Principal Component Analysis (PCA)

Principal Component Analysis (PCA) adalah metode reduksi dimensi linier yang digunakan untuk mengekstraksi fitur-fitur penting dari data berdimensi tinggi dengan mempertahankan sebanyak mungkin varians.

Tujuan

Diberikan matriks data $X \in \mathbb{R}^{n \times d}$ dengan n sampel dan d fitur, PCA bertujuan menemukan proyeksi data ke ruang berdimensi lebih rendah k ($k \leq \min(n, d)$) yang memaksimalkan **varians** dari data hasil proyeksi.

Langkah-Langkah Algoritma

1. Sentralisasi Data

Data diubah agar memiliki mean nol:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i, \quad X_{\text{centered}} = X - \mu$$

2. Decomposisi SVD

Melakukan dekomposisi Singular Value Decomposition (SVD) pada data tersentralisasi:

$$X_{\text{centered}} = U \Sigma V^\top$$

- $U \in \mathbb{R}^{n \times n}$: matriks kiri ortonormal
- $\Sigma \in \mathbb{R}^{n \times d}$: diagonal singular values
- $V^\top \in \mathbb{R}^{d \times d}$: matriks kanan ortonormal

3. Principal Components

Komponen utama diperoleh dari V, yakni baris pertama hingga k dari $V^\top$:

$$\text{components\_} = V^\top[:k, :]$$

4. Transformasi ke Ruang Baru

Untuk mentransformasi data ke ruang baru:

$$Z = X_{\text{centered}} \cdot V_k^\top$$

di mana $V_k^\top \in \mathbb{R}^{k \times d}$ adalah matriks komponen utama.

5. Varians dan Rasio Varians

Varians yang dijelaskan oleh tiap komponen utama dihitung dari singular values:

$$\lambda_j = \frac{\sigma_j^2}{n-1}, \quad \text{explained\_variance\_ratio} = \frac{\lambda_j}{\sum_{i=1}^{d} \lambda_i}$$