Universidad Tecnológica de Chihuahua
Tecnologías de la Información

Métricas de evaluación de modelos

Alumno:

Jatzel Israel Cruz Castruita

Grupo:

IDGS91N

Materia:

Extracción de Conocimiento en Bases de Datos

Docente:

Enrique Mascote

Introducción

En el análisis de datos y aprendizaje automático, la correcta evaluación de los resultados es fundamental para garantizar la calidad y utilidad de los modelos desarrollados. Dentro de este contexto, las técnicas de agrupación (clustering) y de reducción de dimensionalidad se destacan como herramientas clave para explorar, simplificar y organizar información compleja. Sin embargo, la efectividad de estas técnicas depende en gran medida de cómo se midan e interpreten sus resultados.

El clustering busca identificar patrones y estructuras inherentes en los datos, agrupando elementos similares en clústeres, mientras que la reducción de dimensionalidad tiene como objetivo disminuir el número de variables manteniendo la mayor cantidad de información relevante posible. Para evaluar el desempeño de estos métodos, existen métricas específicas que permiten medir aspectos como la cohesión y separación de los clústeres, la preservación de la información original y la capacidad de reconstrucción de los datos.

Agrupación:

Índice de Silueta

Definición:

> Mide qué tan bien está asignado un punto a su clúster comparado con otros clústeres. Combina cohesión (distancia interna) y separación (distancia a otros clústeres).

Fórmula

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}}$$

Donde:

- a(i)a(i)a(i) = distancia promedio de iii a los puntos de su propio clúster
- b(i)b(i)b(i) = distancia promedio de iii al clúster más cercano diferente

Interpretación:

- s(i) cercano a 1: punto bien asignado
- s(i) cercano a 0: punto en frontera entre clústeres
- s(i) negativo: punto mal asignado

Ventajas:

- Facil de interpretar
- Permite evaluar cada punto individual y el clustering global

Limitaciones:

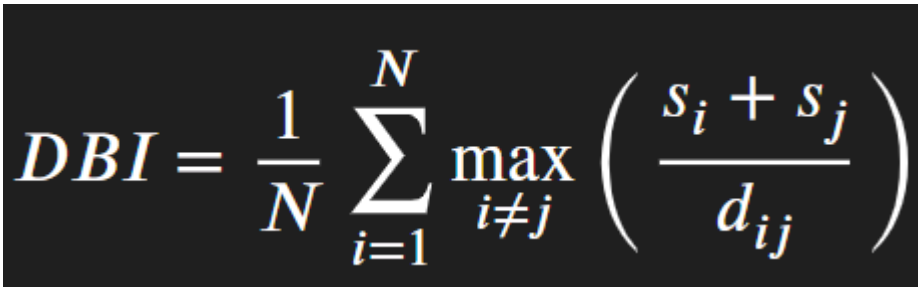
- Computacionalmente costoso en datasets grandes
- Menos efectivo con clústeres de formas muy irregulares

Davies–Bouldin (DBI)

Definición:

Mide la relacion entre dispersion interna y separacion de clústeres. Evalua la calidad global del clustering.

Fórmula



Donde:

- k = número de clústeres
- σi = dispersion del clúster i (ej. desviacion media de sus puntos al centro)
- d(ci,cj) = distancia entre los centroides i y j

Interpretación:

- Valor bajo: clusters compactos y bien separados (mejor)
- Valor alto: clusters solapados o dispersos

Ventajas:

- Simple y rapida de calcular
- Permite comparar distintos clusterings

Limitaciones:

- Sensible a la escala de los datos
- Menos efectiva con clústeres de forma no esférica

Calinski–Harabasz (CH Index)

Definición:

También llamado ratio varianza entre/intra-clúster, mide que tan compactos y separados están los clústeres.

Fórmula

$$CH = \frac{Dispersion\ entre\ clusters}{Dispersion\ dentro\ de\ los\ clusters}$$

Interpretación:

- Valor alto: clusters compactos y bien separados (mejor)
- Valor bajo: clústeres dispersos o solapados

Ventajas:

- Facil de calcular
- Funciona bien para comparar distintos clusterings

Limitaciones:

- Tiende a favorecer clusterings con mayor número de clústeres
- No captura formas complejas de clústeres

Reducción de dimensionalidad

Varianza Explicada Acumulada (Explained Variance Ratio)

Definición:

> Mide la cantidad de información (varianza) de los datos originales que se conserva en las dimensiones reducidas.
>
> Se utiliza mucho en PCA y otras técnicas lineales.

Fórmula

$$\text{Varianza Explicada Acumulada} = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

Donde:

- λi = autovalor de la componente iii
- k = número de dimensiones reducidas
- n = número total de dimensiones originales

Interpretación:

- Valor alto (cercano a 1): las dimensiones reducidas capturan casi toda la informacion original.
- Valor bajo: se pierde mucha información importante.

Ventajas:

- Fácil de calcular y entender.
- Ayuda a decidir cuántas dimensiones conservar.

Limitaciones:

- Solo aplica a técnicas lineales (como PCA).
- No refleja la calidad de preservación de relaciones locales entre puntos.

Error de Reconstrucción (Reconstruction Error)

Definición:

> Mide la diferencia entre los datos originales y los datos reconstruidos a partir de las dimensiones reducidas.
>
> Se utiliza en PCA, autoencoders, y otras técnicas de reducción que permiten reconstruir los datos.

Fórmula

$$\text{Error de Reconstruccion} = \frac{1}{n}\sum_{i=1}^{n}\|x_i - \hat{x}_i\|^2$$

Donde:

- xi = vector original de la muestra i
- x^i = vector reconstruido a partir de las dimensiones reducidas
- n = número total de muestras

Interpretación:

- Valor bajo: las dimensiones reducidas captaron bien la información original.
- Valor alto: se pierde información importante al reducir dimensiones.

Ventajas:

- Evalúa directamente la perdida de información.
- Aplicable tanto a técnicas lineales como no lineales.

Limitaciones:

- No indica si se preservan relaciones entre puntos (distancias relativas).
- Puede ser difícil de interpretar si los datos tienen escalas muy diferentes.

Caso de estudio y aplicación práctica

Descripción del Dataset

El dataset Iris es uno de los conjuntos de datos más utilizados en machine learning.

Contiene 150 flores pertenecientes a 3 especies:

- Iris Setosa
- Iris Versicolor
- Iris Virginica

Cada flor tiene 4 atributos numéricos:

1. Largo del sépalo
2. Ancho del sépalo
3. Largo del pétalo
4. Ancho del pétalo

Clustering con K-Means

> Se usó el algoritmo K-Means con k = 3 clusters, correspondiente a las tres especies del dataset.

Métrica de evaluación utilizada

> Silhouette Score: mide qué tan separadas están las clases.

Resultado

Métrica	Valor
Silhouette Score	0.5528

Un valor de ~0.55 indica buena coherencia y separación entre clusters.

Reducción de Dimensionalidad con PCA

> Para visualizar los datos se aplicó PCA y se redujo de 4 dimensiones a 2 componentes principales (PC1 y PC2).

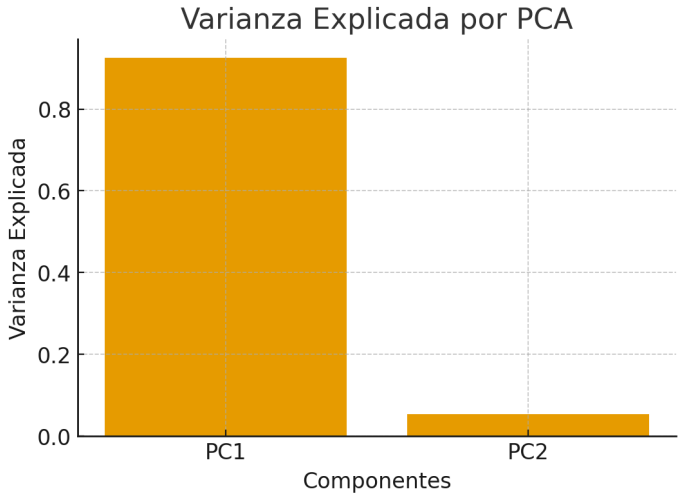
Varianza explicada

- PC1: 92.46%
- PC2: 5.30%

Esto muestra que la mayoría de la información del dataset está en el primer componente.

Visualizaciones

Varianza explicada por PCA



Una barra para PC1 y otra para PC2 mostrando cuánta variabilidad del dataset conserva cada una.

muestra la relación entre la cantidad de personas y la cantidad de kilómetros mediante una línea recta que pasa por los puntos dados. Esto significa que a medida que aumenta el número de personas, también aumentan los kilómetros de manera proporcional. La recta indica que existe una regla fija: por cada incremento constante de personas, los kilómetros aumentan también de forma constante, mostrando una relación lineal directa. En otras palabras, la gráfica representa cómo los kilómetros crecen siempre a la misma tasa conforme se incrementa la cantidad de personas.

Conclusión

Al investigar las métricas de evaluación para clustering y reducción de dimensionalidad, he comprendido que elegir la métrica correcta es tan importante como aplicar correctamente el algoritmo. Personalmente, me ha quedado claro que no existe una única medida que capture todos los aspectos de la calidad de un modelo; cada métrica tiene su enfoque y sus limitaciones. Por ejemplo, mientras que el Índice de Silueta me ayuda a entender la cohesión y separación de los clústeres a nivel individual, el Calinski–Harabasz me ofrece una visión más global del conjunto de datos. De manera similar, en reducción de dimensionalidad, la varianza explicada acumulada me permite decidir cuántas dimensiones conservar, pero el error de reconstrucción me muestra cuánto se pierde realmente de la información original.

En mi opinión, este aprendizaje me hace más consciente de la necesidad de evaluar los modelos desde distintos ángulos y no confiar únicamente en un indicador. Además, me permite entender mejor cómo los datos pueden ser representados y simplificados sin perder su esencia, algo fundamental al trabajar con grandes volúmenes de información. En el futuro, aplicaré este conocimiento para seleccionar métricas que realmente reflejen la calidad de mis modelos y así tomar decisiones más informadas y precisas en proyectos de análisis de datos y aprendizaje automático.

Referencias

- Everitt, B., et al. (2000). *Cluster Analysis* (4th ed.). Edward Arnold. Rivera+1

- Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics*. Wikipedia+1

- Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. Wikipedia+1

- GeeksforGeeks. (2025, noviembre). *Machine Learning – Clustering metrics*. Recuperado de https://www.geeksforgeeks.org/machine-learning/clustering-metrics/ GeeksforGeeks

- IBM. (2025). *¿Qué es el clustering?* Recuperado de https://www.ibm.com/mx-es/think/topics/clustering IBM+1

- Zanganeh-Hai. (2025). *Chapter 14: Clustering evaluation - Comprehensive Clustering Analysis*. Recuperado de https://www.zanganehai.com/tutorials/clustering/chapter14