Guía de PCA y su Uso en Modelos Supervisados

PCA (Análisis de Componentes Principales) es una técnica de aprendizaje no supervisado que se utiliza para reducir la dimensionalidad de un conjunto de datos. Aunque no es un modelo de clasificación, se puede usar como paso previo para modelos supervisados como regresión logística, árboles de decisión o KNN. PCA transforma variables correlacionadas en un conjunto de componentes no correlacionados que explican la mayor varianza posible del conjunto original.

¿Cómo se aplica PCA antes de un modelo supervisado?

1. Se parte de un conjunto de variables predictoras (X).
2. Se estandarizan los datos para que todas las variables tengan media 0 y desviación estándar 1.
3. Se calcula la matriz de covarianza de los datos escalados.
4. Se encuentran los autovalores (λ) y autovectores (v) de la matriz de covarianza.
5. Se seleccionan los autovectores correspondientes a los mayores autovalores para formar las nuevas dimensiones.
6. Se proyectan los datos originales en estas nuevas dimensiones para obtener los componentes principales.
7. Estos componentes se usan como entrada para un modelo supervisado.

Fórmula general

Sean X los datos escalados. La matriz de covarianza se define como:

$$\Sigma = (1/n) * X^T X$$

Luego se resuelve el problema de autovalores:

$$\Sigma v = \lambda v$$

donde v son los autovectores (direcciones principales) y λ los autovalores (magnitud de la varianza en esa dirección).

Los datos transformados se obtienen como:

$$Z = XW$$

donde W es la matriz de los k autovectores principales seleccionados.

Ejemplo sencillo

Supongamos que tenemos un conjunto de datos con 3 variables: edad, ingresos y número de horas trabajadas.

Aplicamos PCA y decidimos quedarnos con los dos primeros componentes que explican el 95% de la varianza.

Luego usamos estos dos componentes como variables de entrada para entrenar un modelo de clasificación binaria que predice si una persona tiene alta experiencia laboral.

Ventajas de usar PCA antes de un modelo supervisado

- - Reduce la cantidad de variables a manejar
- - Elimina redundancia entre variables correlacionadas
- - Mejora el rendimiento de algunos algoritmos
- - Ayuda a visualizar los datos en 2D o 3D

¿Siempre mejora el modelo?

No siempre. Si tus variables originales están bien seleccionadas y son relevantes para el objetivo del modelo, usar PCA puede eliminar información útil. Por eso, es recomendable comparar el rendimiento del modelo con y sin PCA.

PCA es una herramienta muy útil para preprocesamiento de datos y visualización, y puede mejorar el rendimiento de los modelos supervisados si se usa correctamente.