Principal Component Analysis (PCA)

1 Introduction

Principal Component Analysis (PCA) is a dimensionality reduction technique that transforms data into a new coordinate system. The new coordinates, known as principal components, are orthogonal and capture the directions of maximum variance in the data. PCA is widely used for reducing the number of features in a dataset while preserving as much information as possible.

2 Mathematical Foundation

2.1 Standardization

PCA is sensitive to the scale of the data. Therefore, it is often necessary to standardize the data before applying PCA. For a dataset $\mathbf{X}$ with n samples and d features, the standardized data $\mathbf{X}_{\text{centered}}$ is computed as:

$$\mathbf{X}_{\text{centered}} = \mathbf{X} - \mu \tag{1}$$

where μ is the mean vector of the data.

2.2 Covariance Matrix

The covariance matrix $\mathbf{C}$ of the centered data is given by:

$$\mathbf{C} = \frac{1}{n-1}\mathbf{X}_{\text{centered}}^{\top}\mathbf{X}_{\text{centered}} \tag{2}$$

The covariance matrix captures the variance and correlation between different features.

2.3 Eigenvalue Decomposition

To perform PCA, we decompose the covariance matrix into its eigenvalues and eigenvectors:

$$\mathbf{C} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^{\top} \tag{3}$$

where $\mathbf{V}$ contains the eigenvectors (principal components) and $\mathbf{\Lambda}$ is a diagonal matrix with eigenvalues.

The eigenvectors represent the directions of maximum variance, and the eigenvalues represent the magnitude of variance in these directions.

2.4 Principal Components

The principal components are the eigenvectors corresponding to the largest eigenvalues. To reduce the dimensionality of the data, we select the top k principal components:

$$\mathbf{X}_{\text{pca}} = \mathbf{X}_{\text{centered}}\mathbf{V}_k \tag{4}$$

where $\mathbf{V}_k$ contains the top k eigenvectors.

2.5 Explained Variance

The amount of variance explained by each principal component is given by:

$$\text{Explained Variance}_i = \frac{\lambda_i}{\sum_{j=1}^{d} \lambda_j} \tag{5}$$

where λ_i is the i-th eigenvalue. The explained variance helps to understand how much information (variance) is retained by each principal component.

3 Applications

PCA is commonly used in various applications, including:

- Data visualization
- Noise reduction
- Feature extraction
- Preprocessing for machine learning algorithms