UNIVERSIDAD TECNOLÓGICA DE CHIHUAHUA

Tecnologías de la información

Extracción de conocimiento en base de datos

IV.2. Métricas de evaluación de modelos (50%)

IDGS91N

Presenta:

Brahn Raudales

Docente:

Enrique Mascote

sábado, 29 de noviembre de 2025

1. Introducción

En escenarios de análisis de datos sin etiquetas (no supervisado), dos técnicas muy útiles son el **clustering** —para agrupar datos similares— y la **reducción de dimensionalidad** — para simplificar espacios de variables, eliminar redundancia y facilitar visualización. Sin embargo, para evaluar la calidad de los resultados de clustering o reducción es necesario usar **métricas adecuadas**, ya que no existe una "verdad" externa cuando no se tiene etiqueta. Este reporte busca presentar varias de esas métricas, explicar su interpretación y limitaciones, y demostrar su uso práctico aplicándolas a un conjunto de datos real, combinando clustering y reducción de dimensionalidad.

2. Métricas de agrupación (clustering)

A continuación se describen tres métricas internas usadas para evaluar la calidad de agrupamientos sin necesidad de etiquetas verdaderas.

2.1 Índice de silueta (Silhouette Score)

Definición y fórmula
Para cada punto i, se define:

- a_i: la distancia promedio entre el punto y todos los demás puntos del mismo cluster (cohesión interna).
- b_i: la distancia promedio entre el punto y todos los puntos del próximo cluster más cercano (separación).

La puntuación de silueta para el punto $i$es:

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

El **Silhouette Score** del clustering es el promedio de $s_i$sobre todos los puntos. El valor oscila entre -1 y +1. Medium+2ScienceDirect+2

Interpretación

- $s \approx +1$: los puntos están bien agrupados, cerca de su cluster y lejos de otros → buen clustering.
- $s \approx 0$: los clusters se superponen o no están bien definidos.
- $s \approx -1$: los puntos podrían estar mal asignados (tal vez pertenecen a otro cluster).

Ventajas

- Intuitiva y fácil de interpretar.
- Considera tanto cohesión interna como separación externa.
- Útil para comparar diferentes particiones/clustering.

Limitaciones

- Sensible a la densidad y forma de los clusters: si los clusters tienen densidades o tamaños muy distintos, la silueta puede ser engañosa. Codefinity+1
- Solo mide calidad interna; no garantiza que los clusters tengan significado semántico real.

2.2 Índice de Davies–Bouldin (DBI)

Definición y fórmula
Para cada cluster $i$se calcula su dispersión interna S_i(por ejemplo promedio de distancias de puntos al centroide). Para cada par de clusters (i, j), se calcula la separación M_{ij}(por ejemplo la distancia entre centroides). Luego:

$$R_{i,j} = \frac{S_i + S_j}{M_{ij}}$$

Para cada cluster i, se toma $R_i = \max_{j \neq i} R_{i,j}$. Finalmente, el índice Davies–Bouldin se define como:

undefined

donde $p$es la dimensión original. Ciencia de Datos+1

Interpretación

- Un valor alto (por ejemplo ≥ 80-90 %) indica que los primeros componentes retienen la mayoría de la información de variabilidad, lo que sugiere que la dimensión puede reducirse sin perder mucho contenido.
- Un valor bajo indica que se pierde demasiada información si se reduce tanto.

Ventajas

- Fácil de calcular y entender.
- Útil para decidir cuántas dimensiones conservar.

Limitaciones

- Solo considera varianza — no garantiza que las componentes retengan relaciones relevantes para la tarea (por ejemplo clusters, estructura no lineal).
- No aplica bien si la información importante no está en las direcciones de mayor varianza.

3.2 Trustworthiness (distancia media de vecinos en espacio reducido)

Definición conceptual
Cuando se reduce dimensión (por ejemplo con técnicas de proyección no lineal, como t-SNE), *trustworthiness* mide qué tan bien se preservan las relaciones de vecindad local del espacio original en el espacio reducido. Es una métrica de calidad de la proyección. (Nota: la fórmula exacta depende de ranking de vecinos, pero la idea es cuantificar cuántos de los vecinos más cercanos en espacio reducido respetan los vecinos cercanos en espacio original). KDnuggets+1

Interpretación

- Trustworthiness cercano a 1 → la mayoría de los vecinos cercanos en el espacio reducido correspondían a vecinos cercanos en el espacio original → buena preservación de estructura local.
- Valor bajo → la proyección distorsiona relaciones locales, por lo que puede no ser confiable para análisis basado en proximidad.

Ventajas

- Evalúa la preservación de estructura local — útil si lo que importa es proximidad entre puntos.
- Es independiente de etiquetas externas.

Limitaciones

- No mide preservación global de relaciones distantes.
- Depende del número de vecinos considerados.

4. Caso de estudio y aplicación práctica

4.1 Descripción del dataset

Para este ejemplo uso el clásico conjunto de datos Iris dataset — contiene 150 muestras de flores con 4 atributos numéricos: largo y ancho del sépalo, largo y ancho del pétalo. Es ideal para clustering y reducción de dimensión.

(Nota: si prefieres usar otro dataset, solo cambia los datos y variables en el código.)

4.2 Aplicación de clustering + cálculo de métricas de agrupación

- Aplicarás un algoritmo de clustering (por ejemplo, K-Means con $k = 3$— dado que Iris tiene 3 clases verdaderas).
- Después de obtener las etiquetas de cluster, calcularás: Silhouette Score, Davies–Bouldin Index, Calinski–Harabasz Index.

(Código en Python usando scikit-learn — en anexos.)

4.3 Aplicación de reducción de dimensionalidad + métrica

- Aplicarás PCA sobre el dataset para reducir de 4 dimensiones originales a 2 componentes.
- Calcularás la **varianza explicada acumulada** por las 2 componentes.
- Opcional: también puedes calcular *trustworthiness* si usas una proyección no lineal (por ejemplo t-SNE), para evaluar la calidad de la reducción.

4.4 Resultados — tablas y gráficas

- Gráfico scatter de los clusters en el espacio original reducido a 2D (por PCA o t-SNE), coloreado por cluster.
- Tabla con los valores de las 3 métricas de agrupación.

- Gráfica de varianza explicada acumulada por componente principal (para mostrar cuánta información retienes al reducir).

5. Comparativa y análisis

- Discutir qué tan “buenos” se ven los clusters según cada métrica: por ejemplo, un alto Silhouette, un alto índice CH y un bajo DBI indican un buen agrupamiento.
- Ver si la reducción de dimensionalidad con PCA conserva suficiente variabilidad (o vecindad) para que clustering siga siendo válido.
- Reflexionar sobre fortalezas y limitaciones: por ejemplo, si los clusters se ven bien en 2D pero la métrica DBI sube, puede que la reducción haya distorsionado relaciones.

6. Conclusiones y recomendaciones

- La combinación de métricas de clustering permite evaluar desde distintos ángulos: cohesión, separación, separación respecto a variabilidad global.
- Las métricas de reducción permiten decidir cuántas dimensiones conservar y evaluar la fidelidad de la proyección.
- Es recomendable usar **varias métricas a la vez** y **visualización** para validar resultados.
- En datasets reales, puede ser útil experimentar con distintos algoritmos, distintos valores de ko distintos métodos de reducción, y comparar.

7. Referencias

Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (2ª ed.). O'Reilly Media.
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*.
Van der Maaten, L., & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9*, 2579–2605.
OpenAI. (2025). *ChatGPT* [Modelo de lenguaje]. OpenAI. https://chat.openai.com/

Anexo: Pseudocódigo / pasos sugeridos (o código en Python)

Puedes incluir un bloque de código como este (o similar) para ilustrar la implementación del clustering + PCA + cálculo de métricas. Si quieres, te lo genero completo.