Curso Corto y Profundo: Análisis de Componentes Principales (PCA)

Dr. Manuel Moreira

Introducción

El Análisis de Componentes Principales (PCA) es una técnica de reducción de dimensionalidad que transforma un conjunto de variables correlacionadas en un conjunto de variables no correlacionadas llamadas componentes principales. Este curso cubre los fundamentos matemáticos, aplicaciones y aspectos históricos.

1. Generalidades de Álgebra Lineal: Eigenvectores y Eigenvalores

Espacios vectoriales y transformaciones lineales

Un espacio vectorial es un conjunto de vectores que pueden ser sumados y multiplicados por escalares. Una transformación lineal $T : V \to W$ satisface:

$$T(\mathbf{u} + \mathbf{v}) = T(\mathbf{u}) + T(\mathbf{v}), \quad T(c\mathbf{u}) = cT(\mathbf{u})$$

Las matrices representan transformaciones lineales. Dada una matriz A, un vector no nulo $\mathbf{v}$ es un eigenvector si:

$$A\mathbf{v} = \lambda\mathbf{v}$$

donde λ es el eigenvalor asociado. La ecuación característica es:

$$\det(A - \lambda I) = 0$$

Relación con PCA

En PCA, los componentes principales son los eigenvectores de la matriz de covarianza Σ. Los eigenvalores representan la varianza explicada por cada componente.

2. Distancia de Mahalanobis y Distribuciones de Probabilidad

Distancia de Mahalanobis

Para un vector $\mathbf{x}$ en un espacio con media $\boldsymbol{\mu}$ y matriz de covarianza Σ, la distancia de Mahalanobis es:

$$d_M(\mathbf{x}) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

Esta distancia considera la correlación entre variables.

Distribución normal multivariada

La función de densidad de una normal multivariada $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ es:

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$$

Las isodensas forman elipsoides, cuya forma depende de Σ.

Link con PCA

Al aplicar PCA, se transforma $\mathbf{X}$ a $\mathbf{Y} = \mathbf{X}W$, donde W es la matriz de eigenvectores de Σ. En el nuevo espacio, la matriz de covarianza es diagonal y la distancia de Mahalanobis se convierte en euclídea.

3. Matriz de Covarianza vs. Matriz de Correlación

Matriz de covarianza

La matriz de covarianza Σ tiene elementos:

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j) = \mathbb{E}[(X_i - \mu_i)(X_j - \mu_j)]$$

La varianza total es la traza de Σ: $\sum_{i=1}^{d} \sigma_{ii} = \sum_{i=1}^{d} \lambda_i$.

Matriz de correlación

La matriz de correlación ρ se define como:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}, \quad \sigma_i = \sqrt{\sigma_{ii}}$$

Esta matriz estandariza las variables.

Elección en PCA

- Usar Σ cuando las variables tienen unidades similares.
- Usar ρ cuando las variables tienen escalas muy diferentes.

4. Deducción Matemática del Modelo PCA

Formulación del problema

Buscamos direcciones $\mathbf{w}$ que maximicen la varianza de la proyección:

$$\max_{\mathbf{w}} \mathrm{Var}(\mathbf{w}^\top \mathbf{X}) = \mathbf{w}^\top \Sigma \mathbf{w}, \quad \text{sujeto a } \mathbf{w}^\top \mathbf{w} = 1.$$

Usamos multiplicadores de Lagrange:

$$\mathcal{L}(\mathbf{w}, \lambda) = \mathbf{w}^\top \Sigma \mathbf{w} - \lambda(\mathbf{w}^\top \mathbf{w} - 1)$$

Derivando respecto a $\mathbf{w}$:

$$\nabla_{\mathbf{w}} \mathcal{L} = 2\Sigma\mathbf{w} - 2\lambda\mathbf{w} = 0 \implies \Sigma\mathbf{w} = \lambda\mathbf{w}.$$

Así, $\mathbf{w}$ es un eigenvector de Σ y λ es el eigenvalor.

Componentes principales

La matriz W de eigenvectores (ordenados por eigenvalores decrecientes) permite transformar los datos:

$$\mathbf{Y} = \mathbf{X}W$$

Cada columna de Y es un componente principal.

5. Rotación de Ejes con Matrices de Rotación

Geometría de PCA

PCA realiza una rotación ortogonal del sistema de coordenadas original. La matriz de rotación W es ortogonal: $W^{-1} = W^\top$.

Matriz de rotación en 2D

En 2D, una rotación por ángulo θ se representa como:

$$W = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

Los loadings son los elementos de W y representan los cosenos directores.

6. Ejemplos en Genómica y Competencia Empresarial

Genómica

Se tiene una matriz de expresión génica X de tamaño $n \times p$ (n muestras, p genes). Tras aplicar PCA:

- Los primeros componentes (PC1, PC2) permiten visualizar agrupamientos de muestras.
- Los loadings de PC1 indican los genes con mayor variabilidad.

Competencia empresarial

Se analizan 50 variables de clientes. PCA permite:

- Construir índices (PC1: riqueza", PC2: "tecnofilia").
- Segmentar clientes en el plano PC1-PC2.

7. Historia del Modelo PCA

- **1901**: Karl Pearson introduce los .[ej]es principales"para ajustar planos a datos.
- **1933**: Harold Hotelling formaliza y acuña el término çomponentes principales".
- **1950-60**: Uso en psicometría (Thurstone) y ciencias sociales.
- **1970**: Aplicaciones en ecología y quimiometría.
- **2000**: Base para técnicas como eigenfaces.

Bibliografía básica

1. Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.