Grupo 4
Integrantes:
Sebastián Romero
Juan Calderón
Kristian Mendoza

Proyecto 8: Clustering

Expectation-Maximization (EM)

El algoritmo de esperanza-maximización (EM) es una herramienta útil cuando necesitamos estimar parámetros desconocidos en un modelo estadístico, como la media o la varianza. A menudo, estos modelos implican variables que no podemos observar directamente, pero que influyen en nuestros cálculos.

Puedes imaginar el proceso de EM como resolver un rompecabezas con piezas faltantes. Aunque no podemos ver todas las piezas, podemos hacer suposiciones sobre cómo deberían encajar basándonos en las que sí podemos ver. EM funciona de manera similar. Comenzamos con suposiciones sobre los valores desconocidos, usamos esas suposiciones para hacer predicciones sobre lo que no podemos ver, ajustamos nuestras suposiciones en función de esas predicciones, y repetimos este proceso hasta que nuestras predicciones y suposiciones coincidan lo mejor posible.

El truco radica en que, aunque no podemos ver todas las piezas del rompecabezas, podemos inferirlas a partir de las que sí podemos ver. Eventualmente, EM nos lleva a una solución que se ajusta lo mejor posible a los datos que conocemos. Sin embargo, es importante tener en cuenta que a veces podemos obtener soluciones que no tienen sentido en el contexto del problema que estamos abordando.

Clustering Using REpresentatives (CURE):

CURE (Clustering Using REpresentatives) es un algoritmo eficiente para agrupar datos en grandes bases de datos. A diferencia del método K-means, es más resistente a valores atípicos y capaz de identificar grupos con formas irregulares y tamaños variables.

Para lidiar con la variabilidad en la forma y tamaño de los grupos, CURE utiliza un enfoque jerárquico que combina elementos de los métodos basados en centroides y aquellos que consideran todos los puntos. Selecciona un número constante de puntos bien distribuidos dentro de cada grupo y los acerca hacia el centroide del grupo. Estos puntos reducidos se utilizan como representantes del grupo. Los grupos se fusionan en cada paso del proceso de agrupamiento jerárquico basándose en la cercanía entre los representantes.

El costo computacional de CURE es significativo (O(n^2 log n)) y su complejidad espacial es lineal (O(n)). Sin embargo, su implementación directa en bases de datos grandes puede ser problemática debido a esta complejidad.

Para abordar este desafío, se han desarrollado mejoras:

- Muestreo aleatorio: Se utiliza para manejar grandes volúmenes de datos, balanceando precisión y eficiencia.
- Particionamiento: Se divide el espacio de muestra en particiones más pequeñas, lo que facilita la reducción de tiempos de ejecución.
- Etiquetado de datos en disco: Para asignar datos a grupos, se seleccionan representantes aleatorios y se asignan datos basados en su cercanía a estos representantes.
- Estas mejoras permiten que CURE sea más escalable y aplicable a conjuntos de datos más grandes, superando así las limitaciones de su enfoque original.

T-distributed stochastic neighbor embedding TSNE

El t-distributed stochastic neighbor embedding (t-SNE) es un método estadístico para visualizar datos de alta dimensionalidad asignando a cada punto de datos una ubicación en un mapa de dos o tres dimensiones.
El algoritmo t-SNE consta de dos etapas principales. Primero, construye una distribución de probabilidad sobre pares de objetos de alta dimensionalidad de manera que objetos similares sean asignados a una probabilidad más alta mientras que puntos disímiles son asignados a una probabilidad más baja. Segundo, define una distribución de probabilidad similar sobre los puntos en el mapa de baja dimensionalidad, y minimiza la divergencia Kullback-Leibler (KL divergence) entre las dos distribuciones con respecto a las ubicaciones de los puntos en el mapa. Mientras que el algoritmo original utiliza la distancia euclidiana entre objetos como base de su métrica de similitud, esto puede cambiarse según sea necesario.

Aunque los gráficos de t-SNE a menudo parecen mostrar agrupamientos, estos agrupamientos visuales pueden ser influenciados fuertemente por la parametrización elegida y, por lo tanto, es necesario tener un buen entendimiento de los parámetros para t-SNE.

Para un conjunto de datos con n elementos, t-SNE se ejecuta en tiempo O(n^2) y requiere espacio O(n^2).

Bibliografía:

Wikipedia contributors. (2022, April 29). CURE algorithm. Wikipedia. https://en.wikipedia.org/wiki/CURE_algorithm

Wikipedia contributors. (2024, February 10). T-distributed stochastic neighbor embedding. https://en.wikipedia.org/wiki/T-distributed_stochastic_neighbor_embedding

Wikipedia contributors. (2024, March 23). Expectation–maximization algorithm. Wikipedia. https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm