Resumen de lo aprendido 2

Aprendizaje No Supervisado y Clustering: K-means

Aprendizaje No Supervisado

El aprendizaje no supervisado es un tipo de aprendizaje automático donde el modelo se entrena en un conjunto de datos que no tiene etiquetas de salida. A diferencia del aprendizaje supervisado, que aprende a partir de un conjunto de datos etiquetado para hacer predicciones o decisiones sin intervención humana, los algoritmos de aprendizaje no supervisado se usan para encontrar patrones o relaciones en los datos.

Aplicaciones Comunes

- Agrupación o Clustering
- Reducción de Dimensionalidad
- Reglas de Asociación

Clustering

El clustering es una técnica que tiene el objetivo de dividir un conjunto de datos en grupos que contengan elementos similares. No utiliza ninguna etiqueta, y se basa en la estructura de los datos para realizar la agrupación.

Métricas Comunes

- Distancia Euclidiana
- Similitud Coseno
- Distancia de Manhattan

K-means

K-means es uno de los algoritmos de clustering más utilizados. Es simple y eficiente, aunque también tiene sus desafíos y limitaciones.

Cómo Funciona

1. **Inicialización**: Se seleccionan K puntos como centroides iniciales.
2. **Asignación**: Cada punto se asigna al centroide más cercano.
3. **Actualización**: Los centroides se recalculan como el promedio de todos los puntos asignados a ese cluster.

4. **Iteración**: Los pasos 2 y 3 se repiten hasta que los centroides ya no cambien significativamente.

Parámetros Importantes

- **K**: Número de Clústeres. Un valor incorrecto de K puede dar resultados poco útiles.
- **Función de Distancia**: Generalmente es la distancia euclidiana, pero puede variar.
- **Condiciones Iniciales**: Los centroides iniciales pueden afectar el resultado final.

Métodos de Validación

- **Método del Codo**: Ayuda a encontrar el valor óptimo de K.
- **Coeficiente de Silueta**: Mide cuán similares son los objetos en el mismo cluster frente a otros clusters.

Este resumen proporciona una visión general de lo que es el aprendizaje no supervisado y cómo se utiliza el algoritmo K-means para el clustering. La técnica tiene una amplia gama de aplicaciones, desde la segmentación del cliente hasta el análisis de genes.

Otros Algoritmos de Clustering en Aprendizaje No Supervisado

Además de K-means, existen varios otros algoritmos de clustering en aprendizaje no supervisado, cada uno con sus propias ventajas, desventajas y casos de uso. A continuación, se presentan algunos de ellos:

1. Hierarchical Clustering (Clustering Jerárquico)

Este método crea una jerarquía o un árbol de clusters. Hay dos enfoques principales:

- **Aglomerativo**: Comienza con cada punto como su propio cluster y fusiona los más cercanos iterativamente.
- **Divisivo**: Comienza con un único cluster que contiene todos los puntos y lo divide iterativamente.

Ventajas

- No es necesario especificar el número de clusters a priori.
- Facilita la visualización y la interpretación mediante dendrogramas.

Desventajas

- Costoso en términos de tiempo de cálculo para grandes conjuntos de datos.

3. Gaussian Mixture Models (Modelos de Mezclas Gaussianas)

Es una generalización del algoritmo K-means que incluye información sobre la estructura de covarianza de los datos, así como los centros de los clusters gaussianos latentes.

Ventajas

- Proporciona una "suavidad" en la asignación de puntos a clusters.
- Funciona bien para clusters elípticos.

Desventajas

- Sensible a la inicialización.
- Requiere más parámetros.

2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN agrupa puntos cercanos en función de su densidad, y puede encontrar cualquier forma de cluster, a diferencia de K-means que solo encuentra clusters convexos.

Ventajas

- No requiere el número de clusters a priori.
- Puede encontrar clusters de forma arbitraria.
- Capaz de manejar "ruido" y puntos "anómalos".

Desventajas

- No funciona bien cuando los clusters tienen diferentes densidades.

Existen muchos algoritmos pero solamente vamos a revisar algunos de ellos.