UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

TECNOLOGÍAS DE LA INFORMACIÓN

Extracción de Conocimiento en Bases de Datos

IV.1. Algoritmos de agrupación

IDGS91N

PRESENTA:

JORGE ALEJANDRO HERNANDEZ CONTRERAS

DOCENTE:

Enrique Mascote

Chihuahua, Chih.28 de Noviembre de 2025

Introducción

En la extracción de conocimiento a partir de datos, dos procesos fundamentales son el clustering y la reducción de dimensionalidad. El clustering permite agrupar observaciones sin necesidad de etiquetas, encontrando patrones subyacentes, similitudes y estructuras ocultas dentro de un conjunto de datos. Por su parte, la reducción de dimensionalidad transforma un conjunto de variables originales en un espacio de menor dimensión, conservando la mayor parte de la información relevante; esto facilita el análisis, mejora el rendimiento computacional y reduce el ruido. Ambas técnicas son esenciales en áreas como minería de datos, ciencia de datos, análisis exploratorio y aprendizaje automático, permitiendo comprender datos complejos, visualizar estructuras y preparar información para modelos predictivos más eficientes.

Algoritmos de agrupación

A continuación se describen tres algoritmos de clustering: K-Means, DBSCAN y Gaussian Mixture Models.

K-Means

Principio de funcionamiento

K-Means es un método de agrupación basado en particiones. Divide los datos en k clusters, minimizando la distancia entre los puntos y su centroide. Funciona iterativamente con dos pasos:

- **Asignación:**
  cada punto se asigna al centroide más cercano.
- **Actualización:**
  se recalculan los centroides promediando los puntos asignados.

El proceso continúa hasta que los centroides dejan de cambiar significativamente.

Parámetros clave

- **k:**
  número de clusters deseado.
- **Criterio de parada:**
  número máximo de iteraciones o cambio mínimo en los centroides.
- **Distancia usada:**
  normalmente Euclidiana.

Ventajas

- Rápido y eficiente con grandes volúmenes de datos.
- Fácil de implementar e interpretar.

Limitaciones

- Se debe elegir k previamente.
- No funciona bien con clusters de formas no esféricas.
- Sensible a valores atípicos.

Ejemplo

```
Inicializar k centroides aleatorios
Repetir:
    Para cada punto:
        Asignar al centroide más cercano
    Actualizar cada centroide como el promedio de su cluster
Hasta convergencia

```

DBSCAN

Principio de funcionamiento

DBSCAN agrupa puntos basándose en densidad. Clasifica los puntos en:

- Core points: tienen suficientes puntos vecinos.
- Border points: están cerca de un punto núcleo.
- Noise points: no pertenecen a ningún cluster.

Expande clusters conectando puntos densamente alcanzables.

Parámetros clave

- eps: radio para buscar vecinos.
- minPts: mínimo de puntos dentro del radio ε para formar un punto núcleo.

Ventajas

- Detecta clusters de formas arbitrarias.
- Identifica ruido y outliers.
- No requiere especificar k.

Limitaciones

- Sensible a la selección de ε y minPts.
- Difícil de usar con datos de alta dimensión (la distancia pierde significado).

Ejemplo

```
Para cada punto no visitado:
    Marcar como visitado
    Obtener vecinos dentro de ε
    Si #vecinos < minPts → marcar como ruido
    Si no:
        Crear nuevo cluster
        Expandir cluster con vecinos densamente conectados
```

Gaussian Mixture Models

Principio de funcionamiento

GMM asume que los datos proceden de una mezcla de distribuciones gaussianas.
Utiliza el algoritmo EM:

- E-step: calcula la probabilidad de que cada punto pertenezca a cada componente gaussiana.
- M-step: actualiza los parámetros de cada gaussiana.

A diferencia de K-Means, GMM asigna probabilidades de pertenencia, no asignaciones duras.

Parámetros clave

- Número de componentes (k).
- Covariance type: full, tied, diagonal o spherical.
- Criterios de convergencia.

Ventajas

- Identifica clusters elípticos o complejos.
- Asignación probabilística más flexible.
- Mejor que K-Means cuando los clusters tienen formas diferentes.

Limitaciones

- Debe elegirse k.
- Puede converger a soluciones locales.
- Requiere más tiempo computacional.

Ejemplo

```
Inicializar parámetros de k gaussianas
Repetir:
    E-step: calcular probabilidad de pertenencia de cada punto
    M-step: actualizar medias, covarianzas y pesos
Hasta convergencia

```

Algoritmos de reducción de dimensionalidad

A continuación se describen PCA y t-SNE.

Análisis de Componentes Principales

Fundamento matemático

PCA transforma los datos a un nuevo sistema de coordenadas calculando las direcciones de máxima varianza. Matemáticamente:

- Centrar los datos.
- Calcular la matriz de covarianza.
- Obtener vectores propios y valores propios.
- Proyectar los datos en los componentes principales.

Parámetros clave

- Número de componentes a conservar.
- Estandarización previa.

Ventajas

- Muy eficiente.
- Reduce ruido.
- Útil para visualización en 2D o 3D.

Limitaciones

- Solo capta relaciones lineales.
- Difícil de interpretar componentes.
- Escala de variables afecta resultados.

Ejemplo

```
1   Calcular matriz de covarianza
2   Obtener eigenvectors y eigenvalues
3   Seleccionar los componentes principales
4   Proyectar datos
5
```

t-SNE

Fundamento conceptual

t-SNE es una técnica de reducción no lineal diseñada para visualización.
Convierte distancias entre puntos en probabilidades de similitud, preservando estructuras locales. Usa una distribución t para evitar el crowding problem.

Parámetros clave

- Perplexity: controla el equilibrio entre vecinos cercanos y globales.
- Learning rate.
- Número de iteraciones.

Ventajas

- Excelente para visualizar datos complejos en 2D/3D.
- Captura relaciones no lineales.

Limitaciones

- No conserva estructura global.
- Lento con conjuntos grandes.
- No sirve para modelos predictivos, solo visualización.

Ejemplo

```
Calcular similitud entre cada par de puntos
Generar mapa inicial en 2D
Optimizar posiciones minimizando la divergencia KL
```

Comparativa

Objetivo	**Clustering**	**Reducción de dimensionalidad**
¿Qué hace?	Agrupa puntos similares en clusters.	Reduce variables manteniendo información clave.
Tipo de técnica	No supervisada	Preprocesamiento o visualización
Salida	Etiquetas de cluster	Nuevas dimensiones
Cuándo usar	Encontrar grupos ocultos, segmentación	Visualizar datos, eliminar ruido, acelerar modelos

Conclusión

El clustering y la reducción de dimensionalidad son fundamentales para comprender estructuras internas de los datos. Algoritmos como K-Means, DBSCAN y GMM permiten identificar grupos con diferentes supuestos y formas. Por su parte, PCA y t-SNE facilitan la visualización y simplificación de datos. Elegir la técnica adecuada depende del tipo de datos, su dimensionalidad, la presencia de ruido y los objetivos del análisis. Su uso conjunto potencia la exploración y el descubrimiento de patrones complejos.

Bibliografía

Google Developers. (2024). *Clustering: K-Means*. https://developers.google.com/machine-learning/clustering/k-means

Scikit-Learn. (2024). *DBSCAN clustering*. https://scikit-learn.org/stable/modules/clustering.html#dbscan

TensorFlow. (2024). *Understanding PCA and dimensionality reduction*. https://www.tensorflow.org/tutorials/text/word_embeddings