Rizqika Mulia Pratama
13522126

PCA

1. Cara kerja algoritma

Algoritma Principal COmponent Analysis (PCA) adalah metode machine learning yang digunakan untuk reduksi dimensi data, yang bertujuan untuk mengurangi jumlah variabel dalam dataset sambil mempertahankan sebanyak mungkin informasi yang ada. PCA sering digunakan dalam EDA dan preprocessing sebelum menerapkan algoritma machine learning lainnya.

1. Standardisasi data

   Langkah pertama dalam PCA adalah menstandarisasi data. Ini dilakukan untuk memastikan bahwa setiap fitur memiliki skala yang sama, sehingga tidak ada fitur yang mendominasi hasil analisis. Standardisasi biasanya dilakukan dengan mengurangi rata-rata dan membagi dengan standar deviasi untuk setiap fitur.

2. Menghitung matrix covariance

   Setelah data distandarisasi, langkah berikutnya adalah menghitung matrix covariance. Matriks ini menunjukkan bagaimana dua variabel berhubungan satu sama lain. Jika variabel-variabel tersebut berkorelasi, maka nilai-nilai dalam matrix covariance akan menunjukkan hubungan tersebut.

3. Menghitung Eigenvalue dan eigenvector

   Dari matrix covariance, kita kemudian menghitung eigenvalue dan eigenvector. Eigenvector menunjukkan arah dari varians maksimum dalam data, sedangkan eigenvalue menunjukkan seberapa besar variansi yang ada pada arah tersebut.

4. Memilih komponen utama

   Setelah mendapatkan eigenvector dan eigenvalue, langkah selanjutnya adalah memilih sejumlah komponen utama yang akan digunakan untuk merepresentasikan data. Biasanya, kita memilih komponen dengan eigenvalue terbesar yang menjelaskan sebagian besar variansi dalam data

5. Transformasi data

   Langkah terakhir adalah mentransformasikan data asli ke dalam ruang baru yang dibentuk oleh komponen utama yang dipilih. Ini dilakukan dengan ruang baru yang dibentuk oleh komponen utama yang dipilih. Ini dilakukan dengan mengalikan data yang telah distandarisasi dengan matriks eigenvector yang dipilih.

6. Evaluasi

   Setelah transformasi, kita dapat mengevaluasi hasil PCA dengan memvisualisasikan data dalam dimensi yang lebih rendah. Ini membantu dalam memahami struktur data dan mengidentifikasi pola atau cluster.

Rizqika Mulia Pratama
13522126

2. Perbandingan model buatan sendiri dan library

Berdasarkan hasil evaluasi, model dari library scikit learn memperoleh score train dan test mean squared error di kisaran angka 7 x 10^-31, sementara untuk model manual yang dibangun dari scratch, mendapat score di kisaran 1 x 10^-30. Ini menunjukkan bahwa model PCA scikit learn lebih baik dalam merekonstruksi data asli setelah reduksi dimensi, yang berarti bahwa model tersebut lebih akurat dalam menangkap variansi dalam data. Hal ini kemungkinan disebabkan oleh adanya perbedaan kecil implementasi algoritma dalam cara perhitungan yang dilakukan, terutama dalam hal presisi numerik atau cara pengolahan data. Misalnya, penggunaan np.cov dan np.linalg.eig harus dilakukan dengan hati-hati untuk memastikan bahwa hasilnya konsisten dengan scikit-learn.