Nama: Nella Aprilia

NIM: 1103210185

Tugas 5: Membuat Catatan PCA

StatQuest: Principal Component Analysis (PCA), Step-by-Step

Konseptual untuk PCA StatQuest memecahnya menjadi potongan-potongan melalui analisis komponen utama (PCA) langkah-langkah menggunakan dekomposisi nilai singular (SVD). Gen sebagai variabel yang akan diukur untuk setiap sampel, jika hanya mengukur satu gen maka dapat memplot datanya pada garis bilangan. Jika mengukur 2 gen maka dapat memplot datanya pada grafik XY dua dimensi. Gen 1 adalah sumbu x dan merentang salah satu dari dua dimensi pada grafik ini Pada Gen 2 adalah sumbu y dan merentang dimensi lain. Jika mengukur tiga gen maka akan menambahkan sumbu lain pada grafik dan membuatnya terlihat 3D yaitu 3 dimensi. Titik-titik yang lebih kecil memiliki nilai yang lebih besar untuk Gen 3 dan letaknya lebih jauh, titik yang lebih besar mempunyai nilai yang lebih kecil untuk Gen 3 dan lebih dekat namun, jika mengukur 4 gen maka tidak dapat lagi memplot datanya. dibutuhkan 4 gen. 4 dimensi akan membahas bagaimana PCA dapat melakukan 4 atau lebih pengukuran gen dan dengan demikian 4 atau lebih dimensi data dan buat plot PCA 2 dimensi. PCA dapat memberi tahu mengenai gen atau variabel yang paling berharga untuk mengelompokkan data. Misalnya PCA mungkin memberi tahu bahwa Gen 3 bertanggung jawab untuk memisahkan sampel sepanjang sumbu x.

PCA bekerja untuk data 2 Dimensi Terakhir akan membahas bagaimana PCA dapat memberi tahu seberapa akurat grafik 2D. Untuk memahami apa yang dilakukan PCA dan cara kerjanya, cara kerjanya dengan mengkumpulan data yang hanya memiliki 2 gen dan akan mulai dengan memplot datanya, kemudian akan menghitung pengukuran rata-ratanya gen 1, dan pengukuran rata-rata untuk Gen 2 dengan nilai rata-rata dapat menghitung pusat datanya. pada yang terjadi pada grafik tidak lagi memerlukan data asli. Sekarang, akan menggeser datanya sehingga pusatnya berada di atas titik asal grafik. Catatan: Pergeseran data tidak mengubah posisi titik data relatif satu sama lain, titik ini masih yang tertinggi, ini masih titik paling kanan, dan seterusnya. Sekarang data sudah terpusat pada titik asal dan dapat mencoba memasukkan garis ke titik tersebut. Untuk melakukan ini, dimulai dengan menggambar garis acak yang melewati titik asal kemudian putar garis tersebut hingga sesuai dengan datanya, mengingat harus melalui titik asal pada akhirnya, baris ini paling cocok.

Menemukan PC1 kembali ke garis acak asli yang melewati titik asal untuk mengukur seberapa cocok garis ini dengan data, PCA memproyeksikan data ke dalamnya dan kemudian ia dapat mengukur jarak dari data ke garis dan mencoba menemukannya garis yang meminimalkan jarak tersebut, atau dapat mencoba mencari garis yang memaksimalkan jarak dari titik yang diproyeksikan ke titik asal. Jika pilihan tersebut tampaknya tidak setara, kemudian dapat membangun intuisi dengan melihat bagaimana jarak ini menyusut ketika garisnya lebih pas, sementara jarak ini menjadi lebih besar jika garisnya lebih pas. Sekarang, untuk memahami apa yang terjadi secara matematis, mari perhatikan satu titik data.Titik ini tetap dan begitu pula jaraknya dari titik asal dengan kata lain, jarak titik ke titik asal tidak berubah ketika garis putus-putus merah berputar. Saat memproyeksikan suatu titik ke garis,

akan mendapatkan sudut siku-siku di antara titik-titik hitam. Garis dan garis putus-putus merah. Artinya jika memberi label pada sisi-sisinya seperti ini: A,B, dan C, maka dapat menggunakan teorema Pythagoras untuk menunjukkan bagaimana B dan C berbanding terbalik. Karena A dikuadratkan dan tidak berubah jika B bertambah besar maka C harus mengecil begitu pula jika C semakin besar maka B harus semakin kecil.

PCA dapat meminimalkan jarak ke garis, atau memaksimalkan jarak dari titik proyeksi ke titik asal. PCA memproyeksikan data ke dalamnya dan kemudian mengukur jaraknya titik ini ke titik asal, sebut saja D1 dan kemudian PCA mengukur jarak dari titik ini ke titik asal, menyebutnya D2 kemudian diukur D3, D4, D5, dan D6. Berikut enam jarak yang diukur, yang dilakukan adalah mengkuadratkan semuanya, Jaraknya dikuadratkan sehingga nilai negatif tidak menghilangkan nilai positif lalu jumlahkan semua jarak kuadrat tersebut, dan hasilnya Singkatnya akan menyebutnya jarak kuadrat. Salah satu cara untuk memikirkan PC1 adalah dari segi resep koktail untuk membuat PC1, campurkan empat bagian Gen 1 dengan satu bagian Gen 2.

Menemukan PC2 ini hanya dua dimensi grafik, PC2 hanyalah garis yang melalui titik asal yang tegak lurus dengan PC1 tanpa optimasi lebih lanjut yang harus dilakukan dan ini berarti resepnya. PC2 adalah -1 bagian Gen 1 hingga 4 bagian Gen 2. Jika akan menskalakan semuanya sehingga akan mendapatkan vektor satuan, resepnya adalah -0,242 bagian Gen 1 dan 0,97 bagian Gen 2 ini adalah vektor tunggal untuk PC2 atau vektor eigen untuk PC2. Skor untuk PC2, mereka memberi tahu hal itu, dalam hal bagaimana nilai diproyeksikan PC2, Gen 2 merupakan 4 kali lebih penting dari Gen 1. Terakhir untuk nilai eigen untuk PC2 adalah rata-rata jumlah kuadrat jarak diantara titik proyeksi dan titik asal. Menggambar grafik PCA Untuk menggambar plot PCA akhir, cukup memutar semuanya sehingga PC1 horizontal kemudian menggunakan titik-titik yang diproyeksikan untuk menemukan kemana perginya sampel di plot PCA. Untuk titik-titik yang diproyeksikan ini harus sesuai. Sampel PCA dilakukan dengan menggunakan dekomposisi nilai tunggal.

PCA bekerja untuk data 3 Dimensi cara kerjanya sama dengan 2 variabel dengan memusatkan data kemudian menemukan garis yang paling pas untuk melewati titik asal. Sama seperti sebelumnya, jalur yang paling pas adalah PC1. Tetapi resep untuk PC1 sekarang ada 3 bahan. Dalam hal ini Gen 3 merupakan unsur yang paling penting untuk PC1. kemudian menemukan PC2 pada jalur pas terbaik setelah melewati titik asal dan tegak lurus PC1. Berikut resep untuk PC2. Dalam hal ini Gen 1 adalah bahan terpenting untuk PC2. Terakhir, kemudian menemukan PC3, garis yang melalui titik asal dan tegak lurus PC1 dan PC2 punya lebih banyak gen kemudian menemukan lebih banyak komponen utama dengan menambahkan tegak lurus garis dan memutarnya. Secara teori, ada 1 per gen atau variabel, namun dalam praktiknya jumlah PC adalah jumlah variabel atau jumlah sampel, yang mana lebih kecil. Komponen utama yang diketahui dapat menggunakan nilai eigen, yaitu jumlah dari kuadrat jarak untuk menentukan proporsi variasi yang diperhitungkan setiap PC dalam hal PC1 menyumbang 79% variasi, PC2 menyumbang 15% pada variasi PC3 menyumbang 6% dari variasi. Berikut plot layarnya. Menggunakan PC1 dan PC2 untuk menggambar grafik dua dimensi dengan datanya.