Pengenalan Principal Component Analysis

Teknik Fisika
Institut Teknologi Bandung

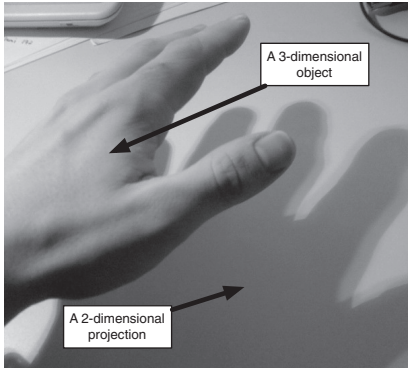
Bahan-bahan berikut ini diambil dari:

Simon Rogers and Mark Girolami. A First Course in Machine Learning. 2nd Edition. CRC Press. 2017.

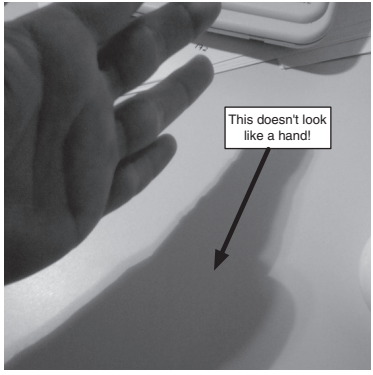
Proyeksi data

- Misalkan kita memiliki dataset yang terdiri dari N objek, y_n, $n = 1, 2, \ldots, N$.
- Setiap objek adalah vektor dengan dimensi M (M dapat berupa jumlah fitur dari data yang kita miliki).
- Banyak model mesin pembelajar yang memiliki parameter yang bertambah banyak jika dimensi dari data semakin banyak.
- Data dengan dimensionalitas yang tinggi juga sulit untuk divisualisasi.
- Terkadang transformasi data M-dimensi ke representasi D dimensi (dengan $D < M$) diperlukan. Proses ini dikenal dengan nama proyeksi.
- Idealnya proyeksi ini tetap memiliki properti menarik dari data yang kita ingin pelajari.

Contoh proyeksi



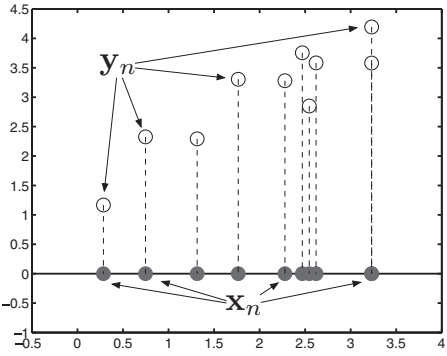
(a)



(b)

Objek tangan (3d) diproyeksikan ke 2d (bayangan).

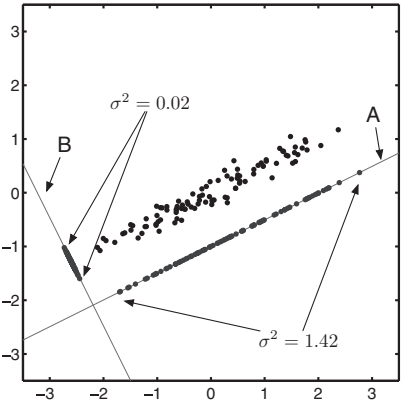
Contoh proyeksi



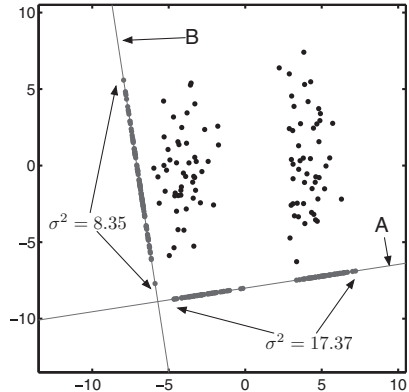
(c)

Proyeksi suatu data ke sumbu x.

Bagaimana cara mengukur derajat "menarik/penting" dari data?



(a) Data from a single, elongated Gaussian.



(b) Data from two Gaussians.

Contoh proyeksi data ke dua arah (A dan B).

Principal Component Analysis (PCA)

- PCA adalah metode yang sering digunakan untuk melakukan proyeksi data ke dimensi yang lebih rendah.
- PCA adalah proyeksi linear: setiap dimensi proyeksi adalah kombinari linear dari dimensi asli. Jika kita melakukan proyeksi dari M ke D dimensi, PCA akan mendefisikan D vektor, $\mathbf{w}_d$, yang masing-masingnya berdimensi M. Elemen ke-d dari proyeksi x_{nd} (di mana $[x_n = x_{n1}, \ldots, x_{nD}]^\mathsf{T}$) adalah:

$$x_{nd} = \mathbf{w}_d^\mathsf{T} \mathbf{y}_n$$

PCA

- PCA menggunakan variansi pada ruang proyeksi sebagai kriteria untuk memilih $\mathbf{w}_d$.
- Misalnya: $\mathbf{w}_1$ adalah proyeksi yang akan membuat variansi pada x_{n1} semaksimal mungkin.
- Dimensi proyeksi kedua juga dipilih untuk memaksimalkan variansi, namum $\mathbf{w}_2$ harus ortogonal terhadap $\mathbf{w}_1$:

$$\mathbf{w}_2^{\mathrm{T}}\mathbf{w}_1 = 0$$

- Begitu juga untuk dimensi proyeksi yang ketiga dan seterusnya. Secara umum:

$$\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_j = 0 \,\forall\, j \neq i$$

- PCA juga menambahkan konstrain bahwa tiap $\mathbf{w}_i$ memiliki panjang 1.

$$\mathbf{w}_i^{\mathrm{T}}\mathbf{w}_i = 1$$

Prosedur PCA

Dapat ditunjukkan bahwa $\mathbf{w}_i$ dapat diperoleh dari persamaan eigen:

$$\mathbf{C}\mathbf{w} = \lambda\mathbf{w}$$

dengan $\mathbf{C}$ adalah matriks kovariansi:

$$\mathbf{C} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{y}_n - \bar{\mathbf{y}})(\mathbf{y}_n - \bar{\mathbf{y}})^{\mathsf{T}}$$

atau:

$$\mathbf{C} = \frac{1}{N}\mathbf{Y}^{\mathsf{T}}\mathbf{Y}$$

dan nilai eigen λ adalah variansi dari data yang diprojeksikan ke arah $\mathbf{w}$.

Prosedur PCA

- Transformasi data sehingga memiliki rata-rata nol dengan cara mengurangi setiap titik data dengan rata-rata sampel:

$$\bar{\mathbf{y}} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{y}_n$$

- Hitung matriks kovariansi.
- Cari pasangan nilai eigen dari matriks kovariansi.
- Cari eigenvektor dengan D nilai eigen tertinggi.
- Buat proyeksi data:

$$\mathbf{X} = \mathbf{YW}$$

di mana $\mathbf{W}$ adalah matriks $M \times D$ yang dibuat dari D vektor eigen dari matriks kovariansi.

Tugas: PCA pada data sintetik

- Buat data sintetik yang memiliki struktur klaster. Misalnya

```
Y_1 = np.random.randn(20,2) # jumlah data adalah 20.
Y_2 = np.random.randn(20,2) + 5.0
Y_3 = np.random.randn(20,2) - 5.0
Y = np.concatenate( (Y_1, Y_2, Y_3), axis=0 )
```

- Tambahkan beberapa dimensi randam pada data yang tidak memiliki struktur klaster.

```
Ndata = Y.shape[0]
Y = np.concatenate( (Y, np.random.randn(Ndata,5)), axis=1) # tambah
↪  data 5 dimensi
```

- Aplikasikan prosedur PCA pada data tersebut. Plot data yang sudah tereduksi dimensionalitasnya.

Bandingkan hasil yang Anda peroleh dengan menggunakan pustaka Scikit Learn. Apakah ada perbedaan yang Anda amati? Jelaskan apa yang mungkin menyebabkan perbedaan tersebut jika ada.

Hint

- Lakukan visualisasi pada data sintetik yang dibuat. Buat scatterplot untuk pasangan data pada tiap dimensi (fitur).
- Untuk menghitung nilai dan vektor eigen: `np.linalg.eig`
- Untuk melakukan perkalian matriks: `np.matmul` (jika tipe data adalah `ndarray`).