

Hodnotiace miery

Miery vhodnosti príznakov

Filter:

Konzistencia

Medzitriedna vzdialenosť

Štatistická závislosť

Informačno-teoretické miery

Obálka:

Dosiahnutá chyba klasifikátora

Informačno-teoretické miery

Hartleyho miera informácie

Správa dĺžky n

Počet rôznych symbolov s

Množstvo informácie v správe je funkciou počtu možných správ $N = s^n$:

$$\mathcal{I} = f(N)$$

Informačno-teoretické miery

Uvažujme dve správy dĺžky n_1 a n_2

Správy spojíme do jednej správy

$$\mathfrak{I} = \mathfrak{I}_1 + \mathfrak{I}_2$$

$$f(s^{n_1+n_2}) = f(s^{n_1}) + f(s^{n_2})$$

$$f(N_1 \cdot N_2) = f(N_1) + f(N_2)$$

Ktorá funkcia spĺňa?

Informačno-teoretické miery

Hartleyho miera informácie

$$\mathfrak{I} = \log(N) = \log(s^n) = n \cdot \log(s)$$

Shannonova miera informácie

Majme diskretnú náhodnú premennú A s možnými výstupmi $\{a_1, \dots, a_n\}$.

Pravdepodobnostná funkcia je $P(A=a_i)=p_i$

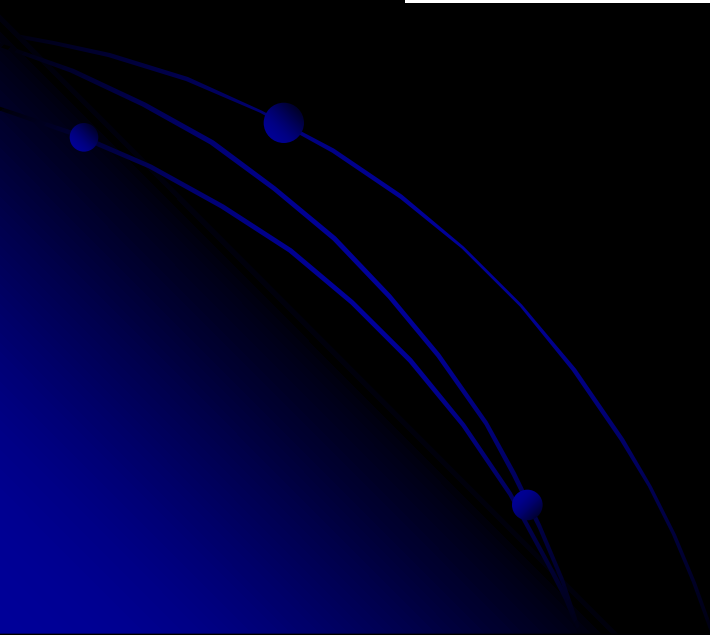
Informácia, ktorú dostaneme, keď pozorujeme výstup a_i je

$$\mathfrak{I}(a_i) = -\log_2(P(A = a_i))$$

Shannonova entropia

Entropia (neistota) = stredná hodnota informácie

$$\begin{aligned} H(A) &= E(\mathfrak{I}(A)) \\ &= -E(\log_2(P(A))) = \\ &= -\sum_{a \in \Omega} P(A=a) \cdot \log_2(P(A=a)) \end{aligned}$$

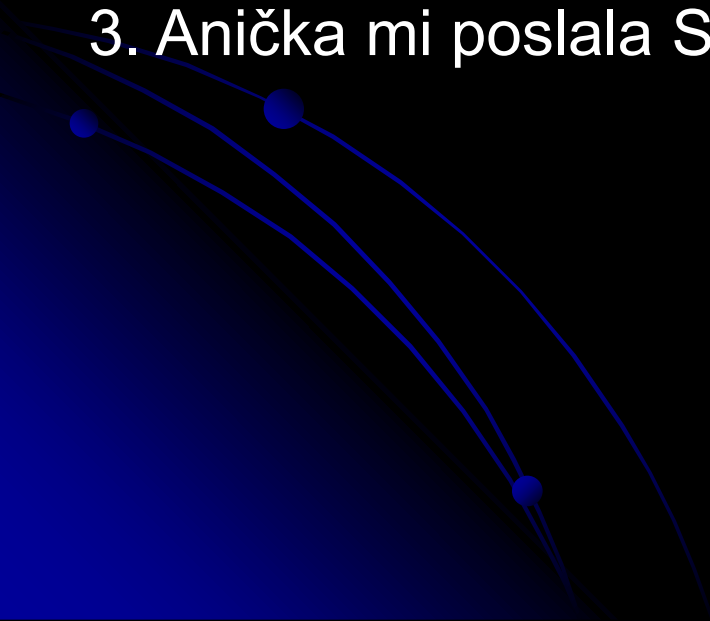


Shannonova entropia

Jožko: „Dostal som na skúške A.“

Koľko informácie nesie táto správa?

1. Jožko má väčšinou A
2. Jožko je flákač
3. Anička mi poslala SMS, že Jožko dostal A



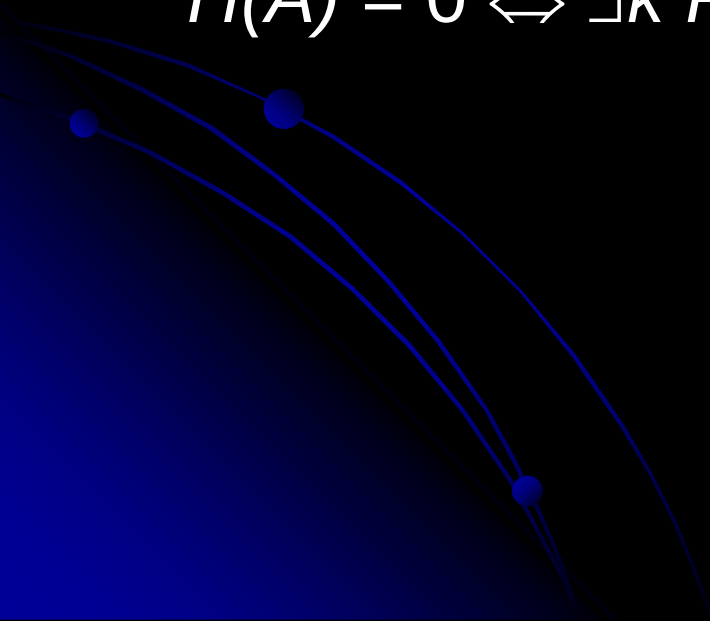
Vlastnosti Shannonovej entropie

$$H(A) \leq \log(N)$$

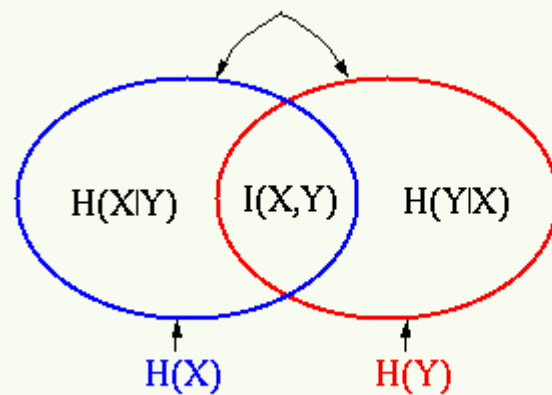
$$H(A) = \log(N) \Leftrightarrow \forall i \, P(A=a_i) = 1/N$$

$$H(A) \geq 0$$

$$H(A) = 0 \Leftrightarrow \exists k \, P(A=a_k) = 1$$



$$H(X,Y) = H(X|Y) + H(Y|X) + I(X,Y)$$



$$I(X,Y) = H(X) + H(Y) - H(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Informačno-teoretické miery

Ohodnotenie množiny príznakov

$$J(\tilde{X}) = I(\tilde{X}; y)$$

Celá podmnožina

$$X^* = \arg \max_{X_S \subset X} I(X_S; y)$$

Jednotlivé príznaky

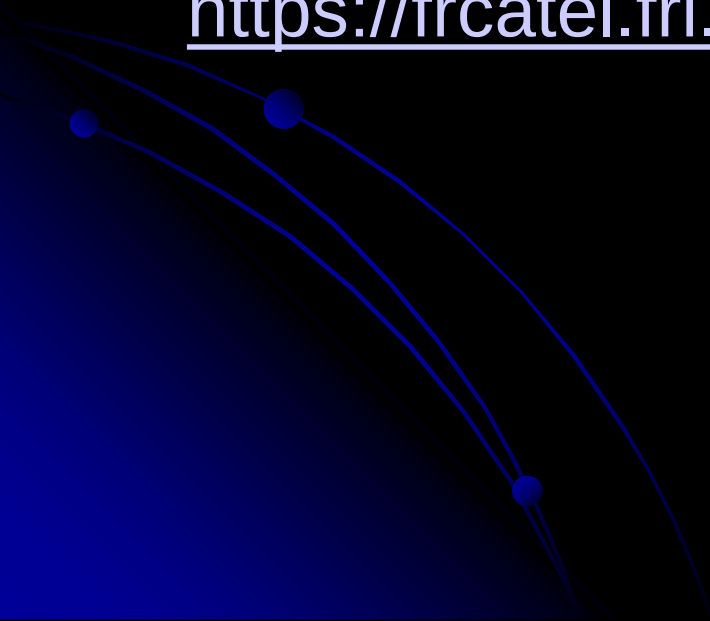
$$x^* = \arg \max_{x_k \in X \setminus X_S} I(\{X_S, x_k\}; y)$$

Elements Of Information Theory

Thomas M. Cover, Joy A. Thomas

Odvodenie formuly pre entropiu

<https://frcatel.fri.uniza.sk/users/paluch/ti.pdf>



Medzitriedna vzdialenosť

$$D_{\tilde{X}}(\omega_i, \omega_j) = \frac{1}{|\omega_i| |\omega_j|} \sum_{\mathbf{x} \in \omega_i} \sum_{\mathbf{y} \in \omega_j} d_{\tilde{X}}(\mathbf{x}, \mathbf{y})$$

$$J(\tilde{X}) = \sum_{i=1}^C P(\omega_i) \sum_{j=i+1}^C P(\omega_j) D_{\tilde{X}}(\omega_i, \omega_j)$$

Vzdialenosť

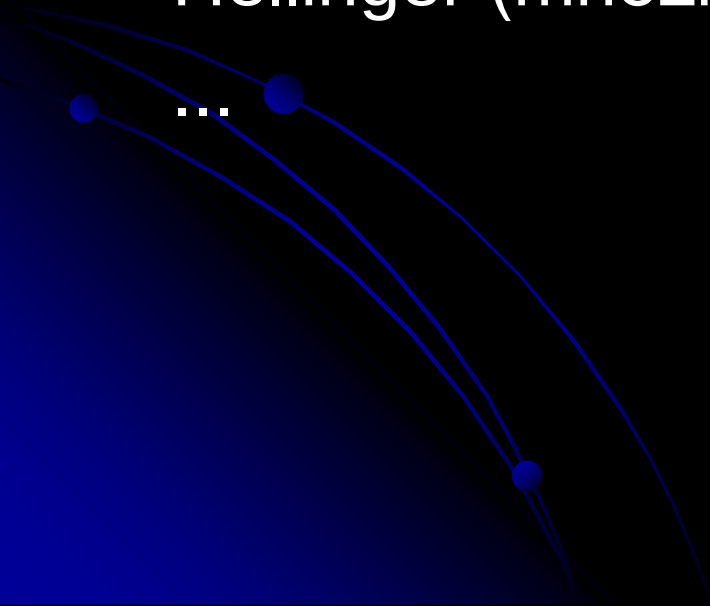
Euklidovská (bod – bod)

Mahalanobis (bod – množina)

Bhattacharya (množina – množina)

Hellinger (množina – množina)

...



Hodnotiace miery

Miery vhodnosti príznakov

Filter:

Konzistencia

Medzitriedna vzdialenosť

Štatistická závislosť

Informačno-teoretické miery

Obálka:

Dosiahnutá chyba klasifikátora

Hodnotiace miery

Miery vhodnosti príznakov

Wrapper:

Prediktívna schopnosť množiny trénovacích príznakov
(kvalita rozpoznávania pre testovacie dáta)

krížová validácia



Hľadanie optimálnej podmnožiny

Forward selection

```
Initialize s={}
Do:
Add feature to s
which improves OF(s) most
While OF(s) can be improved
```

Backward elimination

```
Initialize s={1,2,...,n}
Do:
remove feature from s
which improves OF(s) most
While OF(s) can be improved
```

- Spätná eliminácia nachádza lepšie modely
- Problém pri veľkých dátach na začiatku procesu
- Obe môžu byť priveľmi greedy

Simulované žíhanie

1. Určenie chladiacej funkcie $T_{i+1} = f(T_i)$

2. Určenie počiatočného riešenia $Y_0 = \{0,1\}^D$

3. Kým $T_i > T_{MIN}$

1. Nájdenie nového riešenia Y_{i+1} lokálnym prehľadávaním okolia Y_i

2. Výpočet $\Delta E = J(Y_i) - J(Y_{i+1})$

3. Ak $\Delta E < 0$

Akceptovanie riešenia

Inak

Akceptovanie s pravdepodobnosťou

$$P = \min(1, e^{-\Delta E / T_i})$$

4. $T_{i+1} = f(T_i)$

Genetické algoritmy

1. Vytvorenie počiatočnej náhodnej populácie $\{Y_i\}_{i=1}^N$
2. Ohodnotenie jedincov populácie $J(Y_i)$
3. Opakuj

1. Náhodný výber jedincov – rodičov na základe ohodnotenia
2. Vytvorenie populácie potomkov krížením

0100	1010111
1101	0001110

1101	1010111
0100	0001110

3. Mutácia potomkov

1101	1010111
0100000111	0

1101	0010111
0100000111	1

4. Ohodnotenie populácie

2 prístupy

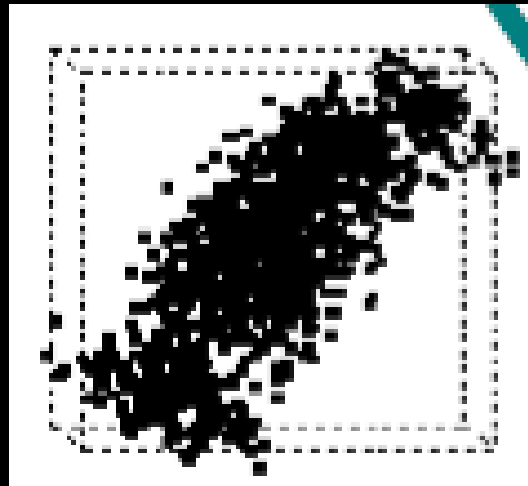
Výber príznakov:

vyberieme podmnožinu z originálnych príznakov

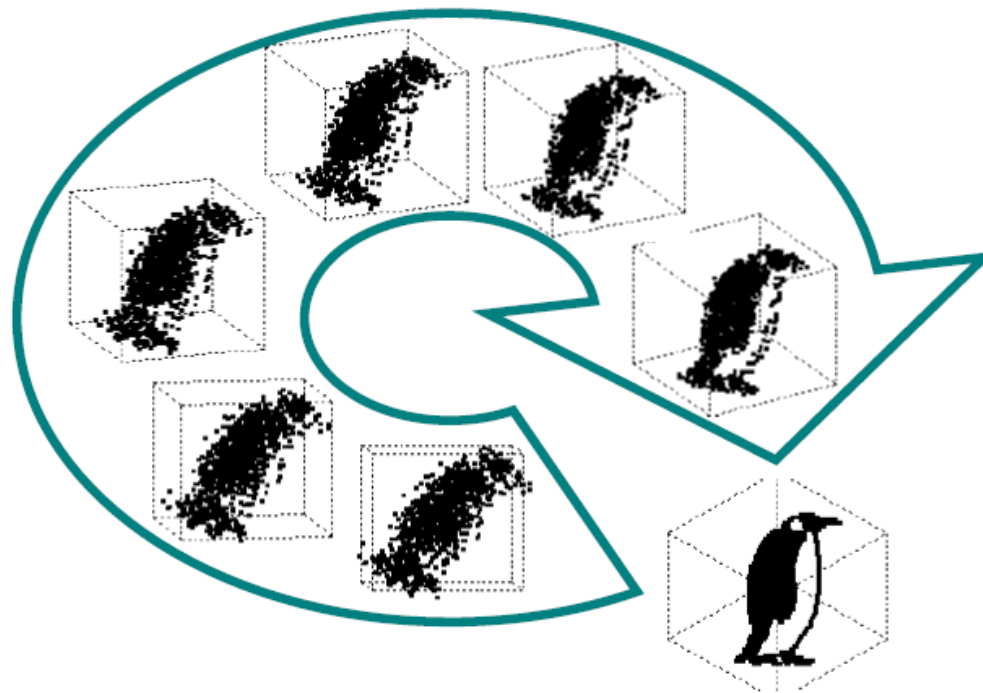
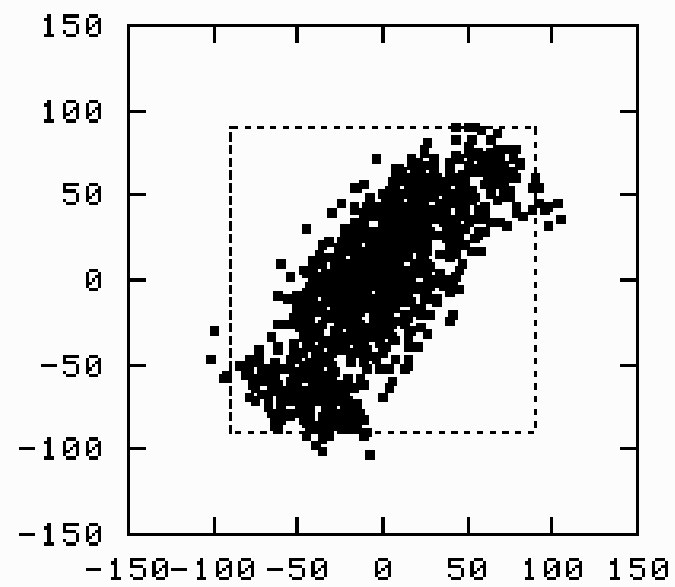
Redukcia príznakov:

transformujeme pôvodnú množinu do menej-dimenzionálnej

Dáta v 3D priestore



Zmysluplný smer premietania do 2D?



Redukcia príznakov

Neriadené (minimalizujú stratu informácie)

Principal Component Analysis (PCA)

Latent Semantic Indexing (LSI)

Independent Component Analysis (ICA)

Riadené (maximalizujú medzitriedne rozdiely)

Linear Discriminant Analysis (LDA)

Canonical Correlation Analysis (CCA)

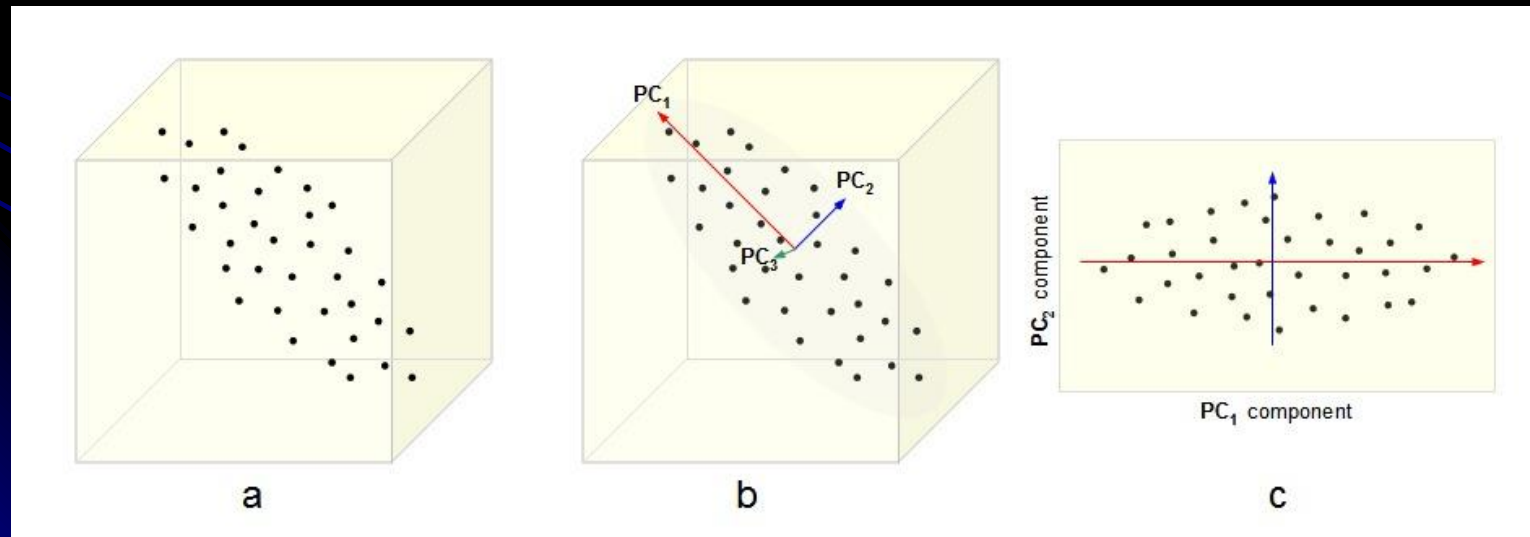
Partial Least Squares (PLS)

Principal Component Analysis (PCA)

Metóda hlavných komponentov

Karhunen-Loeve, K-L method

PCA hľadá „podpriestor“, ktorý zachytáva čo najviac variácie v dátach



PCA

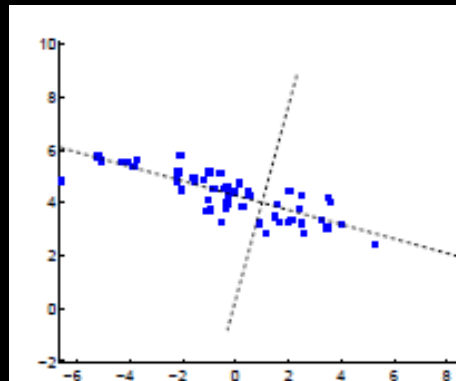
otočí a posunie súradnicovú sústavu tak, aby prvá os bola v smere najväčšej variability dát a ďalšie boli na ňu kolmé v smeroch najväčšej zvyšnej variability

Máme N D -rozmerných vektorov príznačkov $[\mathbf{x}_1, \dots, \mathbf{x}_N]$

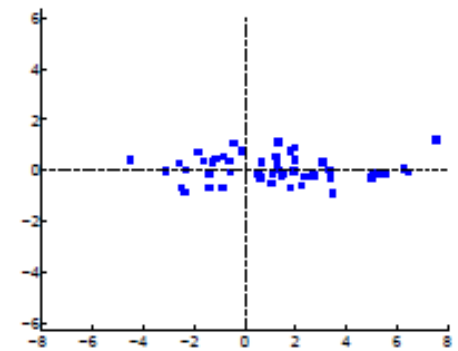
nová ortonormálna báza

$$\{\mathbf{b}_1, \dots, \mathbf{b}_D\}$$

$$\mathbf{b}_i^T \mathbf{b}_j = \delta_{ij}$$



(a) Originálne dáta.



(b) Transformované dáta.

priemet vektora $x'_{i1} = \mathbf{b}_1^T \mathbf{x}_i$

priemer $\bar{x}'_1 = \mathbf{b}_1^T \bar{\mathbf{x}}$ $\bar{\mathbf{x}} = \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i$

variancia

$$\begin{aligned} Var_1 &= \frac{1}{N} \sum_{i=1}^N (x'_{i1} - \bar{x}'_1)^2 = \frac{1}{N} \sum_{i=1}^N (\mathbf{b}_1^T \mathbf{x}_i - \mathbf{b}_1^T \bar{\mathbf{x}})^2 = \\ &= \frac{1}{N} \sum_{i=1}^N (\mathbf{b}_1^T (\mathbf{x}_i - \bar{\mathbf{x}}))^2 = \mathbf{b}_1^T \boldsymbol{\Sigma} \mathbf{b}_1, \end{aligned}$$

prvá os v smere najväčšej variability –
maximalizačná úloha s viazným extrémom

Lagrangeova metóda neurčitých koeficientov

$$\begin{aligned} \max_{\mathbf{b}_1} \mathbf{b}_1^T \Sigma \mathbf{b}_1, \\ \text{aby } \mathbf{b}_1^T \mathbf{b}_1 = 1. \end{aligned}$$

$$L = \mathbf{b}_1^T \Sigma \mathbf{b}_1 - \lambda(\mathbf{b}_1^T \mathbf{b}_1 - 1)$$

$$\frac{\partial L}{\partial \mathbf{b}_1} = 2\Sigma \mathbf{b}_1 - 2\lambda \mathbf{b}_1 \equiv 0$$

$$\Sigma \mathbf{b}_1 = \lambda \mathbf{b}_1$$

vlastné číslo

$$\mathbf{b}_1^T \Sigma \mathbf{b}_1 = \lambda \mathbf{b}_1^T \mathbf{b}_1 = \lambda$$

$$Var_2 = \mathbf{b}_2^T \Sigma \mathbf{b}_2.$$

$$\begin{aligned} \max_{\mathbf{b}_2} \mathbf{b}_2^T \Sigma \mathbf{b}_2, \\ \text{aby } \mathbf{b}_2^T \mathbf{b}_2 = 1, \\ \mathbf{b}_1^T \mathbf{b}_2 = 0. \end{aligned}$$

$$L = \mathbf{b}_2^T \Sigma \mathbf{b}_2 - \lambda(\mathbf{b}_2^T \mathbf{b}_2 - 1) - \mu \mathbf{b}_1^T \mathbf{b}_2$$

$$\frac{\partial L}{\partial \mathbf{b}_2} = 2\Sigma \mathbf{b}_2 - 2\lambda \mathbf{b}_2 - \mu \mathbf{b}_1 \equiv 0$$

$$\Sigma \mathbf{b}_2 = \lambda \mathbf{b}_2,$$

vlastné číslo

...

smery nových bázových vektorov $\{\mathbf{b}_1, \dots, \mathbf{b}_D\}$

Nech počiatok súradnicovej sústavy: $\mathbf{p}$

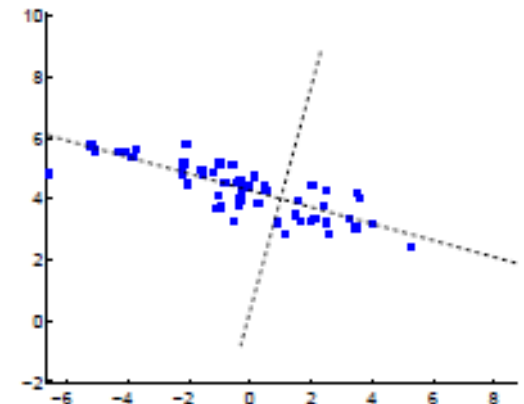
$$\mathbf{y}_i = \mathbf{p} + \sum_{j=1}^N y_{ij} \mathbf{b}_j.$$

chceme aby suma štvorcov vzdialeností medzi pôvodnými a premietnutými vektormi bola minimálna

$$\begin{aligned} E &= \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{y}_i\|^2 = \sum_{i=1}^N \left\| \mathbf{x}_i - \left(\mathbf{p} + \sum_{j=1}^N y_{ij} \mathbf{b}_j \right) \right\|^2 = \\ &= \sum_{i=1}^N \|\mathbf{x}_i - \mathbf{p}\|^2 - 2 \sum_{j=1}^N \sum_{i=1}^N y_{ij} \mathbf{b}_j^T (\mathbf{x}_i - \mathbf{p}) + \sum_{j=1}^N \sum_{i=1}^N y_{ij}^2. \end{aligned}$$

$$y_{ij} = \mathbf{b}_j^T (\mathbf{x}_i - \bar{\mathbf{x}})$$

$$\mathbf{p} = \bar{\mathbf{x}}.$$



Použitie PCA

Vypočítame maticu Σ

$$\Sigma = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T = \frac{1}{N} \mathbf{X} \mathbf{X}^T,$$

kde $[\mathbf{X}]_{D \times N} = [\mathbf{x}_1 - \bar{\mathbf{x}}, \dots, \mathbf{x}_N - \bar{\mathbf{x}}]$

Vypočítame vlastné vektory matice Σ

$$\Sigma \mathbf{b}_j = \lambda_j \mathbf{b}_j$$

Vypočítame súradnice premietnutých vektorov

$$\mathbf{x}'_i = \sum_{j=1}^D \mathbf{b}_j^T (\mathbf{x}_i - \bar{\mathbf{x}}) = \mathbf{B}^T (\mathbf{x}_i - \bar{\mathbf{x}}) = \mathbf{B}^T \mathbf{X},$$

kde $\mathbf{B} = [\mathbf{b}_1, \dots, \mathbf{b}_D]$.

$$N < D$$

$$r = \text{rank}(\mathbf{X}) = \text{rank}(\mathbf{\Sigma}) \leq N$$

najviac r vlastných hodnôt je nenulových

$$\mathbf{\Sigma} = \frac{1}{N} \mathbf{X} \mathbf{X}^T$$

$$D \times D$$

$$O(D^3)$$

$$\mathbf{\Sigma} \mathbf{b}_j = \lambda_j \mathbf{b}_j$$

$$\frac{1}{N} \mathbf{X} \mathbf{X}^T \mathbf{b}_j = \lambda_j \mathbf{b}_j$$

$$\frac{1}{N} \mathbf{X}^T \mathbf{X} \mathbf{X}^T \mathbf{b}_j = \lambda_j \mathbf{X}^T \mathbf{b}_j$$

$$\frac{1}{N} \mathbf{X}^T \mathbf{X} \mathbf{p}_j = \lambda_j \mathbf{p}_j,$$

$\mathbf{p}_j = \mathbf{X}^T \mathbf{b}_j$ je vlastný vektor $N \times N$ matice $\frac{1}{N} \mathbf{X}^T \mathbf{X}$

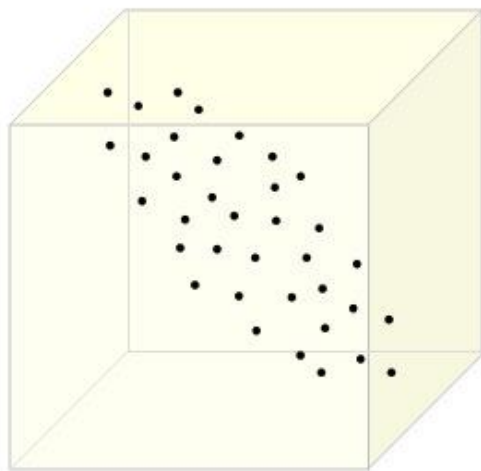
$$O(N^3)$$

$$\begin{aligned}
 \frac{1}{N} \mathbf{X}^T \mathbf{X} \mathbf{p}_j &= \lambda_j \mathbf{p}_j \\
 \frac{1}{N} \mathbf{X} \mathbf{X}^T \mathbf{X} \mathbf{p}_j &= \lambda_j \mathbf{X} \mathbf{p}_j \\
 \Sigma \mathbf{X} \mathbf{p}_j &= \lambda_j \mathbf{X} \mathbf{p}_j
 \end{aligned}
 \Rightarrow \mathbf{b}_j \propto \mathbf{X} \mathbf{p}_j$$

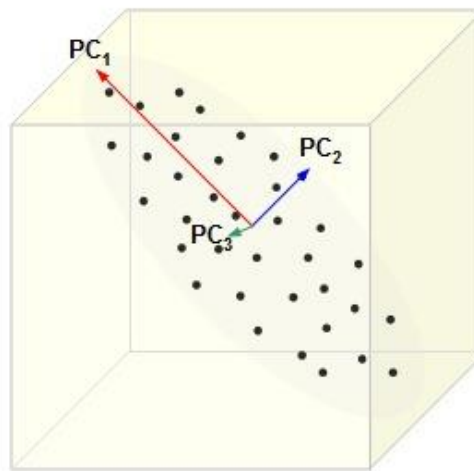
Za predpokladu $\|\mathbf{p}_j\| = 1$ hľadáme také $\mathbf{b}_j$, aby $\|\mathbf{b}_j\| = 1$

$$\mathbf{b}_j = \frac{1}{\sqrt{N \lambda_j}} \mathbf{X} \mathbf{p}_j$$

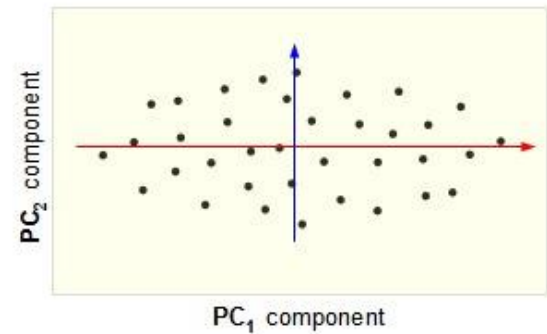
Zníženie dimenzie



a

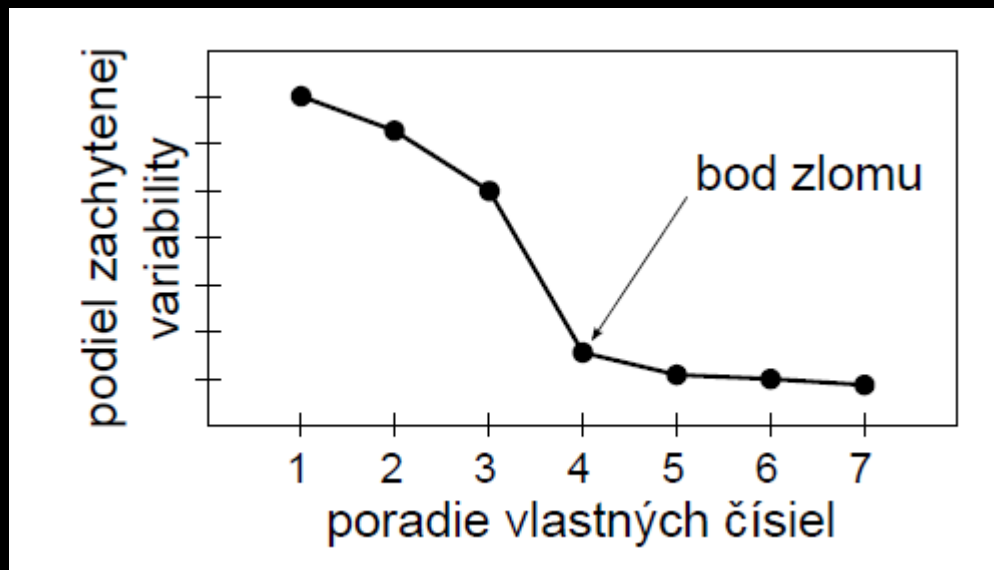


b



c

počet hlavných komponentov



Scree graf

v bode zlomu, od štvrtého najväčšieho vlastného čísla začínajú vlastné čísla zachycujúce málo variability

optimálny počet vlastných čísiel je 3

Podiel variability zachytenej v j -tom hlavnom komponente

$$\frac{\lambda_j}{\sum_{j=1}^D \lambda_j}$$

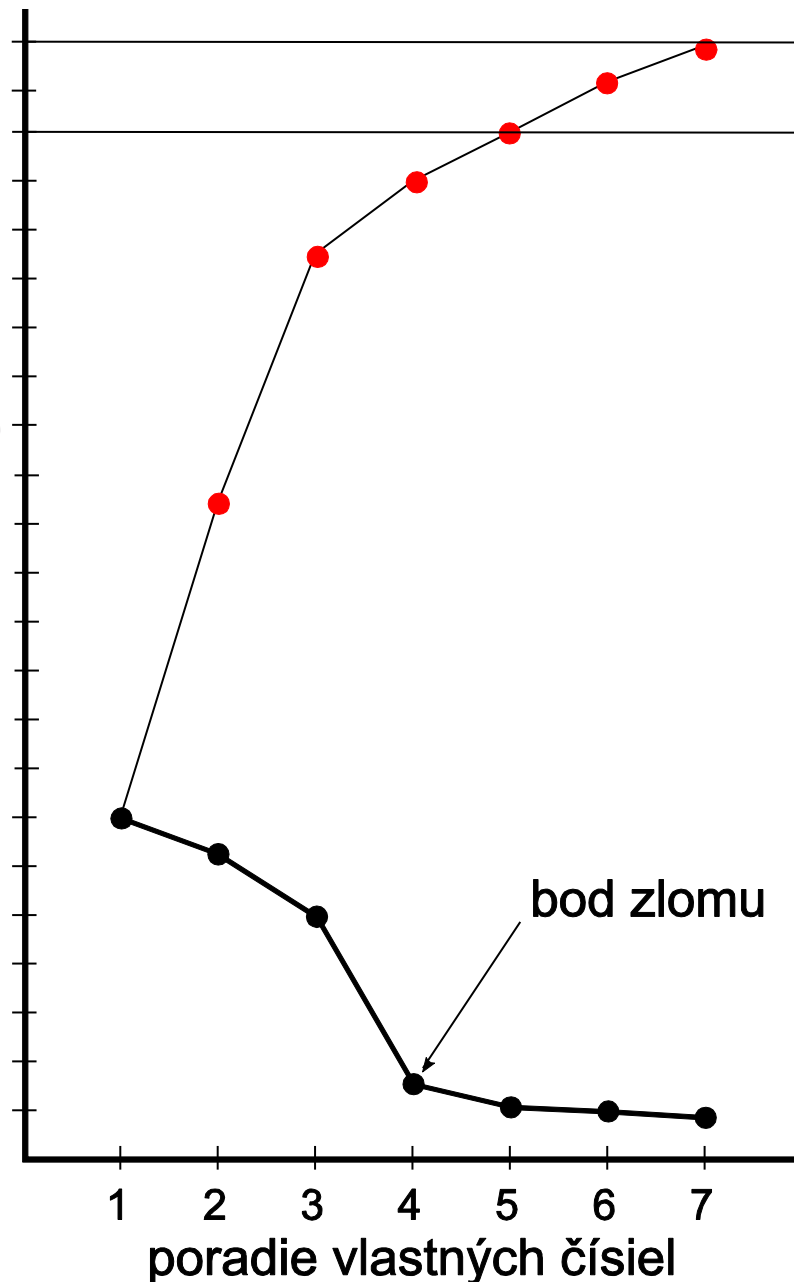
kumulatívny pomer zachytenej variability

$$\frac{\sum_{j=1}^K \lambda_j}{\sum_{j=1}^D \lambda_j} > 0.9 \text{ alebo } 0.95.$$

podiel
variability

poradie vlastných čísel

bod zlomu



Andersonov test

testovanie hypotézy

$$H_0 : \lambda_{K+1} = \lambda_{K+2} = \dots = \lambda_D$$

$$H_1 : \text{neplatí } H_0.$$

Testovacia štatistika je $V = n(D - K) \ln \frac{a}{c}$

$$a = \frac{\sum_{j=K+1}^D \lambda_j}{D-K}$$

$$c = \left(\prod_{j=K+1}^D \lambda_j \right)^{\frac{1}{D-K}}$$

χ^2 rozdelenie s $0.5(D - K + 2)(D - K - 1)$ stupňami voľnosti

SVD

Naspät' k problému $N < D$

singulárny rozklad matice $\mathbf{A}$

Každá matice $\mathbf{A}$ sa dá rozložiť na súčin

$$\mathbf{A} = \mathbf{U} \mathbf{S} \mathbf{V}^T$$

$$[\mathbf{A}]_{N \times D} = [\mathbf{U}]_{N \times N} [\mathbf{S}]_{N \times D} [\mathbf{V}^T]_{D \times D}$$

$$\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_N], \mathbf{V} = [\mathbf{v}_1, \dots, \mathbf{v}_D] \text{ a } \mathbf{S} = \text{diag}(\sigma_1, \dots, \sigma_{\text{rank}(\mathbf{A})})$$

→ $\mathbf{v}$ je vlastný vektor matice $\mathbf{A}^T \mathbf{A}$

$\mathbf{u}$ vlastný vektor matice $\mathbf{A} \mathbf{A}^T$

$$\lambda = \sigma^2$$

$$\Sigma = \frac{1}{N} \mathbf{X} \mathbf{X}^T$$

Vzťah medzi PCA a SVD

$$\mathbf{X} = [\mathbf{x}_1 - \bar{\mathbf{x}}, \dots, \mathbf{x}_N - \bar{\mathbf{x}}]$$

$$\mathbf{A} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

$\mathbf{v}$ je vlastný vektor matice $\mathbf{A}^T \mathbf{A}$

$$\lambda = \sigma^2$$

$$\mathbf{Y} = \frac{1}{\sqrt{N}} \mathbf{X}^T$$



$$\mathbf{Y}^T \mathbf{Y} = \frac{1}{N} \mathbf{X} \mathbf{X}^T = \mathbf{\Sigma}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{S}\mathbf{V}^T$$

$\mathbf{V}$ sú vlastné vektory matice $\mathbf{Y}^T \mathbf{Y} = \mathbf{\Sigma}$

použitie SVD namiesto PCA

SVD algoritmus je numericky stabilnejší ako výpočet PCA