Nama : Anindita Khusnul Oktavia

Program : Introduction to Python for Data Science

SUMMARY

Unsupervised

Unsupervised machine learning algorithm digunakan untuk mengelompokkan data tidak terstruktur menurut kesamaan dan pola yang berbeda dalam kumpulan data. Unsupervised machine learning ini juga mendeskripsikan informasi yang ada - menelusuri seluk-beluknya dan mengidentifikasi data apa itu sebenarnya. Algoritme pada kasus ini tidak dipandu seperti supervised learning algorithm.

Unsupervised algorithm menangani data tanpa pelatihan sebelumnya dan dibiarkan di perangkatnya sendiri untuk menyelesaikan masalah sesuai keinginannya. Unsupervised algorithm berfungsi dengan data tak berlabel. Tujuannya adalah eksplorasi. Jika supervised machine learning berfungsi di bawah aturan yang ditetapkan dengan jelas, unsupervised learning berfungsi dalam kondisi hasil yang tidak diketahui dan karenanya perlu didefinisikan dalam proses.

Unsupervised machine learning algorithm digunakan untuk:

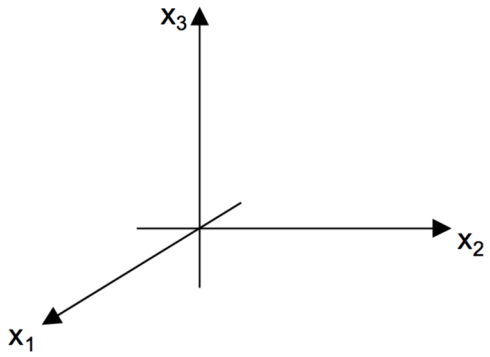
1. **explore** the structure of the information;
2. **extract** valuable insights;
3. **detect** patterns;
4. **implement** this into its operation in order to increase efficiency.

Principal Component Analysis

Principal Component Analysis (PCA) adalah teknik linear dimensionality reduction yang dapat digunakan untuk mengekstraksi informasi dari ruang dimensi tinggi dengan memproyeksikannya ke dalam sub-ruang berdimensi lebih rendah. PCA mencoba untuk mempertahankan bagian penting yang memiliki lebih banyak variasi data dan menghapus bagian yang tidak penting dengan variasi yang lebih sedikit.

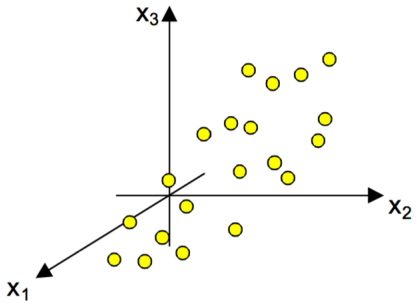
How PCA works

Pertimbangkan matriks X dengan baris N (alias "observasi") dan kolom K (alias "variabel"). Untuk matriks ini kita membangun ruang variabel dengan dimensi sebanyak variabel (lihat gambar di bawah). Setiap variabel mewakili satu sumbu koordinat. Untuk setiap variabel, panjangnya telah distandarisasi menurut kriteria penskalaan, biasanya dengan penskalaan ke varian unit.



K-dimensional variable space. Untuk mempermudah, hanya tiga sumbu variabel yang ditampilkan. "Panjang" dari setiap sumbu koordinat telah distandarisasi sesuai dengan kriteria tertentu, biasanya skala varian unit.

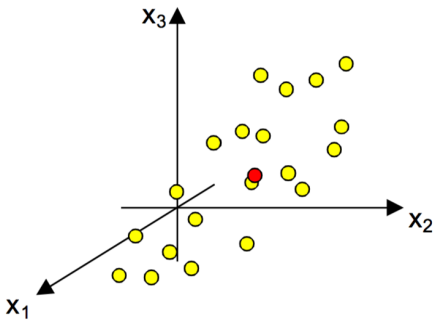
Pada langkah berikutnya, setiap observasi (baris) dari matriks X ditempatkan di ruang variabel K-dimensional. Akibatnya, baris di tabel data membentuk sekumpulan titik di ruang ini.



Pengamatan (rows) dalam matriks data X dapat dipahami sebagai sekumpulan titik dalam ruang variabel (K-space).

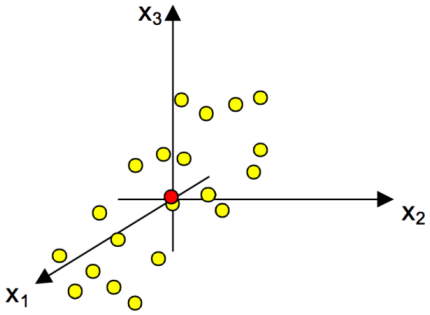
Mean centering

Selanjutnya, mean-centering melibatkan pengurangan rata-rata variabel dari data. Vektor rata-rata sesuai dengan titik di ruang-K.



Dalam prosedur mean-centering, pertama kita menghitung rata-rata variabel. Vektor rata-rata ini dapat diinterpretasikan sebagai titik (di sini berwarna merah) di ruang. Titik tersebut terletak di tengah-tengah titik *swarm* (di pusat gravitasi).

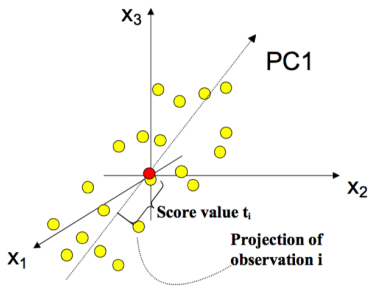
Pengurangan rata-rata dari data dilanjutkan dengan pemosisian ulang sistem koordinat, sehingga titik rata-rata sekarang adalah titik asal.



Prosedur mean-centering dilanjutkan dengan memindahkan asal dari sistem koordinat untuk bertepatan dengan titik rata-rata (di sini berwarna merah).

The first principal component

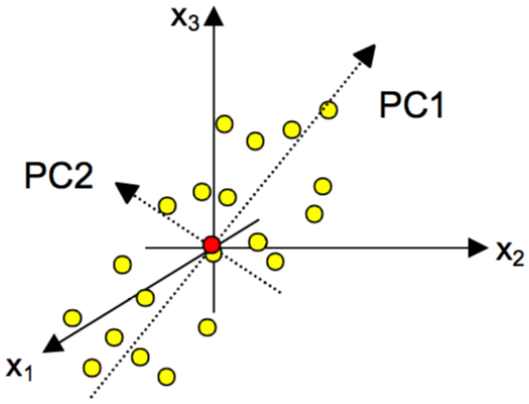
Setelah pemusatan rata-rata dan penskalaan ke varian unit, kumpulan data siap untuk penghitungan summary index pertama, principal component pertama (PC1). Komponen ini adalah garis dalam ruang variabel dimensi-K yang paling mendekati data dalam least squares. Garis ini melewati titik rata-rata. Setiap pengamatan (titik kuning) sekarang dapat diproyeksikan ke garis ini untuk mendapatkan nilai koordinat di sepanjang PC-line. Nilai koordinat baru ini juga dikenal sebagai *score*.



Principal component pertama (PC1) adalah garis yang paling sesuai untuk point swarm. PC1 mewakili maximum variance direction dalam data. Setiap pengamatan (titik kuning) dapat diproyeksikan ke garis ini untuk mendapatkan nilai koordinat di sepanjang PC-line. Nilai ini dikenal sebagai score.

The second principal component

Biasanya, satu summary index atau principal component tidak cukup untuk memodelkan systematic variation dari suatu kumpulan data. Jadi, summary index kedua - principal component kedua (PC2) - dihitung. PC kedua juga diwakili oleh garis dalam ruang variabel K-dimensional, yang ortogonal ke PC pertama. Garis ini juga melewati titik rata-rata, dan meningkatkan perkiraan data X sebanyak mungkin.

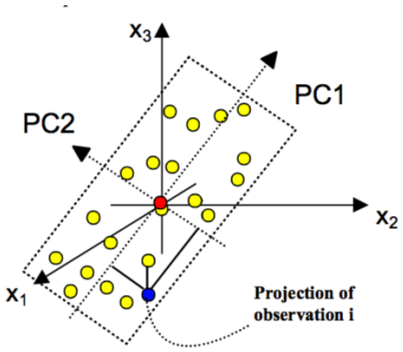


Principal component kedua (PC2) diorientasikan sedemikian rupa sehingga mencerminkan sumber variasi terbesar kedua dalam data, sementara ortogonal terhadap PC pertama. PC2 juga melewati titik rata-rata.

Two principal components define a model plane

Ketika dua principal components ditemukan, mereka bersama-sama menentukan place, window ke ruang variabel K-dimensional. Dengan memproyeksikan semua pengamatan ke sub-ruang berdimensi rendah

dan memplot hasilnya, dimungkinkan untuk memvisualisasikan struktur kumpulan data yang diselidiki. Nilai koordinat dari pengamatan pada bidang ini disebut skor, dan karenanya penggambaran konfigurasi yang diproyeksikan seperti itu dikenal sebagai **score plot**.



Dua PC membentuk plane. Bidang ini merupakan window ke dalam ruang multidimensi, yang dapat divisualisasikan secara grafis. Setiap pengamatan dapat diproyeksikan ke bidang ini, memberikan skor untuk masing-masing.

Untuk implementasinya ada pada file .ipynb