Análisis de Grandes Volúmenes de Datos
Carlos Espinoza Peraza - B92786
Reporte 24 de mayo: Clustering

Resumen de los temas estudiados

Clustering

Esta es una técnica no supervisada que divide el conjunto de datos en grupos, los cuales siguen una esructura natural. Estos grupos pueden ser significativos, útiles, ambos o ninguno. Aportan valor al proceso de `exploración de datos`

```
- Métrica de similitud
- Codificar atributos categóricos
- Transformar atributos númericos
- Cantidad de clusters a realizar
```

> **Similitud:** Indica el grado de similitud que hay entre objetos (1 - similitud)

Tipos de Clustering

Particionamiento

- Segmentos mutualemente excluyentes
- Segmentos de forma esférica
- Basado en mediciones de distancia
- Puede usar el promedio (mean) o medoides (centro del cluster)
- **Útil en conjuntos pequeños**

> **K-Means, K-moids, K-Meoids**

Jerárquico

- Segmentación que tienen niveles (jerarquías)
- No puede corregir merges o splits incorrectos
- Incorpora técnicas de micro-clustering

> **BIRCH, Chamaleon**

Densidad

- Segementación con regiones densas, separadas por regiones de baja densidad
- **Densidad: cada objeto debe tener un número mínimo de vecinos**

- Encuetra objetos de formas arbitrarias
- Ignora outliers

> **DBSCAN, OPTICS, DENCLUE**

Rejilla

- Estructura tipo rejillas
- Rápido procesamiento
- Independiente del número de puntos
- Depende del tamaño de la rejilla

> **STING, CLIQUE**

Comentarios sobre la materia estudiada

Clustering es una técnica poderosa cuando se requiere tener un entendimiento más profundo de los datos sobre los que estamos trabajando. Permite hacer agrupaciones para encontrar similitudes o categorías dentro del conjunto. Al ser una técnica no supervisada, sus resultados pueden ser variables.

Dudas sobre la materia estudiada

> Ninguna

Posible uso como profesional

Usar clusters como mecanismo para analizar las tendencias o para encontrar patrones en datos que a simple vista no son fáciles de categorizar. Conocer las posibles clases en las que se puedan agrupar los datos, con esto se puede tener una mejor idea de que algoritmo de ML podemos utilizar dependiendo de las dimensiones de las clases encontradas.

Problemas que podría resolver con las técnicas estudiadas

Analizar un registro de ventas para intentar determinar que conjunto de productos se vendieron en un rendimiento similar. Con esto se podrían generar las sugerencias de que se podría vender en conjunto o hasta saber cuale son las carecterísticas del grupo con rendimiento más bajo.