PDV - 2.10.2026

piatok 2. októbra 2026 10:50

Náhradná hodina → utorok 6.10., 17:10

Minule: Typy dát, typy analýz

Typy dát: - kvant.
- kategorické / ordinálne : usporiadané
nominálne

nominálne: nedobré kódovať 0, 1, ..., k-1

SK, HU, CZ → rozumné: one-hot-encoding: každú kategóriu kódovať samostatnou 0/1 premennou

| Krajina | SK | HU | CZ |
|---|---|---|---|
| SK | 1 | 0 | 0 |
| SK | 1 | 0 | 0 |
| HU | 0 | 1 | 0 |
| CZ | 0 | 0 | 1 |

alternatíva: one-hot-encoding, ale vynechám jeden zo stĺpcov (iba k-1 prem., ak bolo k kateg): dummy premenné

↑ jedna vieme dodedukovať

| Krajina | HU | CZ |
|---|---|---|
| SK | 0 | 0 |
| SK | 0 | 0 |
| HU | 1 | 0 |
| CZ | 0 | 1 |

cor(G3, Hjob_at_home) = -0,12; t's matkou doma majú mierne horšie výsledky

Typy analýz:

- supervised learning – máme cieľovú premennú (výška platu; spam / normálny mail) (učenie s učiteľom)
- unsupervised learning – nemáme cieľovú prem., chceme "niečo" získať z dát (učenie bez učiteľa)
  - → exploratívna analýza
  - → zhlukovanie
  - → redukcia dimenzie (PCA)

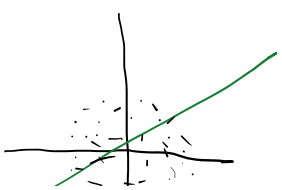
supervised → predikcia: cieľová prem. — kvantitatívna (výška platu) => regresný problém
kategorická => klasifikačný problém (spam – nespam)

Metóda hlavných komponentov

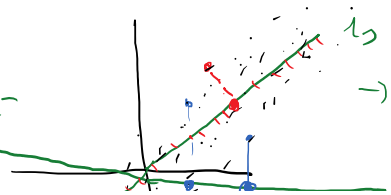
(Principal component analysis – PCA)

→ máme veľkorozmerné dáta => znížiť # rozmerov (napr. na 2/3)

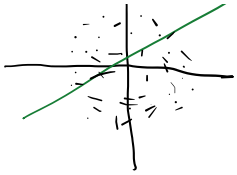
Zrušíme: 2D → 1D

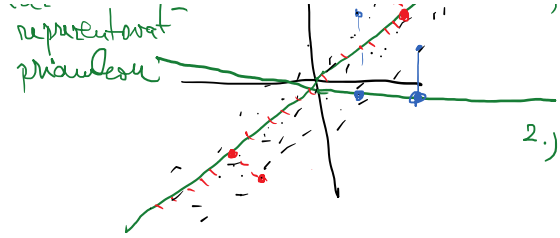


→ ťažké reprezentovať priamkou



→ dá sa rozumne reprezentovať priamkou






reprezentovať priamkou

reprezentovať priamkou

1.) je lepšia ako 2.)
- body sú bližšie k 1.) (súčet vzdialeností bodov od priamky)
- v smere 1.) sú dáta viac rozptýlené!

PCA: zníženie dimenzie
: tak, aby sme stratili čo najmenej informácie
: informáciu meria rozptýlenosťou

matematicky: $X = \begin{bmatrix} — & x_1^T & — \\ & \vdots & \\ — & x_n^T & — \end{bmatrix}$ $n \times p$

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ : vektor priemerov

napr. $\bar{x} = \begin{pmatrix} \overline{\text{vek}} \\ \overline{\text{výška}} \end{pmatrix}$

variančná matica $S = \begin{bmatrix} s_{x_1}^2 & \cdots & \hat{cov}(X_1, X_p) \\ \vdots & \ddots & \vdots \\ & \cdots & s_{x_p}^2 \end{bmatrix} = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^T$

1.) vycentrujeme dáta: $x_i \longmapsto x_i - \bar{x}$ (napr. $\begin{pmatrix} \text{vek}_i \\ \text{výška}_i \end{pmatrix} - \begin{pmatrix} \overline{\text{vek}} \\ \overline{\text{výška}} \end{pmatrix}$)

$\Rightarrow S = \frac{1}{n-1}\sum_{i=1}^{n} x_i x_i^T = \frac{1}{n-1} \underbrace{\begin{bmatrix} | & & | \\ x_1 & \cdots & x_n \\ | & & | \end{bmatrix}}_{X^T} \underbrace{\begin{bmatrix} — & x_1^T & — \\ & \vdots & \\ — & x_n^T & — \end{bmatrix}}_{X} = \frac{1}{n-1} X^T X$ : $p \times p$

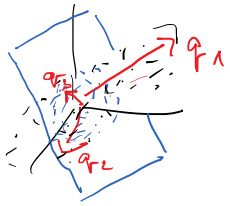
$p \times n \quad n \times p$

$p \times p$

2.) smer s max. rozptylom (disperziou):

$S \rightarrow$ vl. hodnoty $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$

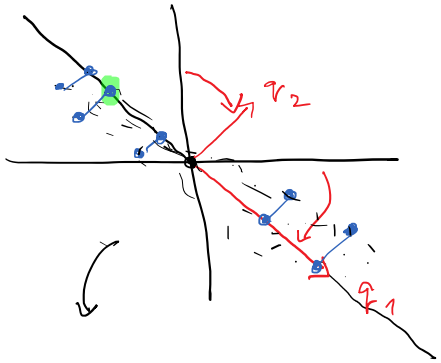
$\rightarrow$ vl. vektory $q_1, q_2, \ldots, q_p$ - ortonormálne




$q_1$: smer max. disperzie

$q_2$: smer max. disperzie vo svete kolmom na $q_1$

$q_3$: — " — kolmom na $q_1, q_2$



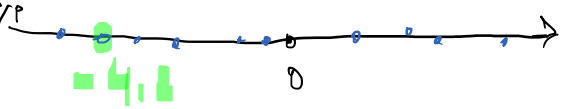

3.) dáta premietneme do nájdeného smeru
$\rightarrow$ pre 1-D reprezentáciu

viac rozmerov: vyjadríme dáta v súradnicovom systéme $q_1, \ldots, q_p$

$\rightarrow$ natočíme súr. sústavu, aby $q_1, \ldots, q_p$ boli súradnicové osi

$\Rightarrow$ osi budú zoradené podľa dôležitosti

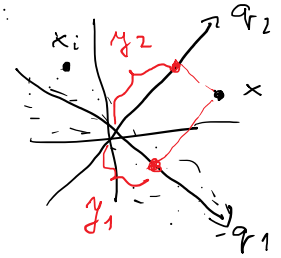
$\Rightarrow$ zoberiem prvých "s" rozmerov (ak chcem s-rozmernú reprezentáciu)



$\Rightarrow$ osi budú zoradené podľa dôležitosti

$\Rightarrow$ zoberiem prvých "s" rozmerov (ak chcem s-rozmernú reprezentáciu)

2-D:




$y_1, y_2$: súradnice $x$ voči $q_1, q_2$

$$x = y_1 \cdot q_1 + y_2 q_2$$

$$\underset{2\times 1}{x} = \underbrace{\begin{bmatrix} | & | \\ q_1 & q_2 \\ | & | \end{bmatrix}}_{2\times 2} \underbrace{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}}_{2\times 1}$$

$Q$ - ortogonálna mat.: $Q^T Q = I$

$$x = Q y \quad / \cdot Q^T$$

$$Q^T x = y$$

vo $p$-rozmere: $y = Q^T x$, $\quad Q = \begin{bmatrix} | & & | \\ q_1 & \cdots & q_p \\ | & & | \end{bmatrix}_{p\times p}$

$\forall x_i$: $y_i = Q^T x_i$, $i = 1, \ldots, n$

↑ nová reprezentácia

všetky: $Y = XQ$ (nxp = nxp pxp)

$$Y = \begin{bmatrix} — & y_1^T & — \\ & \vdots & \\ — & y_n^T & — \end{bmatrix}_{n\times p}$$

4.)

- stĺpce $Y$ sú zoradené podľa dôležitosti $\Rightarrow$ zoberiem prvých s stĺpcov (napr. $s=2$)

$$Y_{\text{čiastočné}} = \begin{bmatrix} — & y_1^T & — \\ & \vdots & \\ — & y_n^T & — \end{bmatrix}$$

s

- $Y$: skóre
  $Q$: rotation (matrix)
- platí: $\lambda_1$: variancia v smere $q_1$
  $\vdots$
  $\lambda_p$: variancia v smere $q_p$
  : $Y_1, \ldots, Y_p$: navzájom kolmé (nekorelované)

$\lambda_1 + \ldots + \lambda_p$: variancia v dátach celková

$\Rightarrow \dfrac{\lambda_1 + \ldots + \lambda_s}{\lambda_1 + \ldots + \lambda_p}$: akú časť variancie zachytáva prvých $s$ hlavných komponentov

$Y_1, \ldots, Y_p$: nové premenné - hlavné komponenty (principal components - PC)

