

Rangkuman PCA

PCA (Principal Component Analysis) adalah sebuah metode untuk mengurangi dimensi dataset yang kompleks dengan menggunakan Singular Value Decomposition (SVD). Metode ini berguna dalam mengidentifikasi pola dan hubungan antara variabel dalam dataset yang besar. PCA memungkinkan kita untuk mengambil banyak pengukuran gen dan menghasilkan plot PCA dua dimensi. Ini membantu mengelompokkan data dan menyoroti gen mana yang paling berkontribusi terhadap perbedaan antar sampel. Proses perhitungan PCA melibatkan penghitungan pusat data dan memproyeksikannya ke garis acak, di mana garis yang paling sesuai ditemukan dengan meminimalkan jumlah kuadrat jarak dari titik proyeksi ke titik asal.

Tujuan utamanya adalah untuk mengidentifikasi pola tersembunyi dalam data dengan mereduksi jumlah dimensi dari dataset yang besar, sehingga memudahkan analisis lebih lanjut.

PCA Untuk Data 2-Dimensi

PCA membantu mengurangi dimensi data ke 2 dimensi untuk memvisualisasikan pola atau keterkaitan antara sampel. Proyeksi data ke garis yang sesuai kemudian dihitung untuk menemukan pola. Dimulai dengan melakukan plotting data, setelah itu dilakukan perhitungan rata-rata untuk tiap variabel. Melalui rata-rata tersebut, kita melakukan perhitungan untuk mencari titik tengah dari data. Setelah itu dilakukan pergeseran data dengan titik tengah menjadi titik (0,0). Setelah itu, kita dapat membuat garis yang sesuai untuk data tersebut dengan cara membuat garis acak dari garis asal dan melakukan perputaran sehingga garis paling sesuai dengan data tersebut.

Komponen Utama (PC1)

Komponen Utama pertama (PC1) adalah vektor eigen yang paling signifikan yang menjelaskan variasi terbesar dalam data. Ini dihitung dengan memproyeksikan data ke garis yang memaksimalkan variasi, atau minimalisasi

jarak antara data dan garis tersebut. Secara matematis, PC1 adalah vektor eigen yang sesuai dengan eigen value terbesar dari matriks kovariansi data. Vektor eigen ini mengarah ke arah di mana variasi data terbesar terjadi. Dengan memproyeksikan data ke dalam arah vektor eigen PC1, kita mendapatkan proyeksi data yang menjelaskan variasi terbesar dalam satu dimensi.

Komponen Utama Kedua (PC2)

Komponen Utama berikutnya (PC2) adalah garis terbaik yang tegak lurus dengan PC1. Ini membantu menangkap variasi tambahan dalam data yang tidak dijelaskan oleh PC1. Dengan memiliki PC2, kita dapat menangkap lebih banyak variasi dalam data yang tidak dapat dijelaskan oleh PC1 saja

Plot PCA

Plot PCA akhir dibuat dengan memutar PC1 menjadi sumbu horizontal dan PC2 menjadi sumbu vertikal. Titik-titik data kemudian diproyeksikan ke plot ini untuk menemukan lokasi sampel di dalamnya. Ini membantu kita memvisualisasikan sebaran data dalam ruang dua dimensi yang ditentukan oleh PC1 dan PC2.

Vektor Eigen adalah vektor yang tidak berubah arah ketika dioperasikan oleh suatu transformasi linier, kecuali dikalikan dengan skalar. Dalam PCA, vektor eigen dari matriks kovariansi data menunjukkan arah di mana variasi dalam data terbesar. Komponen utama atau principal components dari data dihitung dari vektor eigen ini.

Nilai Eigen adalah skalar yang menunjukkan besarnya variasi yang dijelaskan oleh vektor eigen yang sesuai. Nilai eigen mewakili jumlah variasi dalam data yang dapat dijelaskan oleh komponen utama yang bersangkutan. Nilai eigen diurutkan dalam urutan menurun, sehingga nilai eigen yang lebih tinggi menunjukkan komponen utama yang lebih signifikan.