Nama: Irawan Mardiansyah

Kelas: TK-45-04

NIM: 1103210174

Lecture: 5 (Membuat Catatan PCA)

Dalam video "StatQuest: Principal Component Analysis (PCA), Step-by-Step" mempelajari tentang suatu teknik bernama Analisis Komponen Utama (PCA). PCA membantu memahami tentang data yang kompleks dengan mengurangi dimensinya. Misalnya, data tentang transkripsi dua gen pada enam tikus yang berbeda. Gen-gen ini dijadikan representasi variabel yang diukur pada setiap sampel (tikus). Misalnya, membandingkan gen-gen ini dengan variabel seperti nilai ujian matematika dan membaca pada siswa sekolah atau kapitalisasi pasar dan jumlah karyawan dalam sebuah perusahaan. Dalam contoh tersebut, jika hanya ada satu gen yang diukur maka bisa memplot data pada grafik satu dimensi (garis bilangan). Jika ada dua gen yang diukur, data bisa dipetakan pada grafik dua dimensi (grafik X dan Y). Begitu juga jika ada tiga gen, data bisa dipetakan pada grafik tiga dimensi. Namun, ketika jumlah gen yang diukur bertambah menjadi empat atau lebih, memplot data menjadi sulit karena membutuhkan dimensi yang lebih tinggi.

PCA membantu dalam situasi seperti mengubah data ke dimensi yang lebih kecil sehingga dapat dipetakan menjadi grafik dua dimensi. Teknik ini tidak hanya membantu memvisualisasikan data, tetapi juga membantu mengidentifikasi variabel-variabel yang paling penting dalam membedakan sampel dalam dataset. Proses PCA diawali dengan menghitung nilai rata-rata setiap variabel, setelah itu pusat data dipindahkan ke titik awal grafik. Selanjutnya garis grafik diputar hingga sesuai dengan data namun masih melewati titik asal. Garis ini menunjukkan arah variasi data yang paling besar. PCA kemudian menentukan seberapa cocok baris tersebut dengan data. Untuk menilai seberapa cocok suatu garis dengan data, PCA menggunakan proyeksi data pada garis dan mengukur jarak titik data ke garis.

Tujuan PCA adalah mencari garis yang meminimalkan jarak tersebut, dengan kata lain garis yang memaksimalkan jarak dari titik proyeksi ke titik asal. Untuk memahami secara intuitif, jika garis lebih cocok dengan data, maka jarak dari titik garis berkurang, sedangkan jarak dari titik proyeksi ke titik asal bertambah. Secara matematis, PCA menggunakan teorema Pythagoras untuk mengukur jarak antara titik data dan garis proyeksi. Ketika titik data diproyeksikan ke dalam garis lurus, maka terbentuk sudut siku-siku antara garis proyeksi dan garis yang ditarik dari titik asal ke titik data. Ia dapat membagi jarak menjadi dua bagian yang bergerak berlawanan arah seiring bertambahnya jarak.

Jarak yang diukur adalah jarak dari titik proyeksi ke titik asal. Dalam proses PCA, jarak dari setiap titik data ke titik asal diukur setelah diproyeksikan ke garis. Jarak ini kemudian diukur untuk semua titik data, dikuadratkan, dan dijumlahkan. Proses ini diulangi dengan memutar garis hingga ditemukan garis dengan jumlah kuadrat jarak terbesar antara titik proyeksi dan titik asal. Garis ini disebut komponen utama 1 (PC1) dan mewakili arah variasi data yang lebih besar. Secara umum, PC1 menunjukkan sebagian besar variasi data pada satu variabel (misalnya generasi 1) dan sedikit variasi pada variabel lain (misalnya generasi 2). Salah satu cara untuk memahami PC1 adalah dengan membandingkannya dengan resep koktail di mana campuran

empat bagian Gen 1 dan satu bagian Gen 2 menggambarkan hubungan antara variabel-variabel tersebut. Ini juga dikenal sebagai kombinasi linier dari variabel-variabel ini. Ini juga dikenal sebagai kombinasi linier dari variabel-variabel tersebut.

Untuk menggambar plot PCA, yaitu dengan cara memutar PC1 hingga menjadi horizontal, kemudian plot sampel pada plot tersebut menggunakan titik proyeksi. Plot penghalusan adalah representasi grafis dari persentase variasi untuk setiap komponen utama dan membantu memahami kontribusi setiap komponen terhadap variasi total. Untuk PCA tiga variabel (misalnya tiga gen), prosesnya hampir sama dengan PCA dua variabel. Jika datanya terpusat, maka garis yang paling sesuai adalah PC1, disusul PC2 yang tegak lurus PC1, dan seterusnya untuk PC3. Dalam contoh ini, gen dengan kontribusi terbesar pada masing-masing komponen utama menjadi fokus utama pembentukan resep komponen utama tersebut. Jadi, dengan menambahkan lebih banyak variabel (seperti gen), proses PCA menemukan lebih banyak komponen utama dengan menambahkan garis diagonal dan memutarnya agar lebih mencerminkan struktur data dengan lebih baik.

Secara teoritis, jumlah komponen utama (PC) dalam PCA seharusnya sama dengan jumlah variabel atau jumlah sampel, tergantung pada mana yang lebih kecil. Namun, dalam praktiknya, ini tidak selalu terjadi dan akan dijelaskan lebih lanjut dalam video lainnya di masa mendatang. Yang lebih penting untuk dipahami saat ini adalah bagaimana menggunakan nilai eigen untuk menentukan seberapa banyak variasi yang dijelaskan oleh setiap PC. Dengan mengetahui semua komponen utama, kita bisa menggunakan nilai eigen untuk mengetahui proporsi variasi yang dijelaskan oleh setiap PC. Misalnya, jika PC1 menjelaskan 79% variasi, PC2 menjelaskan 15%, dan PC3 menjelaskan 6%, maka PC1 dan PC2 adalah komponen utama yang paling signifikan karena mereka mencakup sebagian besar variasi data, yakni 94%. Untuk mengubah grafik PCA tiga dimensi menjadi grafik dua dimensi, bisa dengan cara menghapus komponen lainnya dan hanya menyisakan data serta PC1 dan PC2. Kemudian, memproyeksikan sampel ke PC1 dan PC2, dan memutar plot sehingga PC1 menjadi horizontal dan PC2 menjadi vertikal. Langkah ini mempermudah visualisasi data. Jika PC3 dan PC4 mempunyai variasi yang signifikan, maka hanya menggunakan dua PC pertama tidak akan memberikan representasi data yang akurat. Namun, plot PCA tetap berguna untuk mengidentifikasi pola dalam data meskipun tidak sempurna.