radio eps:
  - eps = 0.2
  - minPts = 10



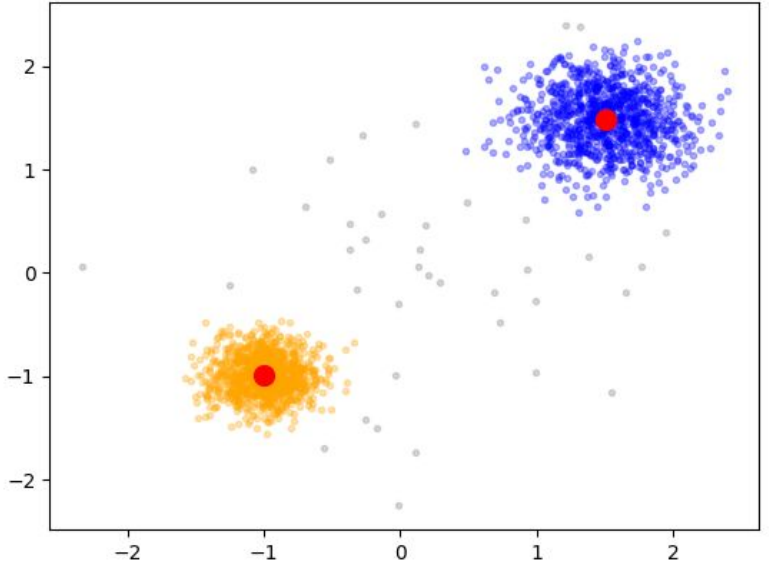
DBSCAN (5) - Ejemplo de ejecución

- Como es un punto de ruido, es decir, hay menos de minPts dentro del radio, no hago nada
  - eps = 0.2
  - minPts = 10



DBSCAN (5) - Ejemplo de ejecución

- Así sucesivamente hasta que todos los puntos sin color han sido visitados.
- Se acaba el algoritmo y tenemos 2 clusters



DBSCAN (6)

- Ventajas:
  - No requiere el valor K
  - Puede encontrar cualquier forma de cluster
  - Discrimina muy bien el ruido y los valores atípicos
  - Es excelente para separar clusters de alta densidad frente a otros
  - Es visualmente atractivo e intuitivo
- Inconvenientes:
  - Es muy sensible a los hiperparámetros eps y minPts

DBSCAN (7)

- Notebook *05_dbscan.ipynb*