

## Principal Component Analysis (PCA)

PCA (Principal Component Analysis) adalah metode statistik yang digunakan untuk mereduksi dimensi data dengan mempertahankan sebanyak mungkin informasi yang relevan. Proses ini dapat membantu kita memahami struktur dalam data yang kompleks dengan mengidentifikasi pola-pola yang tersembunyi di dalamnya. PCA dapat dijelaskan secara konseptual dengan menguraikan langkah-langkahnya menggunakan dekomposisi nilai singular (SVD). SVD adalah teknik matematis yang mendasari PCA untuk memecah data menjadi komponen-komponen utama.

Mari kita mulai dengan pemahaman dasar tentang PCA. Jika kita mempertimbangkan data yang hanya memiliki dua gen, kita dapat memplotnya pada grafik dua dimensi. Gen 1 akan menjadi sumbu x dan Gen 2 akan menjadi sumbu y. Titik-titik pada grafik mewakili sampel-sampel yang diukur. Dalam konteks ini, PCA akan membantu kita memahami gen mana yang paling penting dalam mengelompokkan data. Misalnya, PCA dapat menunjukkan bahwa Gen 3 berperan penting dalam memisahkan sampel di sepanjang sumbu x.

Langkah pertama dalam PCA adalah memusatkan data pada titik asal grafik, sehingga nilai rata-rata dari setiap gen menjadi nol. Kemudian, PCA mencari garis yang paling cocok untuk memproyeksikan data. Garis ini merupakan komponen utama pertama (PC1), yang mencerminkan arah dengan variasi terbesar dalam data. Proyeksi data ke PC1 membantu mengurangi dimensi data. Proses ini dilakukan secara iteratif untuk menemukan komponen utama berikutnya (PC2, PC3, dan seterusnya), yang tegak lurus satu sama lain dan memperjelas pola-pola yang lebih kompleks dalam data.

Penilaian kualitas komponen utama dapat dilakukan dengan menggunakan nilai eigen. Nilai eigen mewakili jumlah dari kuadrat jarak antara titik proyeksi dan titik asal untuk setiap komponen utama. Semakin besar nilai eigen, semakin banyak variasi yang dijelaskan oleh komponen utama tersebut. Dengan memahami konsep-konsep ini, kita dapat menggambar plot PCA dua dimensi yang menggambarkan bagaimana sampel-sampel berada dalam ruang berdasarkan komponen utama utama (misalnya, PC1 dan PC2). Ini membantu kita memvisualisasikan pola-pola yang mungkin sulit dilihat dalam data asli.

Proses PCA serupa untuk data tiga dimensi, hanya saja dengan menambahkan dimensi ekstra, dan penggunaan PC1, PC2, dan PC3 untuk mereduksi dimensi dan memahami struktur dalam data. Dengan memanfaatkan nilai eigen, kita dapat menilai kontribusi relatif dari setiap komponen utama terhadap variasi dalam data.

Dengan demikian, PCA merupakan alat yang kuat untuk mereduksi dimensi data, memahami pola-pola yang tersembunyi, dan mengidentifikasi variabel yang paling penting dalam mengelompokkan data.