



TÉCNICAS DE APRENDIZAJE NO SUPERVISADO

Nicolás García
Hernández

Principal Component Analysis(PCA)

El Análisis de Componentes Principales es una técnica que reduce la dimensionalidad de datos transformando las variables potencialmente correlacionadas en un conjunto más pequeños de variables denominadas componentes principales.

Un algoritmo que simplifica datos complicados eliminando la información menos importante, manteniendo solo lo esencial

Problemas donde se aplica

1

Visualizacion de datos de alta dimension

Permite proyectar datos multidimensionales en espacios de 2 o 3 dimensiones para visualizacion e identificacion de tendencias, patrones y valores atipicos.

2

Preprocesamiento de datos

Se utiliza como paso preliminar antes de entrenar algoritmos de machine learning, extrayendo caracteristicas mas informativas.

3

Procesamiento de imagenes y señales

Facilita la compresion de datos y extraccion de caracteristicas relevantes.

4

Analisis financiero

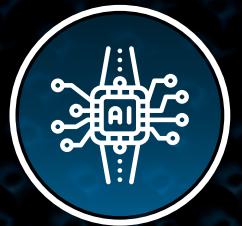
Ayuda a identificar patrones en datos bursatiles y a prever comportamientos del mercado.

5

Analisis metabolico y genomico

En bioinformatica, se emplea para explorar e interpretar conjuntos de datos complejos de expresion genica y estudios metabolicos.

Ventajas



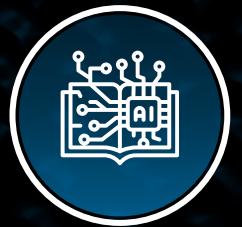
Reducción de la complejidad

Disminuye el numero de variables eliminando características que representan ruido, lo que tambien reduce el sobreajuste y la maldicion de la dimensionalidad.



Mejora el rendimiento computacional

Al trabajar con menos características, los algoritmos de machine learning son más eficientes y tienen mejor rendimiento.



Naturaleza no supervisada

No requiere datos etiquetados, lo que lo hace más flexible y parametrización más simple.



Limitaciones

Limitación a variables numéricas: PCA solo funciona con datos numéricos, no con variables categóricas o texto

Asunción de linealidad: El método es estrictamente lineal, lo que limita su efectividad cuando las relaciones entre variables son no lineales. Para abordar esto existe el Kernel PCA, que captura características no lineales mediante funciones kernel

Dependencia de correlación: PCA no funciona efectivamente cuando las variables no están fuertemente correlacionadas entre sí

Sensibilidad a la escala: Los resultados se ven significativamente afectados por la escala de las características y la presencia de valores atípicos, requiriendo normalización y escalado previos

Gaussian Mixture Models (GMM)

Es un algoritmo que agrupa datos automáticamente en categorías flexibles, asignando a cada dato una probabilidad de pertenencia a cada grupo en lugar de obligarlo a estar en uno solo

Problemas donde se aplica

1

Clustering probabilístico

Agrupa datos en clústeres donde cada observación tiene una probabilidad de pertenencia a múltiples clústeres, permitiendo transiciones suaves entre grupos.

2

Segmentación de clientes

Identifica segmentos de clientes en marketing, capturando grupos con características similares pero también la variabilidad dentro de cada grupo.

3

Detección de anomalías

Mediante GMM, se pueden identificar observaciones que tienen baja probabilidad bajo el modelo aprendido, indicando comportamientos anómalos o fraudes

4

Procesamiento de imágenes médicas

Se utiliza para segmentación de imágenes PET identificando regiones con diferentes intensidades de actividad

5

Análisis de datos con clústeres de diferentes tamaños y formas

Es superior a K-Means cuando los clústeres tienen características heterogéneas

Ventajas



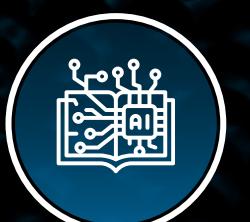
Captura de variabilidad dentro de clústeres

Una ventaja significativa sobre K-Means es la capacidad de modelar la variabilidad y la forma de los clústeres mediante la matriz de covarianza de cada componente



Asignaciones probabilísticas

Proporciona probabilidades de pertenencia a múltiples clústeres, no solo asignaciones binarias, lo que es más realista para muchas aplicaciones del mundo real



Flexibilidad en formas de clúster

GMM puede identificar clústeres de diferentes tamaños y formas, mientras que K-Means tiende a encontrar clústeres esféricos



Limitaciones



Mayor complejidad computacional: GMM requiere más cálculos que K-Means, especialmente al estimar matrices de covarianza completas, lo que lo hace más lento en grandes conjuntos de datos

Dificultad en estimación de parámetros: La optimización de la probabilidad de mezcla es algebraicamente más compleja que en K-Means, lo que puede dificultar la convergencia

Necesidad de especificar el número de componentes: Como K-Means, debe especificarse de antemano cuántos componentes gaussianos se desean, aunque se pueden usar criterios como BIC para seleccionar este número óptimamente

Sensibilidad a inicialización: La elección de los valores iniciales para las medias y covarianzas afecta significativamente los resultados finales, pudiendo converger a soluciones subóptimas

Muchas
Gracias

