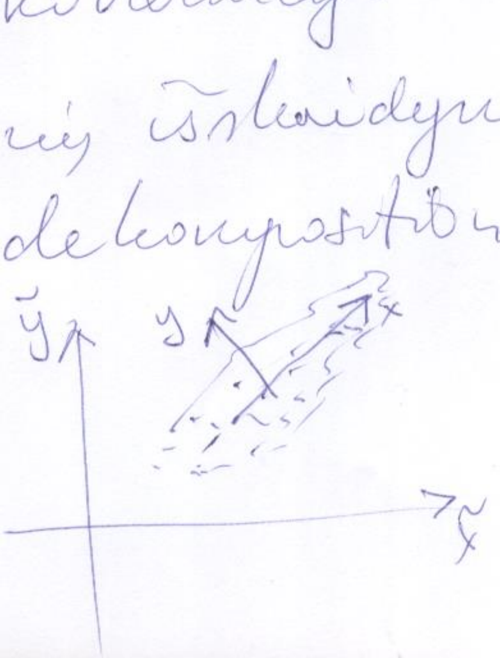
Pagrindinių komponenčių analizė 7 paskaita

PCA – principal component analysis

Surandame naują koordinačių sistemą, kuri aprašo duotus daugiamačius duomenis

Pirmoji kryptis (pirmasis bazinis vektorius) aprašo tą kryptį, kuria judant duomenų variacija yra didžiausia. Antroji randa antrą didžiausios variacijos kryptį ir t.t.

Trumpai galime ~~teigti~~ – kad ieškome <u>duomenų</u> kovariacijos matricos tikrinių vektorių išskaidymo (arba singular value decomposition bendruoju atveju.)



PCA yra apibrėžiama kaip tiesinė ortogonali transformacija

0. Nagrinėkime duomenų matricą X, dimensija $n \times p$ $\left(X[i][j] \right.$ $1 \le i \le n$, $1 \le j \le p$

turime p - požymių

viso atlikta n matavimų (eksperimentų)

1. Matricą normuojame (pervedame duomenis standartinėn formą)

a) skaičiuojame kiekv. požymio vidurkį

$$v[j] = \frac{1}{n} \sum_{i=1}^{n} X[i][j]$$

b) skaičiuojame standartinio nuokrypio koeficientą

$$\sigma[j] = \left(\frac{1}{n} \sum_{i=1}^{n} \left(X[i][j] - V[j] \right)^2 \right)^{1/2}$$

c) Duomenų normavimas

$$M[i][j] = (X[i][j] - V[j]) / \sigma[j]$$

2. Skaičiuojame ~~koreliacijų~~ matricą (kovariacijų matricą)

$$C = M^T M$$

M yra $n \times p$ matrica

M^T yra $p \times n$ matrica

C yra $p \times p$ matrica

$$C[i][j] = \sum_{k=1}^{n} M^T[i][k] M[k][j]$$

$$= \sum_{k=1}^{n} M[k][i] \, M[k][j]$$

i-tasis stulpelis

j-tasis stulpelis

Matrica C yra simetrinė

$$C^T = (M^T M)^T = M^T \cdot (M^T)^T = M^T M.$$

Todėl jos tikrinės reikšmės yra realios, o tikriniai vektoriai yra ortogonalūs

$$\boxed{C W_j = \lambda_j W_j}, \quad \lambda_j \in \mathbb{R}. \quad j = 1, \dots, p.$$

$$(W_i^T, W_j) = 0, \quad \text{kai } i \neq j.$$

$\uparrow$

skaliarinė sandauga.

Sakykime, kad $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p$.

Paliekame tik L-pirmųjų komponenčių (tikrinių vektorių) kurie ir apibrėžia naują koordinačių sistemą

$$\lambda_1 + \dots + \lambda_L \approx 0.99\,(\lambda_1 + \dots + \lambda_p).$$

Tikriniais vektoriais normuojame, kad

$\|W_j\| = 1.$

Tada apibrėžiame L-pirmųjų tikrinių vektorių matrica

$$\tilde{W}_L = (W_1\, W_2, \dots W_L) \quad p \times L \text{ matrica}$$

Duomenis X tada transformuojame į duomenis

$n \times p \to M$

$p \times L - \tilde{W}_L$

$$T_L = M \tilde{W}_L$$

T_L vėl turi n-eilučių (matavimų skaičius).

bet tik L-stulpelių

Pvz. DNT biočiptiklių uždavinys:
sąrašo matricos kai 1 s, turime 400 s.
$n = 1000$ (koncentr. ribinių). $L \approx 30$

Kadangi $\tilde{W}_L$ apibrėžia ortonormuotą transformaciją, tai

$$\underset{p\times L}{\tilde{W}_L}\ \underset{L\times p}{\tilde{W}_L^T} = E_p$$ (vienetinė kvadratinė matrica $p \times p$)

Todėl pabaigę analizę (regresija), galime vėl apskaičiuoti atsaką pradinėje koordinačių sistemoje (nereikia spręsti jokių lygčių)

$$\boxed{M = T_L W_L^T}$$