

Definición de cluster: Cluster, o grupo, es un conjunto de objetos que son "similares" entre ellos y "diferentes" de los objetos que pertenecen a los otros grupos. La palabra "cluster" viene del inglés y significa agrupación. Desde un punto de vista general, el cluster puede considerarse como la búsqueda automática de una estructura o de una clasificación en una colección de datos "no etiquetados". Este tipo de análisis se aplica cuando se desconocen las etiquetas asociadas a los datos.

Clustering

Clustering es el proceso de agrupar datos en clases o clusters de tal forma que los objetos de un cluster tengan una similaridad alta entre ellos, y baja (sean muy diferentes) con objetos de otros clusters.

En Machine Learning, clustering es un ejemplo de aprendizaje no supervisado.

A diferencia de la clasificación, el clustering o aprendizaje no supervisado no requiere clases predefinidas. Por esta razón, el clustering es un ejemplo de aprendizaje por observación, mientras que clasificación es un ejemplo de aprendizaje por ejemplos.

Características:

1. Escalabilidad: normalmente corren con pocos datos.
2. Clusters de formas arbitrarias: los basados en distancias numéricas tienden a encontrar cluster esféricos.
3. Capacidad de manejar diferentes tipos de atributos: numéricos (lo más común), binarios, nominales, ordinales, etc.
4. Capacidad de añadir restricciones.
5. Manejo de ruido: muchos son sensibles a datos erróneos.
6. Poder funcionar eficientemente con alta dimensionalidad.
7. Requerimientos mínimos para especificar parámetros, como el número de clusters.
8. Independientes del orden de los datos.
9. Que los clusters sean interpretables y utilizables

Aplicabilidad: El objetivo del clustering es identificar la clasificación intrínseca de un conjunto de datos no etiquetados. Los algoritmos de clasificación de datos tienen numerosas aplicaciones en distintos ámbitos: Biología, Comercialización, Seguros, Sismología, donde una aplicación de interés es el estudio de los seísmos. La reagrupación de los epicentros de los seísmos observados permite determinar las zonas de riesgos, y poder ayudar a evitar catástrofes.