Nama: Loaeza Septavial

NIM: 1103204003

Kelas: Machine Learning

Principal Component Analysis (PCA)

Analisis komponen utama (PCA) atau analisis komponen utama adalah teknik statistik yang mereduksi dimensi data dengan tujuan mengidentifikasi pola dan hubungan penting dalam data. Hal ini dilakukan dengan mentransformasikan data asli ke dalam sistem koordinat baru. Sumbu-sumbu tersebut disusun sedemikian rupa sehingga komponen pertama (komponen utama) memuat variasi data paling banyak, dan komponen kedua memuat variasi berikutnya paling banyak.

Langkah-langkah umum untuk PCA adalah:

1. Standarisasi data: Karena data asli seringkali mempunyai skala yang berbeda-beda, maka harus melakukan normalisasi atau standarisasi data terlebih dahulu agar semua variabel mempunyai dampak yang sama.
2. Menghitung matriks kovarians: PCA menghitung matriks kovarians dari data standar. Matriks kovarians mengukur hubungan antar variabel dalam data. Menghitung vektor eigen dan nilai eigen: Langkah selanjutnya adalah menghitung vektor eigen dan nilai eigen dari matriks kovarians. Nilai eigen mengukur besarnya variasi yang dapat dijelaskan oleh setiap komponen.
3. Pemilihan komponen utama: Komponen utama adalah vektor eigen yang sesuai dengan nilai eigen terbesar. Komponen ini biasanya menjelaskan variasi terbanyak dalam data.
4. Transformasi data: Data asli diubah (ditransformasikan) menjadi sistem koordinat baru yang terdiri dari komponen-komponen utama. Hal ini memungkinkan data dideskripsikan dalam dimensi yang lebih rendah.
5. Menganalisis hasil: Hasil PCA digunakan untuk memahami pola data. Misalnya, dapat melihat bagaimana variabel asli berkontribusi pada komponen utama atau mengidentifikasi pola hubungan antar variabel.

PCA berguna dalam berbagai bidang seperti analisis data, pengenalan pola, dan kompresi data. Hal ini mengurangi kompleksitas data sekaligus mempertahankan sebagian besar informasi penting.

StatQuest: K-nearest neighbors

K-Nearest Neighbors (K-NN) adalah salah satu algoritma pembelajaran mesin dasar yang digunakan untuk klasifikasi dan regresi. Algoritma ini didasarkan pada gagasan bahwa objek-objek serupa cenderung termasuk dalam kelompok yang sama. Berikut penjelasan konsep K-NN:

1. Data dan Label: K-NN mengolah data yang mempunyai atribut (karakteristik seperti tinggi badan, berat badan, dll) dan label (kelas yang nilainya diprediksi atau ditargetkan).
2. K: Parameter yang sangat penting dari K-NN adalah 'K'. Inilah jumlah tetangga terdekat yang harus diperhatikan saat membuat prediksi. Nilai K yang pilih mempengaruhi hasil prediksi.

3. Menghitung jarak: K-NN mengukur jarak antara data yang ingin diprediksi dengan data lain dalam kumpulan data. Hal ini dapat dilakukan dengan menggunakan berbagai metrik jarak seperti jarak Euclidean atau jarak Manhattan.
4. Menentukan tetangga terdekat: Setelah menghitung jarak, algoritma K-NN memilih K tetangga terdekat dari data yang akan diprediksi.
5. Klasifikasi atau Regresi: Jika tujuan adalah klasifikasi, K-NN menanyakan tetangga terdekat untuk menentukan label kelas yang akan ditetapkan ke data yang diprediksi. Jika targetnya regresi, K-NN menghitung mean atau median dari label tetangga terdekat untuk memprediksi nilai target.
6. Menentukan nilai K: Memilih nilai K yang benar merupakan langkah penting dalam K-NN. Jika nilai K terlalu kecil, model mungkin rentan terhadap outlier, dan jika nilai K terlalu besar, model mungkin terlalu umum.

Video ini mungkin berisi contoh grafis dan ilustrasi untuk membantu lebih memahami konsep K-NN. Selain itu, Josh Starmer sering menggunakan bahasa yang jelas dan sederhana dalam video StatQuest-nya untuk membantu memahami statistik dan konsep pembelajaran mesin. Untuk penjelasan lebih detail dan jelas mengenai K-NN, kami sarankan untuk menonton videonya.

Decision and Classification Trees

Pohon keputusan dan pohon klasifikasi merupakan alat penting dalam analisis data yang digunakan untuk pengambilan keputusan dan klasifikasi data. Ini adalah model berbasis aturan yang dapat memahami informasi yang tersedia dan membuat keputusan.

Pohon keputusan:

Pohon keputusan adalah model grafis yang digunakan untuk memvisualisasikan dan memahami serangkaian keputusan yang dibuat berdasarkan kondisi atau karakteristik tertentu dalam data. Pohon ini terdiri dari simpul atau “simpul” yang mewakili keputusan, cabang yang menghubungkan simpul-simpul tersebut, dan “daun” yang mewakili hasil akhir atau keputusan.

Cara Penggunaan:

Pohon dimulai dengan simpul akar yang mewakili seluruh kumpulan data. Setiap node mengajukan pertanyaan atau membuat pernyataan berdasarkan karakteristik data. Data tersebut kemudian dibagi menjadi dua cabang berdasarkan jawaban pertanyaan. Proses ini diulangi pada setiap node hingga node daun tercapai dan keputusan akhir dibuat. Contoh: Misalnya ingin memutuskan apakah pelanggan harus membeli suatu produk berdasarkan faktor seperti usia, pendapatan, dan preferensi. Pohon keputusan membantu memahami bagaimana faktor-faktor ini memengaruhi keputusan pelanggan.

Pohon klasifikasi:

Pohon klasifikasi adalah jenis pohon keputusan yang digunakan untuk memprediksi kategori atau label dari data. Ini sering digunakan dalam masalah klasifikasi dimana data diklasifikasi ke dalam kategori tertentu berdasarkan karakteristiknya.

Cara Penggunaan:

Pohon dimulai dengan simpul akar yang mewakili seluruh kumpulan data. Di setiap node, algoritme mencari fitur yang paling baik membagi data ke dalam kategori berdasarkan metrik seperti pengotor Gini dan entropi. Berdasarkan fitur terbaik tersebut, data dibagi menjadi dua cabang. Proses ini diulangi untuk setiap node hingga kita mencapai node daun yang mewakili label kelas. Contoh: Dalam klasifikasi spam email, pohon klasifikasi membantu memahami bagaimana karakteristik seperti kata kunci, panjang pesan, dan pengirim memengaruhi keputusan apakah suatu email termasuk spam atau bukan. Masu.

Pohon keputusan dan pohon klasifikasi adalah alat yang membantu menganalisis data untuk mendeskripsikan aturan yang tersedia dalam data dan membuat keputusan berdasarkan aturan tersebut. Ini membantu memahami hubungan antar variabel dan membuat prediksi serta klasifikasi berdasarkan informasi yang ada.