Rangkuman Video Materi PCA

Axel David, 1103210017, TK4504

A. Motivasi Konseptual Untuk PCA

PCA adalah teknik statistik yang digunakan untuk mengurangi dimensi dari dataset yang kompleks, dengan tujuan untuk memahami struktur data dengan cara yang lebih sederhana. Metode ini berguna dalam mengidentifikasi pola dan hubungan antara variabel dalam dataset yang besar.

Tujuan utamanya adalah untuk mengidentifikasi pola tersembunyi dalam data dengan mereduksi jumlah dimensi dari dataset yang besar, sehingga memudahkan analisis lebih lanjut. PCA mencari proyeksi linear dari data asli ke ruang dimensi yang lebih rendah sehingga variabilitas maksimal terjelaskan oleh sejumlah komponen utama atau "principal components" yang dibuat. Dengan cara ini, PCA dapat membantu mengidentifikasi pola-pola yang ada dalam data dengan cara yang lebih sederhana dan mudah dipahami.

B. 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. PCA dapat diterapkan pada data 2 dimensi dengan tujuan untuk mengurangi dimensi tersebut menjadi satu dimensi atau bahkan lebih rendah. Misalnya, jika Kita memiliki kumpulan data 2 dimensi yang berisi titik-titik (x, y), maka Kita dapat menggunakan PCA untuk mengurangi dimensi tersebut menjadi satu dimensi, yaitu sumbu utama (principal component) yang paling signifikan dalam variasi data.

Dengan mengurangi dimensi data dari 2D menjadi 1D, kita dapat dengan mudah memvisualisasikan data dalam ruang satu dimensi atau bahkan menggunakan sumbu utama ini untuk menganalisis pola atau kecenderungan dalam data. Secara umum, PCA membantu kita untuk memahami variasi dan struktur data dengan cara yang lebih sederhana dan lebih terkonsentrasi, sehingga kita dapat mengidentifikasi pola atau hubungan yang tersembunyi dalam data 2 dimensi tersebut.

C. Mencari 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 eigenvalue 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.

Mencari vektor eigen PC1 melibatkan proses perhitungan nilai dan vektor eigen dari matriks kovariansi data, dan kemudian memilih vektor eigen yang sesuai dengan eigenvalue terbesar. Setelah vektor eigen PC1 ditemukan, data dapat diproyeksikan ke dalam dimensi yang lebih rendah menggunakan vektor tersebut sebagai sumbu. Proyeksi data ini akan memberikan representasi satu dimensi dari data yang mempertahankan sebagian besar variasi atau informasi yang ada dalam data asli.

Dengan menggunakan PC1, kita dapat menemukan arah di mana variasi terbesar dalam data terjadi, yang dapat membantu dalam pemahaman dan analisis lebih lanjut tentang struktur dan pola dalam dataset tersebut.

D. Vektor dan Nilai Eigen, dan Skor Loading

PCA melibatkan konsep vektor dan nilai eigen, yang menggambarkan arah dan besar variasi dalam data. Loading scores menunjukkan kontribusi variabel terhadap setiap komponen utama. Dengan menggunakan konsep vektor eigen, nilai eigen, dan skor loading, PCA memungkinkan kita untuk mengurutkan dan memahami variasi dalam data, serta mengidentifikasi komponen utama yang paling signifikan dan kontribusi variabel terhadap komponen tersebut. Ini membantu dalam mereduksi dimensi data, analisis pola, dan pemahaman lebih lanjut tentang struktur data yang kompleks.

1. **Vektor Eigen**
   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.
2. **Nilai Eigen**
   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.
3. **Skor Loading**
   Skor loading menunjukkan kontribusi variabel (fitur) terhadap setiap komponen utama. Skor loading menunjukkan seberapa kuat suatu variabel berhubungan dengan komponen utama tertentu. Semakin tinggi nilai loading score, semakin besar kontribusi variabel tersebut terhadap komponen utama. Skor loading ini digunakan untuk memahami struktur komponen utama dan membantu dalam interpretasi hasil PCA.

E. Mencari Komponen Utama Kedua (PC2)

PC2 adalah komponen utama kedua yang menjelaskan variasi kedua terbesar dalam data. Data ni ditemukan setelah PC1, dengan mempertahankan ortogonalitas terhadap PC1. Setelah kita menemukan Komponen Utama pertama (PC1), langkah selanjutnya adalah mencari Komponen Utama kedua (PC2). PC2 adalah komponen utama kedua yang menjelaskan variasi kedua terbesar dalam data. Penting untuk dicatat bahwa PC2 harus

dipertahankan ortogonal terhadap PC1. Artinya, arah PC2 harus tegak lurus atau saling tegak lurus terhadap arah PC1.

F. Menggambar Grafik PCA

Grafik PCA menggambarkan distribusi data dalam dimensi yang lebih rendah (biasanya 2 dimensi) berdasarkan komponen utama yang dipilih. Grafik PCA adalah cara visual yang berguna untuk memahami distribusi data dalam dimensi yang lebih rendah, khususnya 2 dimensi, berdasarkan komponen utama (principal components) yang telah dipilih dari analisis PCA. Grafik ini membantu untuk menggambarkan bagaimana data tersebar dalam ruang yang lebih sederhana dan mudah dipahami dan mendapatkan wawasan tentang struktur data dalam dimensi yang lebih rendah serta mengidentifikasi pola atau hubungan yang mungkin tersembunyi dalam data tersebut.

G. Menghitung Persentase Variasi Setiap PC dan Plot Skrin

Setelah menemukan komponen utama, persentase varians yang dijelaskan oleh setiap komponen dihitung. Plot skrin digunakan untuk memvisualisasikan persentase varians yang dijelaskan oleh setiap komponen dan memutuskan berapa banyak dimensi yang diperlukan untuk mewakili data. Untuk menghitung persentase variansi yang dijelaskan oleh setiap komponen utama (PC), Kita dapat menggunakan nilai eigen dari masing-masing PC. Persentase variansi yang dijelaskan oleh setiap PC dihitung dengan membagi nilai eigen PC tersebut dengan jumlah total nilai eigen dari semua PC.

Grafik batang yang dihasilkan akan menunjukkan persentase variansi yang dijelaskan oleh setiap komponen utama. Dengan melihat plot skrin ini, Kita dapat memutuskan berapa banyak dimensi yang diperlukan untuk mewakili data. Biasanya, Kita dapat memilih jumlah PC yang memiliki persentase variansi yang cukup tinggi sehingga mempertahankan sebagian besar informasi dalam data, tetapi juga meminimalkan dimensi yang digunakan.

H. PCA untuk Data 3-Dimensi

PCA juga dapat diterapkan untuk data yang memiliki lebih dari 2 dimensi, seperti data 3 dimensi. Prosesnya mirip dengan PCA untuk data 2 dimensi, tetapi melibatkan pemilihan lebih banyak komponen utama untuk mengurangi dimensi.

Dengan menggunakan PCA pada data 3 dimensi, Anda dapat mengurangi dimensi data tersebut sehingga lebih mudah dipahami dan diinterpretasikan, sambil mempertahankan sebagian besar informasi yang relevan. Ini membantu dalam pemahaman dan analisis data yang kompleks dalam konteks yang lebih sederhana. Berikut adalah langkah-langkah umum untuk menerapkan PCA pada data 3 dimensi:

1. Standarisasi data: Seperti pada PCA untuk data 2 dimensi, langkah pertama adalah standarisasi data dengan menghitung mean dari setiap dimensi dan menstandarkan data dengan membagi setiap nilai dengan deviasi standarnya.
2. Menghitung matriks kovariansi: Hitung matriks kovariansi dari data standar.
3. Menghitung nilai-nilai dan vektor-vektor eigen: Hitung nilai-nilai eigen dan vektor-vektor eigen dari matriks kovariansi. Anda akan mendapatkan lebih banyak nilai eigen dan vektor eigen untuk data 3 dimensi daripada untuk data 2 dimensi.

4. Pemilihan komponen utama: Pilih komponen utama (principal components) yang akan digunakan untuk mengurangi dimensi data. Anda dapat memilih sejumlah komponen utama berdasarkan nilai-nilai eigen tertinggi yang menjelaskan sebagian besar varians dalam data.
5. Proyeksi data ke komponen utama: Proyeksikan data ke dalam ruang yang lebih rendah berdasarkan komponen utama yang dipilih. Ini akan mengurangi dimensi data dari 3 dimensi menjadi jumlah komponen utama yang dipilih.
6. Analisis lanjutan: Lakukan analisis lebih lanjut terhadap data yang telah direduksi dimensinya, seperti visualisasi, clustering, atau pembuatan model.