Statistika Komputasi

SSD23438

Egi Safitri, S.Mat., M.Si

Pengantar Analisis Komponen Utama (PCA)

Principal Component Analysis (PCA) adalah teknik statistik yang digunakan untuk mengurangi dimensi data dengan tujuan menyederhanakan data tanpa kehilangan informasi yang signifikan.

- Mengurangi Dimensi Data: PCA membantu mereduksi data berdimensi tinggi (memiliki banyak variabel) menjadi data berdimensi lebih rendah sambil mempertahankan variasi sebanyak mungkin dalam dataset. Ini berguna karena data berdimensi tinggi sering kali sulit untuk divisualisasikan dan dapat menyebabkan overfitting dalam model machine learning.
- Identifikasi Pola Variasi: PCA mengidentifikasi arah (komponen utama) di mana data memiliki variasi terbesar. Komponen utama pertama menangkap sebagian besar variasi dalam data, sedangkan komponen-komponen berikutnya menangkap variasi yang tersisa dengan urutan menurun. Dengan demikian, kita dapat merepresentasikan data dengan lebih sedikit komponen tanpa kehilangan informasi penting.

Konsep Dasar PCA

- **Varians**: PCA dirancang untuk menangkap varians maksimum dalam data. Komponen utama pertama (PC1) selalu memiliki varians terbesar dibandingkan komponen lainnya, dan komponen-komponen berikutnya menangkap varians yang tersisa dengan urutan menurun.
- **Orthogonalitas**: Komponen utama yang dihasilkan oleh PCA saling orthogonal (tegak lurus) satu sama lain. Hal ini memastikan setiap komponen utama menangkap variasi unik dalam data tanpa ada redundansi.
- **Reduksi Dimensi**: PCA digunakan untuk mereduksi dimensi data dengan memilih sejumlah komponen utama dengan varians terbanyak. Ini memungkinkan kita mengurangi kompleksitas data sambil mempertahankan informasi yang paling relevan.

Perhitungan Komponen Utama

Langkah-langkah:

1. Standarisasi Data/Normalisasi : Data sering dinormalisasi sehingga setiap fitur memiliki skala yang sama
2. Matriks Kovarians :Matriks kovarians dihitung untuk memahami hubungan antar variabel.
3. Vektor Eigen dan Nilai Eigen: PCA kemudian mencari eigenvalues dan eigenvectors dari matriks kovarians. Eigenvectors menentukan arah komponen utama, sementara eigenvalues mengukur pentingnya atau kekuatan variasi setiap komponen utama.
4. Proyeksi Data: Data asli diproyeksikan ke ruang baru yang dibentuk oleh komponen utama.

Interpretasi Komponen Utama

Memahami Arti dan Signifikansi Setiap Komponen:

- Komponen utama adalah kombinasi linear dari variabel asli yang dibuat oleh PCA. Setiap komponen utama (principal component) menangkap pola variasi dalam data.
- Komponen utama pertama (PC1) selalu menangkap variasi terbesar dalam data, sedangkan komponen utama kedua (PC2) menangkap variasi terbesar berikutnya yang tidak berkorelasi dengan PC1, dan seterusnya.
- Signifikansi setiap komponen diukur menggunakan eigenvalues, yang menunjukkan seberapa besar variasi dalam data dijelaskan oleh masing-masing komponen. Komponen dengan eigenvalues yang lebih tinggi dianggap lebih signifikan karena menangkap lebih banyak informasi dari data asli.

Interpretasi Komponen Utama

Nama Komponen Berdasarkan Bobot Variabel Tertinggi:

- Setiap komponen utama adalah kombinasi dari variabel asli dengan bobot atau koefisien tertentu. Bobot ini menunjukkan kontribusi atau pengaruh setiap variabel terhadap komponen utama tersebut.
- Untuk memahami makna setiap komponen, kita melihat bobot variabel tertinggi. Misalnya, jika PC1 memiliki bobot tertinggi pada variabel "age"dan "income,"PC1 mungkin mewakili faktor yang terkait dengan usia dan pendapatan.
- Dengan mengidentifikasi variabel yang memiliki bobot terbesar, kita bisa memberikan interpretasi atau label yang bermakna pada komponen utama. Ini membantu memahami aspek data mana yang diwakili oleh masing-masing komponen.

Aplikasi PCA dalam Reduksi Dimensi

- **Visualisasi Data**: PCA digunakan untuk mengurangi dimensi data sehingga dapat divisualisasikan dalam 2 atau 3 dimensi. Hal ini membantu dalam memahami pola, tren, atau distribusi data yang kompleks.
- **Pengelompokan**: Dengan mereduksi dimensi data, PCA membantu mengidentifikasi kelompok atau klaster dalam data, mempermudah analisis klaster dan pengenalan pola.
- **Prediksi**: PCA dapat digunakan sebagai langkah pra-pemrosesan dalam machine learning untuk mengurangi fitur yang kurang relevan, sehingga meningkatkan kinerja model prediksi dan mengurangi risiko overfitting.

Pemilihan Jumlah Komponen Utama

- **Scree Plot**: Grafik yang menunjukkan nilai eigen dari setiap komponen utama dalam urutan menurun. Titik "elbow"pada grafik membantu menentukan jumlah komponen yang optimal dengan mempertahankan komponen yang berada sebelum titik tersebut.
- **Aturan Kaiser**: Mempertahankan komponen utama dengan nilai eigen lebih dari 1, karena komponen ini dianggap memiliki kontribusi yang signifikan terhadap varians data.
- **Uji Nilai Eigen**: Menganalisis nilai eigen dari setiap komponen utama. Komponen dengan nilai eigen tinggi lebih signifikan dalam menjelaskan variasi data, sedangkan komponen dengan nilai eigen rendah dapat diabaikan.

Kelebihan PCA

- **Reduksi Dimensi**: PCA membantu mereduksi jumlah variabel dalam data berdimensi tinggi, membuat analisis lebih sederhana tanpa kehilangan terlalu banyak informasi penting.
- **Visualisasi Lebih Baik**: Dengan mengurangi dimensi data menjadi 2 atau 3 dimensi, PCA memungkinkan visualisasi data yang kompleks menjadi lebih mudah dipahami.
- **Meningkatkan Performa Model**: Dengan menghilangkan fitur yang kurang relevan, PCA dapat meningkatkan performa model machine learning dan mengurangi risiko overfitting.

Kekurangan PCA

- **Interpretasi Sulit**: Komponen utama adalah kombinasi linier dari variabel asli, sehingga lebih sulit untuk diinterpretasikan dibandingkan variabel asli.
- **Rentan Terhadap Outlier**: Outlier dapat sangat memengaruhi hasil PCA karena dapat mengubah arah komponen utama.
- **Asumsi Linearitas**: PCA mengasumsikan hubungan antara variabel adalah linear, sehingga kurang efektif jika data memiliki hubungan non-linear.

Persiapan Data untuk PCA

- **Pembersihan Data**: Langkah awal dalam persiapan data adalah menghapus data yang tidak lengkap, duplikat, atau data yang tidak relevan untuk memastikan kualitas analisis.
- **Pengolahan Data yang Hilang**: Mengatasi data yang hilang dengan menghapus baris/kolom yang tidak lengkap atau menggunakan imputasi untuk mengganti nilai yang hilang, misalnya dengan rata-rata atau median.
- **Penskalaan Data**: Menyelaraskan skala data menggunakan teknik standardisasi atau normalisasi untuk memastikan bahwa variabel dengan rentang nilai berbeda tidak mendominasi analisis PCA.

Implementasi PCA dengan Perangkat Lunak

- **R**: Bahasa pemrograman statistik
- **Python**: Digunakan dalam analisis data dan machine learning

Data Asli

Penjelasan:

- Tabel ini menunjukkan data asli dari empat karyawan dengan dua fitur yang diukur, yaitu tinggi badan dalam sentimeter (cm) dan berat badan dalam kilogram (kg).
- Kolom "Karyawan"mewakili ID setiap karyawan.
- Kolom "Tinggi (cm)"berisi data tinggi badan karyawan.
- Kolom "Berat (kg)"berisi data berat badan karyawan.
- Data ini akan menjadi input awal untuk analisis PCA, di mana setiap fitur akan dianalisis untuk menentukan variasi yang paling signifikan.

Karyawan	Tinggi (cm)	Berat (kg)
1	160	55
2	170	65
3	180	75
4	190	85

Standarisasi Data

Formula Z-score:

$$z = \frac{x - \mu}{\sigma}$$

Rata-rata dan standar deviasi untuk tinggi dan berat:

- $\mu_{\text{Tinggi}} = 175$, $\sigma_{\text{Tinggi}} = 11.18$
- $\mu_{\text{Berat}} = 70$, $\sigma_{\text{Berat}} = 11.18$

Standarisasi Data (Lanjutan)

Penjelasan:

- Tabel ini menunjukkan nilai Z-score untuk tinggi dan berat masing-masing karyawan.
- Z-score mengukur seberapa jauh nilai asli dari rata-rata dalam satuan standar deviasi.
- Nilai negatif menunjukkan bahwa data berada di bawah rata-rata, sementara nilai positif menunjukkan bahwa data berada di atas rata-rata.
- Sebagai contoh, Karyawan 1 memiliki Z-score -1.34 untuk tinggi dan berat, yang berarti tinggi dan beratnya 1.34 standar deviasi di bawah rata-rata.
- Standarisasi seperti ini penting dalam PCA karena memastikan semua variabel berada pada skala yang sama sebelum melakukan analisis.

Karyawan	Z-Tinggi	Z-Berat
1	-1.34	-1.34
2	-0.45	-0.45
3	0.45	0.45
4	1.34	1.34

Matriks Kovarian

Rumus matriks kovarian:

$$\text{cov}(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Matriks kovarian:

$$\begin{pmatrix} 1.0 & 1.0 \\ 1.0 & 1.0 \end{pmatrix}$$

Nilai Eigen dan Vektor Eigen

Nilai Eigen (λ):

$$\lambda_1 = 2, \quad \lambda_2 = 0$$

Vektor Eigen:

$$\mathbf{v_1} = \begin{pmatrix} 0.7071 \\ 0.7071 \end{pmatrix}, \quad \mathbf{v_2} = \begin{pmatrix} -0.7071 \\ 0.7071 \end{pmatrix}$$

Proyeksi ke Komponen Utama

Proyeksi ke PC1 (Karyawan 1):

$$PC1_1 = (-1.34 \times 0.7071) + (-1.34 \times 0.7071) = -1.89$$

Proyeksi ke PC1 untuk semua karyawan:

Karyawan	PC1
1	-1.89
2	-0.63
3	0.63
4	1.89

Interpretasi Hasil PCA

- Komponen utama pertama (PC1) menangkap variabilitas terbesar dalam data (100% dari total variansi).
- Komponen utama kedua (PC2) tidak menangkap variabilitas yang signifikan (variansi = 0).
- Data berhasil direduksi dari 2 dimensi menjadi 1 dimensi (PC1) tanpa kehilangan informasi penting.

Contoh Kasus Penerapan PCA

1. **Analisis Data Genetik**:
   - PCA digunakan untuk mengidentifikasi pola dan variasi dalam data genetik yang memiliki ribuan hingga jutaan variabel.
   - Membantu mereduksi dimensi data genetik menjadi beberapa komponen utama untuk visualisasi dan analisis yang lebih mudah.
2. **Pengenalan Pola**:
   - PCA digunakan dalam pengenalan wajah, sidik jari, atau objek dalam gambar dengan mereduksi data menjadi fitur-fitur penting.
   - Membantu sistem pengenalan pola dalam membedakan individu atau objek secara efisien dengan data yang lebih sederhana.
3. **Analisis Keuangan**:
   - PCA mereduksi data keuangan yang memiliki banyak variabel menjadi beberapa komponen utama yang mewakili variabilitas pasar.
   - Mempermudah analisis tren, identifikasi risiko, dan visualisasi korelasi antara aset keuangan.

Kesimpulan

PCA adalah metode yang sangat berguna dalam analisis data untuk reduksi dimensi, pengelompokan, dan prediksi. Pemahaman yang baik tentang cara kerja PCA akan membantu dalam pengolahan dan interpretasi data yang kompleks.

Thank You!

Institut Informatika & Bisnis
DARMAJAYA
Yayasan Alfian Husin