Análisis de Componentes Principales (PCA) y su Fundamento en Álgebra Lineal

Nicole Susan Loza Ticona

December 9, 2024

Introducción

El Análisis de Componentes Principales (PCA) es una técnica ampliamente utilizada en estadística y aprendizaje automático para la reducción de dimensionalidad. Se emplea para transformar un conjunto de variables posiblemente correlacionadas en un nuevo conjunto de variables no correlacionadas, denominadas **componentes principales**. El PCA es particularmente útil para visualizar y reducir la complejidad de grandes conjuntos de datos manteniendo la mayor cantidad de varianza posible.

En este documento, se explicará cómo funciona PCA desde el punto de vista del álgebra lineal, detallando cada paso del proceso matemático involucrado.

Fundamentos Matemáticos de PCA

El PCA se basa en operaciones fundamentales del álgebra lineal, como la descomposición en valores propios y vectores propios. Los pasos involucrados son los siguientes:

1. Estandarización de los Datos

El primer paso en PCA es estandarizar los datos, es decir, centrarlos en torno a cero y, en algunos casos, escalarlos para que todas las variables tengan la misma escala. Dado un conjunto de datos X con n observaciones y m variables, el proceso de estandarización se realiza de la siguiente manera:

$$X_{\text{centrado}} = X - \bar{X}$$

donde $\bar{X}$ es la media de cada columna (característica) de X.

Es fundamental centrar los datos para evitar que las variables con mayor magnitud dominen el análisis.

2. Cálculo de la Matriz de Covarianza

La matriz de covarianza captura la relación lineal entre las diferentes variables. Para calcularla, se utiliza la siguiente fórmula:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}_{\text{centrado}}^T\mathbf{X}_{\text{centrado}}$$

Donde: - $\mathbf{C}$ es la matriz de covarianza ($m \times m$). - $\mathbf{X}_{\text{centrado}}$ es la matriz de datos centrados. - n es el número de observaciones.

La matriz de covarianza es simétrica, y sus entradas C_{ij} representan la covarianza entre las variables i y j.

3. Descomposición en Valores Propios y Vectores Propios

El siguiente paso es obtener la descomposición espectral de la matriz de covarianza. Esta descomposición consiste en encontrar los **vectores propios** y los **valores propios** de la matriz de covarianza. La ecuación es:

$$\mathbf{C}\mathbf{v}_i = \lambda_i\mathbf{v}_i$$

Donde: - $\mathbf{v}_i$ es el i-ésimo vector propio. - λ_i es el valor propio correspondiente, que indica la cantidad de varianza que es explicada por el componente asociado a $\mathbf{v}_i$.

Los vectores propios $\mathbf{v}_i$ definen las direcciones principales en el espacio de características, y los valores propios λ_i indican cuánta varianza es capturada por cada dirección. La importancia de cada componente está dada por su valor propio λ_i.

4. Ordenación de Componentes

Una vez que hemos calculado los vectores propios y sus valores propios, ordenamos los componentes principales de acuerdo a la magnitud de sus valores propios. Esto nos indica qué componentes retienen la mayor cantidad de información (varianza) del conjunto de datos. La ordenación es de la siguiente forma:

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$$

Así, los primeros k componentes principales seleccionados serán los que explican la mayor parte de la varianza de los datos.

5. Proyección de los Datos en el Nuevo Espacio

Una vez seleccionados los componentes principales, el siguiente paso es proyectar los datos originales en el nuevo espacio definido por los vectores propios seleccionados. Este paso se realiza multiplicando la matriz de datos centrados por la matriz de vectores propios:

$$\mathbf{Z} = \mathbf{X}_{\text{centrado}}\mathbf{W}$$

Donde: - $\mathbf{Z}$ es la matriz de datos proyectados en el nuevo espacio ($n \times k$). - $\mathbf{W}$ es la matriz de vectores propios seleccionados ($m \times k$).

El número k representa el número de componentes principales seleccionados para la proyección.

6. Varianza Explicada y Selección del Número Óptimo de Componentes

Para seleccionar el número óptimo de componentes, se analiza la **varianza explicada** por los componentes principales. La varianza total es la suma de los valores propios:

$$\text{Varianza Total} = \sum_{i=1}^{m} \lambda_i$$

La varianza explicada por los primeros k componentes es:

$$\text{Varianza Explicada por } k = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{m} \lambda_i}$$

El objetivo es seleccionar el valor de k tal que la varianza explicada por los k componentes sea lo suficientemente alta (generalmente se busca explicar más del 95% de la varianza).

Conclusión

El PCA es una herramienta poderosa para la reducción de dimensionalidad. A través de los pasos de estandarización, cálculo de la matriz de covarianza, descomposición en valores propios y proyección, podemos transformar los datos originales en un espacio de menor dimensión, preservando la mayor parte de la varianza. La clave del PCA radica en la capacidad de identificar las direcciones (componentes) en las que los datos varían más y reducir las dimensiones al seleccionar los componentes principales que mejor explican esa varianza.