Algorytm PCA - składowe główne

Własności

- pozwala na ***redukcję wymiaru*** problemu
- ***transformuje (liniowo) przestrzeń*** *atrybutów* dostarczając *nowych współrzędnych*
- algorytm ***nie posiada parametrów***
- wykorzystuje macierz korelacji, ***eliminuje kowariancję*** (czyli liniowe zależności między atrybutami)
- składowych głównych można wyznaczyć tyle, ile było pierwotnych składowych

Kryterium *Kaisera-Guttmana*

Popularne kryterium dobierania ilości składowych:
Należy zachować składowe, dla których wartości własne są większe od 1, czyli wkład składowej większy, niż wkład pojedynczej zmiennej

Kroki algorytmu

Dane wejściowe zostały przedstawione w tabeli.

Dane wejściowe:

	atrybut			
pomiar	1	2	...	m
1	x_{11}	x_{12}	...	x_{1m}
2	x_{21}	x_{22}	...	x_{2m}
...	...	...	...	...
n	x_{n1}	x_{n2}	...	x_{nm}

Algorytm składowych głównych dla danych wejściowych został przedstawiony poniżej:

Krok 1: Obliczenie **średniej i odchylenia standardowego** (dla każdego atrybutu)
Krok 2: Normalizacja zgodnie ze wzorem. Po normalizacji wszystkie atrybuty mają parametry : średnia = 0 i odchylenie = 1.

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

Krok 3: Chcemy znaleźć **nowy układ współrzędnych**. Zakładamy tylko przekształcenia liniowe (obroty, odbicia)

Założenia:

- tylko przekształcenia liniowe
- maksymalizowana wariancja (klasyczna miara zróżnicowania)
- nowe kierunki składowych są normalizowane

Krok 3. 1. Wyznaczamy macierz kowariancji Cz:

$$\mathbf{Z} = \begin{bmatrix} z_{11} & z_{12} & \dots & z_{1m} \\ z_{21} & z_{22} & \dots & z_{2m} \\ \dots & \dots & \dots & \dots \\ z_{n1} & z_{n2} & \dots & z_{nm} \end{bmatrix} \mathbf{a_i} = \begin{bmatrix} z_{1i} \\ z_{2i} \\ \dots \\ z_{ni} \end{bmatrix} \mathbf{a_j} = \begin{bmatrix} z_{1j} \\ z_{2j} \\ \dots \\ z_{nj} \end{bmatrix}$$

dane po normalizacji: $\mu_{a_i} = \frac{1}{n}\sum_{k=0}^{n} z_{ki} = 0, \sigma_{a_i}^2 = \frac{1}{n}\sum_{k=0}^{n} z_{ki}^2 = 1$

Kowariancja - miarą liniowej zależności pomiędzy a_i i a_j

$$\sigma_{a_i a_j} = \frac{1}{n}\sum_{k=0}^{n} z_{ik} z_{jk} = \frac{1}{n}\mathbf{a_i}\mathbf{a_j}^T \text{ gdzie } \mathbf{a_j}^T = \begin{bmatrix} z_{1j} & z_{2j} & \dots & z_{nj} \end{bmatrix}$$

$$-1 \leq \sigma_{a_i a_j} \leq 1$$

$$\mathbf{C_z} = \frac{1}{n}\mathbf{Z}\mathbf{Z}^T = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & \sigma_2^2 & \dots & \sigma_{2m} \\ \dots & \dots & \dots & \dots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_m^2 \end{bmatrix} = \begin{bmatrix} 1 & \sigma_{12} & \dots & \sigma_{1m} \\ \sigma_{21} & 1 & \dots & \sigma_{2m} \\ \dots & \dots & \dots & \dots \\ \sigma_{n1} & \sigma_{n2} & \dots & 1 \end{bmatrix}$$

Ponieważ $\sigma_{ij} = \sigma_{ji}$ macierz $\mathbf{C_z}$ jest symetryczna

- sumaryczna wariancja

$$\sum_{j=0}^{m} \sigma_i = m$$

- po zmianie (rotacja, odbicie) układu współrzędnych sumaryczna wariancja nie zmieni się

$\mathbf{Y} = \mathbf{PZ}$, gdzie $\mathbf{P}$ jest macierzą przekształcenia

, macierz $\mathbf{P}$ zawiera wektory, które są kierunkami składowych

$$\mathbf{P} = [\mathbf{p_1}, \mathbf{p_2}, ..., \mathbf{p_m}]$$

Wykorzystując algorytmy algebry liniowej przekształca się przestrzeń, aby macierz kowariancji była diagonalna

$$\mathbf{C_Y} = \begin{bmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \lambda_m \end{bmatrix}$$

$$\mathbf{C_Y} = \frac{1}{n}\mathbf{YY}^T = \frac{1}{n}(\mathbf{PZ})(\mathbf{PZ})^T = \frac{1}{n}\mathbf{PZZ}^T\mathbf{P}^T = \mathbf{PC_Z P}^T$$

dla macierzy symetrycznej $\mathbf{A}$, macierzy jej wektorów własnych $\mathbf{E}$ zachodzi zależność:

$\mathbf{A} = \mathbf{EDE}^T$, gdzie $\mathbf{D}$ jest macierzą diagonalną

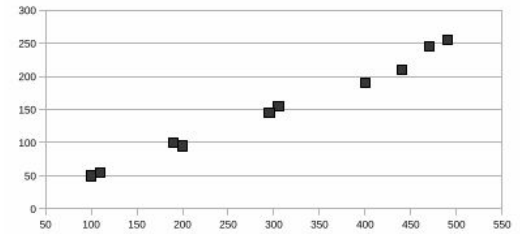
więc: $\mathbf{P}$ jest macierzą wektorów własnych macierzy C_Z

Krok 3. 2. Wyznaczamy z macierzy kowariancji wartości własne i wektory własne: Poszukujemy wartości własnych i wektorów własnych - wektory własne będą naszymi składowymi głównymi, a wartości własne będą mówić jak dużo wariancji przynależy do tego wektora własnego.

Przykład:

(osie x i y odpowiadają atrybutom)
Krok 1: Obliczenie średnich i odchyleń dla atrybutów
Krok 2: Normalizacja

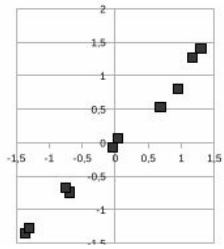
dane wejściowe (10 przykładów, 2 atrybuty):



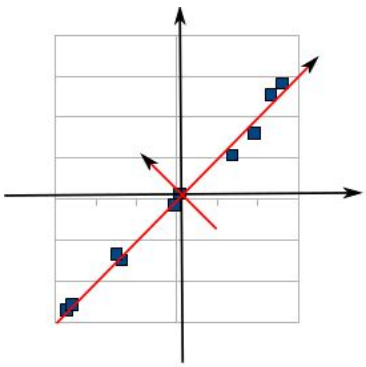
$$\mu_1 = 300$$
$$\sigma_1 = 146.4$$
$$\mu_2 = 150$$
$$\sigma_2 = 74.4$$

normalizacja:

$$z_{i1} = \frac{x_{i1} - \mu_1}{\sigma_1}$$
$$z_{i2} = \frac{x_{i2} - \mu_2}{\sigma_2}$$



Kierunki składowych głównych dla rozpatrywanego przykładu:



Krok 3. Wyznaczamy macierz kowariancji. Obliczamy wartości własne i wektory własne.

Dla rozpatrywanego przykładu:

$$\mathbf{C_z} = \begin{bmatrix} 1 & 0.994 \\ 0.994 & 1 \end{bmatrix}$$

po rozkładzie na wartości własne i wektory własne:

$$\begin{bmatrix} 1 & 0.994 \\ 0.994 & 1 \end{bmatrix} = \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 1.994 & 0 \\ 0 & 0.006 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \end{bmatrix}$$

Nowe kierunki

$$\mathbf{p_1} = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} \end{bmatrix}, \mathbf{p_2} = \begin{bmatrix} \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} \end{bmatrix}, \lambda_1 = 1.994, \lambda_2 = 0.006$$

Źródła:

[1] https://pl.wikipedia.org/wiki/Analiza_g%C5%82%C3%B3wnych_sk%C5%82adowych
[2] http://coin.wne.uw.edu.pl/~jcieciel/FA_PCA_prezentacja%20v2.pdf