

PCA a iné

19. novembra 2014

Vstup: $x_1, x_2, \dots, x_t$, kde $x_i \in R^n$, resp. matica X tvaru $t \times n$.

Cieľ: znížiť počet atribútov a zachovať, čo najviac informácie

Atribúty na nulový priemer:

$$\mu = \frac{1}{t} \sum x_i$$

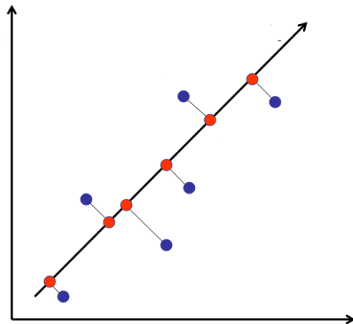
$$x_i = x_i - \mu$$

(Voliteľné) štandardné odchýlka na 1:

$$\sigma_j^2 = \frac{1}{t} \sum x_{i,j}^2$$

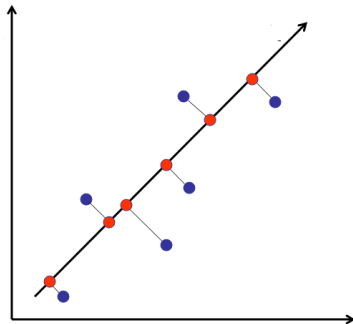
$$x_{i,j} = \frac{x_{i,j}}{\sigma_j}$$

Projekcia



$$x'_i = \langle x_i, u \rangle = x_i u^T$$

Projekcia



$$x'_i = \langle x_i, u \rangle = x_i u^T$$

$$u \cdot u^T = 1$$

Chceme maximalizovať rozptyl:

$$\max \frac{1}{t} \sum (x_i u^T)^2$$

Chceme maximalizovať rozptyl:

$$\max \frac{1}{t} \sum (x_i u^T)^2$$

$$\max \frac{1}{t} \sum u x_i^T x_i u^T$$

Chceme maximalizovať rozptyl:

$$\max \frac{1}{t} \sum (x_i u^T)^2$$

$$\max \frac{1}{t} \sum u x_i^T x_i u^T$$

$$\max u \left(\frac{1}{t} \sum x_i^T x_i \right) u^T$$

Chceme maximalizovať rozptyl:

$$\max \frac{1}{t} \sum (x_i u^T)^2$$

$$\max \frac{1}{t} \sum u x_i^T x_i u^T$$

$$\max u \left(\frac{1}{t} \sum x_i^T x_i \right) u^T$$

Kovariančná matica:

$$S = \frac{1}{t} \sum x_i^T x_i$$

Maximalizujeme: $\max uSu^T$ za predpokladu $uu^T = 1$.

Lagrangeove multiplikátory, chceme nulový gradient:

$$\nabla_u (uSu^T + \lambda(1 - u^T u)) = 0$$

Maximalizujeme: $\max uSu^T$ za predpokladu $uu^T = 1$.

Lagrangeove multiplikátory, chceme nulový gradient:

$$\nabla_u (uSu^T + \lambda(1 - u^T u)) = 0$$

$$uS = \lambda u$$

Maximalizujeme: $\max uSu^T$ za predpokladu $uu^T = 1$.

Lagrangeove multiplikátory, chceme nulový gradient:

$$\nabla_u (uSu^T + \lambda(1 - u^T u)) = 0$$

$$uS = \lambda u$$

Hľa! Vlastné čísla a vlastné vektory (eigenvalue, eigenvector).

Daná štvorcová matica M , hľadáme vektor x a číslo λ , také, že: $xM = \lambda x$

Daná štvorcová matica M , hľadáme vektor x a číslo λ , také, že: $xM = \lambda x$

$$x(M - \lambda I) = 0$$

Daná štvorcová matica M , hľadáme vektor x a číslo λ , také, že: $xM = \lambda x$

$$x(M - \lambda I) = 0$$

$$\det(M - \lambda I) = 0$$

Daná štvorcová matica M , hľadáme vektor x a číslo λ , také, že: $xM = \lambda x$

$$x(M - \lambda I) = 0$$

$$\det(M - \lambda I) = 0$$

Sú aj rýchlejšie spôsoby počítania.

Nájdeme najväčšie vlastné číslo λ_1 , jeho vlastný vektor u_1 a hlavný pre každý prvok komponent spočítame ako $x_i u^T$. Ako ďalšie komponenty?

Nájdeme najväčšie vlastné číslo λ_1 , jeho vlastný vektor u_1 a hlavný pre každý prvok komponent spočítame ako $x_i u^T$. Ako ďalšie komponenty?

Stačí zobrať ďalšie vlastné vektory (v poradí podľa veľkosti vlastných čísel).

Nájdeme najväčšie vlastné číslo λ_1 , jeho vlastný vektor u_1 a hlavný pre každý prvok komponent spočítame ako $x_i u^T$. Ako ďalšie komponenty?

Stačí zobrať ďalšie vlastné vektory (v poradí podľa veľkosti vlastných čísel).

Keď do matice P dáme po stĺpcov prvých k vlastných vektorov, tak nové atribúty dostaneme ako XP .

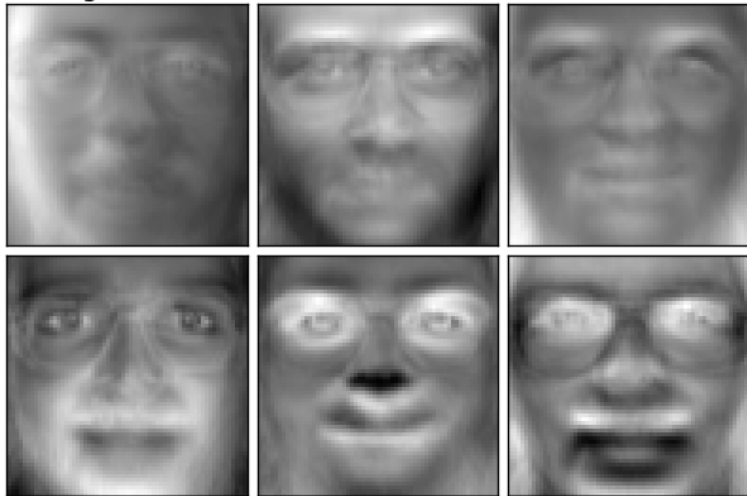
Fancy pictures

First centered Olivetti faces



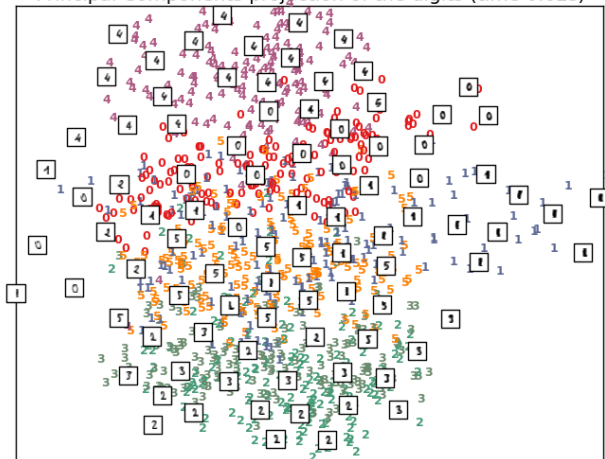
Fancy pictures

Eigenfaces - RandomizedPCA - Train time 0.3s



Fancy pictures

Principal Components projection of the digits (time 0.02s)



Ďalšie vlastnosti

Ak je matica M symetrická (ako naše S), tak sa dá rozložiť ako:
 $M = QDQ^T$, kde Q je ortogonálna a má v stĺpcoch vlastné vektory z M a
 D je diagonálna matica, ktorá má na diagonále vlastné čísla z M .
(Ortogonalna matica: $QQ^T = I$.)

Ďalšie vlastnosti

Ak je matica M symetrická (ako naše S), tak sa dá rozložiť ako:
 $M = QDQ^T$, kde Q je ortogonálna a má v stĺpcoch vlastné vektory z M a D je diagonálna matica, ktorá má na diagonále vlastné čísla z M .
(Ortogonálna matica: $QQ^T = I$.)

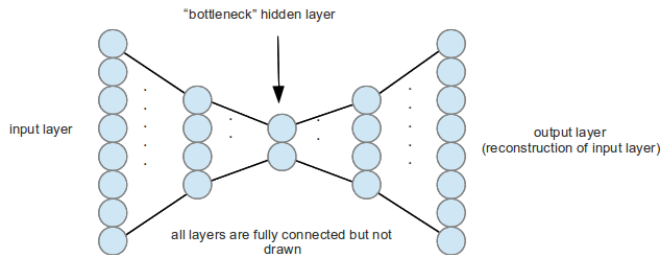
- Ak urobíme PCA na všetkých komponentoch, potom $(XP)^T(XP) = P^T X^T X P = P^T S P = D$. Toto likviduje koreláciu medzi atribútmi.
- ZCA - PCA, kde naškálujeme výsledné atribúty na rovnaký rozptyl (užitočné predspracovanie pre niektoré algoritmy, ktoré nemajú radi korelované atribúty).

Reprezentácia 1: Každý dokument ako riedky vektor dvojíc:
(slovo, počet výskytov)

PCA

Reprezentácia 2: Každý dokument ako vektor zopár (100 - 1000) reálnych čísel.

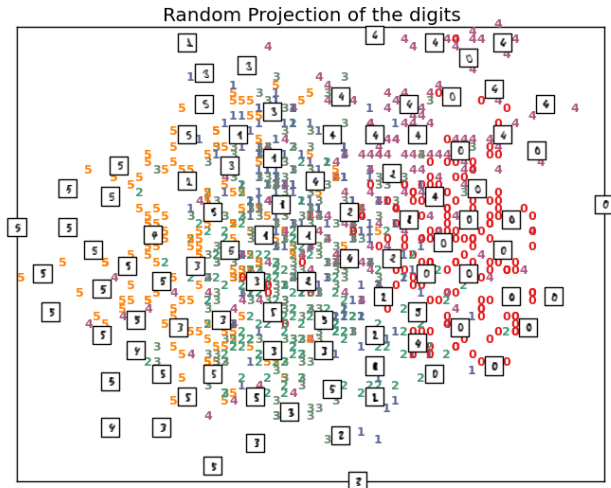
Autoenkóder



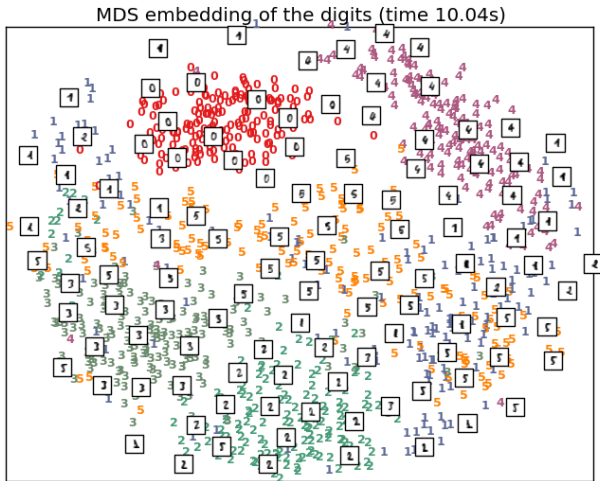
Cieľ: Redukovať dáta do 2 (3) dimenzií. Chceme podobné vstupy mať pri sebe, odlišné ďaleko od seba.

Väčšinou rátame natvrdo projekciu prvku (nedá sa použiť pre nové vstupy).

Random projection

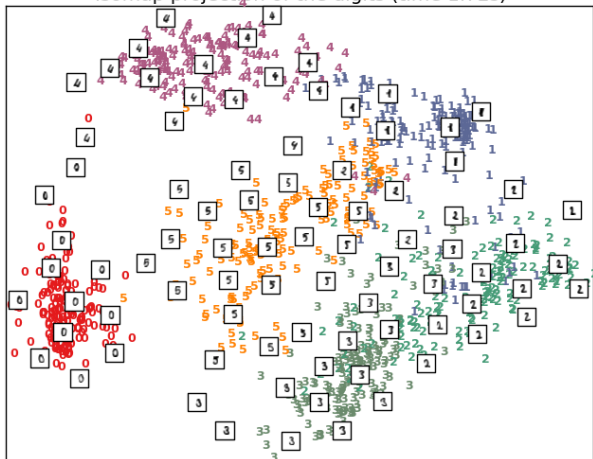


Multidimension scaling

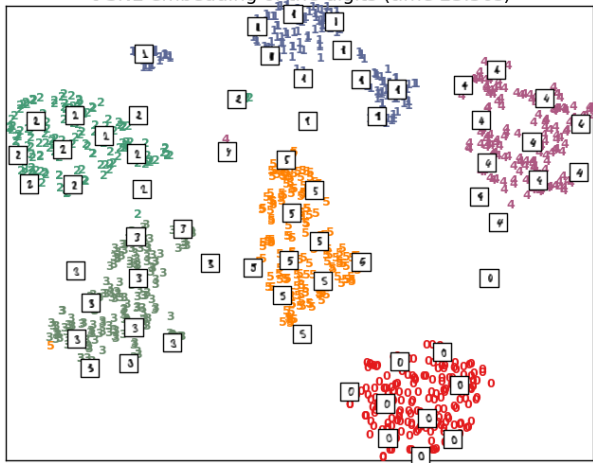


ISOMAP

Isomap projection of the digits (time 1.72s)



t-SNE embedding of the digits (time 23.50s)



Das Ende der Vorlesung