Zajęcie 0.3. Analiza głównych składowych

Abstract

Celem jest nabycie podstawowej znajomości użycia rozkładu SVD w celu uzyskania głównych składowych (PCA).

1. Podstawowe pojęcia

Analiza głównych składowych (PCA) to jedno z głównych zastosowań SVD, zapewniające statystyczną interpretację hierarchicznego układu współrzędnych opartego na danych, używanego do reprezentowania skorelowanych danych wielowymiarowych. Ten układ współrzędnych obejmuje macierze korelacji opisane w zajęciu o SVD a korelacji.

PCA wstępnie przetwarza dane poprzez odejmowanie średniej i ustawianie wariancji na jedynkę przed wykonaniem SVD. Geometria powstałego układu współrzędnych jest określana przez główne składowe (PC), które są ze sobą nieskorelowane (ortogonalne), ale mają maksymalną korelację z pomiarami.

2. Obliczenie

Mamy macierz $X = \{X_{ij}\}_{i=\overline{1,n},j=\overline{1,m}} \in \mathbb{R}^{n\times m}$.

1. Obliczymy średnie wartości w kolumnach

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$$

2. Macierz średnia

$$\bar{X} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \dots & \bar{x}_m \end{bmatrix}$$

3. Odejmujemy średnią macierz

$$B = X - \bar{X} \tag{1}$$

4. Macierz kowariancji

$$C = \frac{1}{n-1} B^\star B$$

> Macierz kowariancji C jest symetryczna i dodatnio półokreślona i ma nieujemne rzeczywiste wartości własne. Każdy element C_{ij} określa ilościowo korelację cech i i j we wszystkich eksperymentach.

5. > **Składowymi głównymi** V są wektory własne macierzy C i definiują one zmianę współrzędnych, w których macierz kowariancji jest diagonalna:
>
> $$CV = VD,$$

   Mamy że

$$C = VDV^\star, \quad D = V^\star CV \tag{2}$$

Kolumny macierzy wektorów własnych V są głównymi składowymi, a elementy macierzy diagonalnej D są wariancjami danych wzdłuż tych kierunków. Ta transformacja ma pewność istnienia, ponieważ C jest hermitowską, a kolumny V są ortonormalne. W tych współrzędnych składowych głównych wszystkie cechy są ze sobą liniowo nieskorelowane.

Macierz głównych składowych V jest także macierzą prawych wektorów osobliwych B. Załużmy że $B = U\Sigma V^T$. Wtedy

$$C = \frac{1}{n-1} B^\star B = \frac{1}{n-1} V\Sigma U^\star U\Sigma V^T = \frac{1}{n-1} V\Sigma^2 V^T$$

Mamy że

$$D = \frac{1}{n-1}\Sigma^2$$

Wariancja danych w tych współrzędnych, określona przez elementy diagonalne λ_k z D, jest powiązana z wartościami osobliwymi jako

$$\lambda_k = \frac{\sigma_k^2}{n-1}$$

Zatem SVD zapewnia solidne numerycznie podejście do obliczania głównych składowych. Przybliżenie $\tilde{B}$ uzyskane poprzez zachowanie tylko pierwszych r składowych głównych będzie miało brakującą wariancję związaną z kwadratowym błędem normy Frobeniusa.

3. Zadanie

Zadanie dotyczy obliczenia środka, osi głównych oraz kątu obrotu danych dwuwymiarowych z pliku .csv zgodnie z wariantem zadania

Sprawozdania w postaci:

1. Sprawozdanie (plik .pdf)
2. plik .ipynb
3. pdf-eksport pliku .pynb

zachować w zdalnym repozytorium (np Github) link na który umieściś w sprawozdaniu. Sprawozdanie należy wysłać na e-uczelnię w ustalonym terminem.

References

References

[pandasUG] Pandas User's Guide `https://pandas.pydata.org/pandas-docs/stable/user_guide/index.html`

[DA2016] Data Analysis with Python and pandas using Jupyter Notebook `https://dev.socrata.com/blog/2016/02/01/pandas-and-jupyter-notebook.html`