



ANALÝZA A KLASIFIKACE DAT



prof. Ing. Jiří Holčík, CSc.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

VIII. ANALÝZA HLAVNÍCH KOMPONENT



ZAČÍNÁME

ANALÝZA HLAVNÍCH KOMPONENT

PRINCIPAL COMPONENT ANALYSIS (PCA)

ROZKLAD PODLE VLASTNÍCH ČÍSEL

SINGULAR VALUE DECOMPOSITION (SVD)

Karhunenova-Loevova transformace

ZAČÍNÁME

- ☑ **extrakce příznaků** - hledání zobrazení (optimálního) Z , které transformuje původní m rozměrný prostor (obraz) na prostor (obraz) n rozměrný ($m \geq n$);
- ☑ **nalezení vhodné transformace** – potřeba optimalizačního kritéria:
 - obrazy v novém prostoru budou aproximovat původní obrazy ve smyslu minimální střední kvadratické odchylky;
 - obrazy v novém prostoru budou minimalizovat odhad pravděpodobnosti chyby

ZAČÍNÁME

- ☑ aby byla úloha řešitelná, hledáme zobrazení v oboru lineárních zobrazení

ZAČÍNÁME

- ☑ aby byla úloha řešitelná, hledáme zobrazení v oboru lineárních zobrazení

Jak poznáme lineární zobrazení?

ZAČÍNÁME

- ☑ aby byla úloha řešitelná, hledáme zobrazení v oboru lineárních zobrazení

Jak poznáme lineární zobrazení?

$$\mathbf{x} = \mathbf{Z} \cdot \mathbf{y}$$

TEORIE

- ☑ předpokládejme, že je dáno K obrazů a nechť existuje m příznakových veličin, které tyto obrazy charakterizují. Tedy k -tý obraz je vyjádřen m rozměrným sloupcovým vektorem $\mathbf{y}_k \in \mathcal{Y}^m$, $k=1, \dots, K$.
- ☑ aproximujme nyní kterýkoliv obraz \mathbf{y}_k lineární kombinací n ortonormálních vektorů \mathbf{e}_i ($m \geq n$)

$$\mathbf{x}_k = \sum_{i=1}^n c_{ki} \mathbf{e}_i. \quad (\text{☺})$$

TEORIE

- ☑ koeficienty c_{ki} lze považovat za velikost i -té souřadnice vektoru \mathbf{y}_k vyjádřeného v novém systému souřadnic s bází \mathbf{e}_i , $i=1,2,\dots,n$, tj. platí

$$c_{ki} = \mathbf{y}_k \cdot \mathbf{e}_i.$$

- ☑ použijeme-li jako kritérium minimální střední kvadratické odchylky, pak je

$$\varepsilon_k^2 = \|\mathbf{y}_k - \mathbf{x}_k\|^2.$$

TEORIE

- ☑ pak pomocí dříve uvedených vztahů pro \mathbf{x}_k a c_{ki} dostaneme

$$\varepsilon_k^2 = \|\mathbf{y}_k\|^2 - \sum_{i=1}^n c_{ki}^2.$$

- ☑ střední kvadratická odchylka pro všechny obrazy \mathbf{y}_k , $k=1, \dots, K$ je

$$\varepsilon^2 = \frac{1}{K} \sum_{k=1}^K \varepsilon_k^2 = \frac{1}{K} \sum_{k=1}^K \|\mathbf{y}_k\|^2 - \sum_{i=1}^n \mathbf{e}_i^T \left[\frac{1}{K} \sum_{k=1}^K \mathbf{y}_k \cdot^T \mathbf{y}_k \right] \cdot \mathbf{e}_i.$$

TEORIE

- ☑ pak pomocí dříve uvedených vztahů pro \mathbf{x}_k a c_{ki} dostaneme

$$\varepsilon_k^2 = \|\mathbf{y}_k\|^2 - \sum_{i=1}^n c_{ki}^2.$$

- ☑ střední kvadratická odchylka pro všechny obrazy \mathbf{y}_k , $k=1, \dots, K$ je

$$\varepsilon^2 = \frac{1}{K} \sum_{k=1}^K \varepsilon_k^2 = \frac{1}{K} \sum_{k=1}^K \|\mathbf{y}_k\|^2 - \sum_{i=1}^n \mathbf{e}_i^T \left[\frac{1}{K} \sum_{k=1}^K \mathbf{y}_k \cdot^T \mathbf{y}_k \right] \cdot \mathbf{e}_i.$$

(je tedy závislá na volbě báze systému \mathbf{e}_i)

TEORIE

- ☑ diskrétní konečný rozvoj podle vztahu (☺) s bázovým systémem \mathbf{e}_i , optimálním podle kritéria minimální střední kvadratické chyby nazýváme diskrétní Karhunenův – Loevův rozvoj;
- ☑ aby střední kvadratická odchylka podle výše uvedeného vztahu byla minimální, musí být odečítaná hodnota na pravé straně rovnice maximální.

TEORIE

☑ musíme tedy maximalizovat výraz

$$\sum_{i=1}^n \mathbf{e}_i^T \kappa(\mathbf{y}) \cdot \mathbf{e}_i, \quad \text{kde} \quad \kappa(\mathbf{y}) = \frac{1}{K} \sum_{k=1}^K \mathbf{y}_k \cdot \mathbf{y}_k^T$$

je autokorelační matice řádu m . Protože je symetrická a semidefinitní, jsou její vlastní čísla λ_i , $i=1, \dots, m$, reálná a nezáporná a vlastní vektory \mathbf{v}_i , jsou buď ortonormální, nebo je můžeme ortonormalizovat (v případě násobných vlastních čísel).

TEORIE

- ☑ uspořádáme-li vlastní čísla sestupně podle velikosti, tj.

$$\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_m \geq 0$$

a podle toho očíslováme i odpovídající charakteristické vektory, lze dokázat, výe uvedený výraz dosahuje maxima, jestliže platí

$$\mathbf{e}_i = \mathbf{v}_i, i=1, \dots, n$$

a pro velikost maxima je

$$\max \sum_{i=1}^n \mathbf{e}_i^T \cdot \kappa(\mathbf{y}) \cdot \mathbf{e}_i = \sum_{i=1}^n \lambda_i$$

TEORIE

- ☑ pro minimální střední kvadratickou odchylku tedy platí

$$\begin{aligned}\epsilon_{\min}^2 &= \frac{1}{K} \sum_{k=1}^K \|\mathbf{y}_k\|^2 - \sum_{i=1}^n \lambda_i = \\ &= \text{tr}(\kappa(\mathbf{y})) - \sum_{i=1}^n \lambda_i = \sum_{i=n+1}^m \lambda_i\end{aligned}$$

TEORIE

- ☑ v některých případech je vhodnější vektory \mathbf{y}_k před aproximací centrovat se střední hodnotou

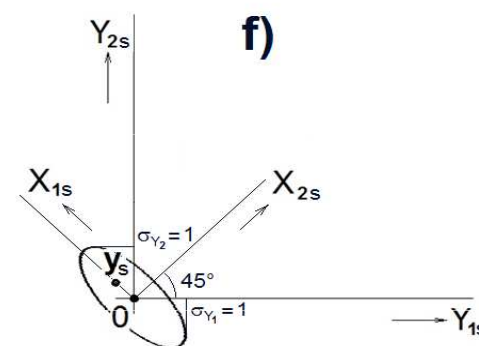
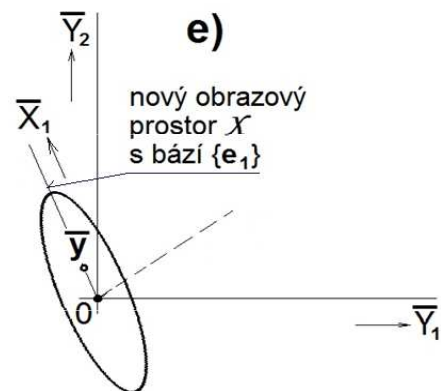
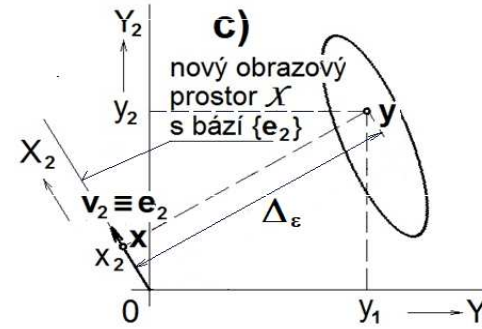
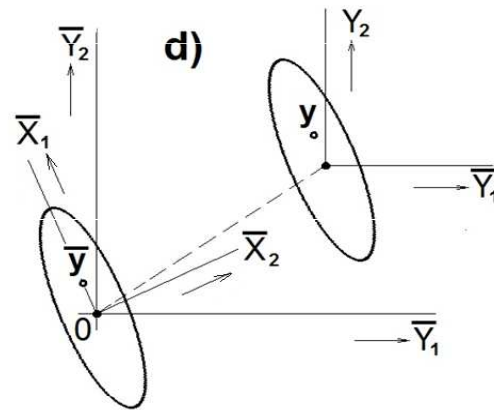
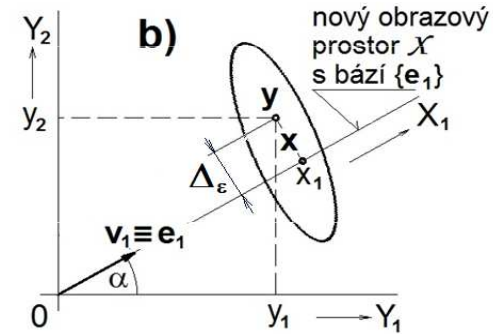
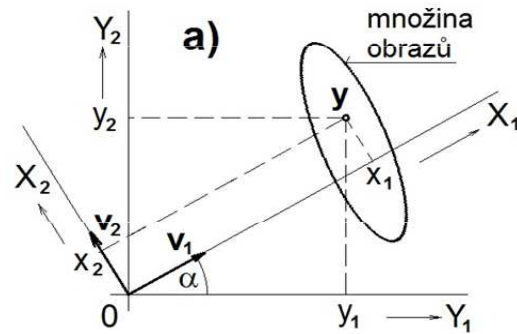
$$\boldsymbol{\mu} = \frac{1}{K} \sum_{k=1}^K \mathbf{y}_k$$

a místo s obrazem \mathbf{y}_k počítáme s jeho centrovanou verzí $\bar{\mathbf{y}}_k = \mathbf{y}_k - \boldsymbol{\mu}$.

Postup výpočtu se nemění, ale místo autokorelační matice používáme disperzní matici ve tvaru

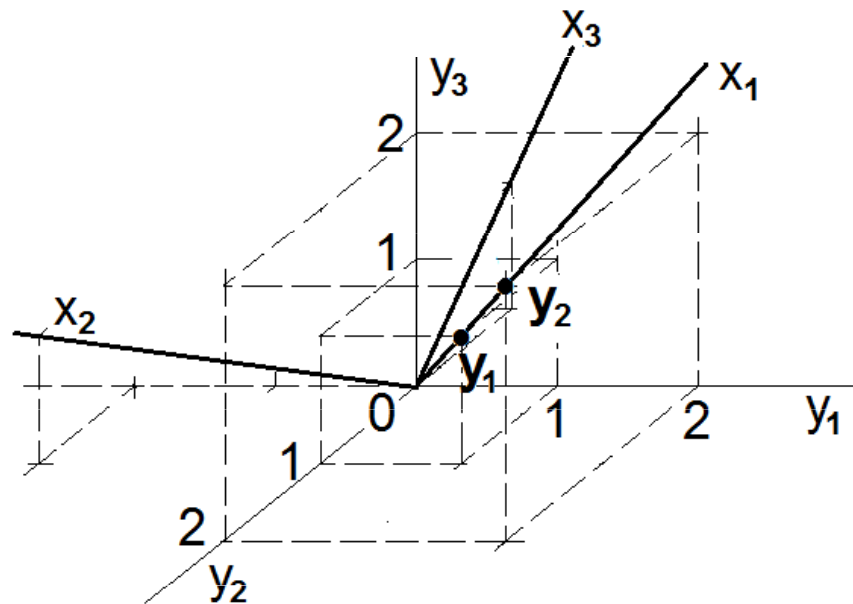
$$D(\mathbf{y}) = \frac{1}{K} \sum_{k=1}^K \bar{\mathbf{y}}_k \cdot \bar{\mathbf{y}}_k^T. \quad \text{Platí} \quad \kappa(\mathbf{y}) = D(\mathbf{y}) + \boldsymbol{\mu} \cdot \boldsymbol{\mu}^T.$$

GEOMETRICKÁ INTERPRETACE



PŘÍKLAD

Předpokládejme, že množinu obrazů γ^3 tvoří dva obrazové vektory $\mathbf{y}_1 = (1, 1, 1)^\top$ a $\mathbf{y}_2 = (2, 2, 2)^\top$. Pomocí Karhunenova – Loevova rozvoje najdeme novou souřadnicovou soustavu, která umožní popsat oba vektory s minimální střední kvadratickou odchylkou.



PŘÍKLAD

autokorelační matice:

$$\mathbf{\kappa} = \frac{1}{2}(\mathbf{y}_1 \cdot \mathbf{y}_1^T + \mathbf{y}_2 \cdot \mathbf{y}_2^T) = \frac{1}{2} \left[\begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \cdot (1 \ 1 \ 1) + \begin{pmatrix} 2 \\ 2 \\ 2 \end{pmatrix} \cdot (2 \ 2 \ 2) \right] = \begin{bmatrix} 2,5 & 2,5 & 2,5 \\ 2,5 & 2,5 & 2,5 \\ 2,5 & 2,5 & 2,5 \end{bmatrix}.$$

vlastní čísla:

$$\det \begin{bmatrix} 2,5 - \lambda & 2,5 & 2,5 \\ 2,5 & 2,5 - \lambda & 2,5 \\ 2,5 & 2,5 & 2,5 - \lambda \end{bmatrix} = 0$$

$$(2,5 - \lambda)^3 + 2,5^3 + 2,5^3 - 3 \cdot 2,5^2 \cdot (2,5 - \lambda) = 0$$

$$\lambda^3 - 7,5\lambda^2 = 0 \Rightarrow \lambda_1 = 7,5 \text{ a } \lambda_{2,3} = 0.$$

PŘÍKLAD

vlastní vektory: $[\kappa - \lambda \cdot \mathbf{I}] \cdot \mathbf{x} = 0.$

Pro $\lambda_1 = 7,5$ dostáváme lineární soustavu tří rovnic

$$-5x_1 + 2,5x_2 + 2,5x_3 = 0;$$

$$2,5x_1 - 5x_2 + 2,5x_3 = 0;$$

$$2,5x_1 + 2,5x_2 - 5x_3 = 0,$$

která obsahuje pouze dvě lineárně nezávislé rovnice a tedy její parametrické řešení je

$$x_1 = \frac{x_2 + x_3}{2}; \quad x_2 = x_3 \quad \text{a} \quad x_3 = t.$$

pro $t = 1$ je k vlastnímu číslu λ_1 vlastní vektor

$$\mathbf{x}_1 = (1, 1, 1)^T,$$

PŘÍKLAD

pro $\lambda_{2,3} = 0$

$$2,5x_1 + 2,5x_2 + 2,5x_3 = 0$$

$$2,5x_1 + 2,5x_2 + 2,5x_3 = 0$$

$$2,5x_1 + 2,5x_2 + 2,5x_3 = 0$$

$$x_1 = -x_2 - x_3; \quad x_2 = t \quad \text{a} \quad x_3 = u.$$

Parametry t a u volíme tak, aby vlastní vektory byly navzájem ortogonální,

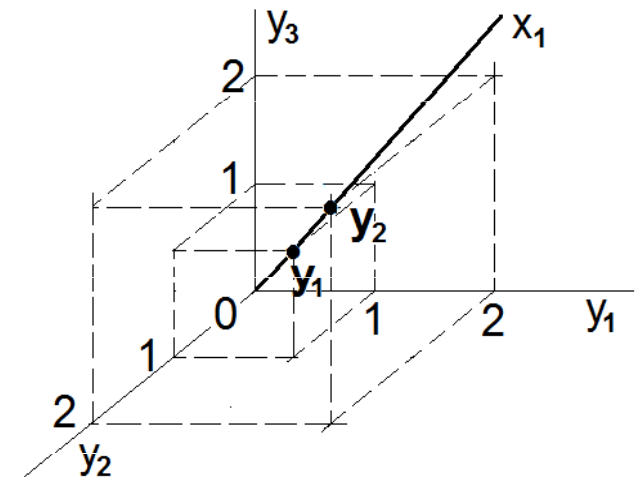
pro \mathbf{x}_2 např. $t = 1$ a $u = 1$, pak $\mathbf{x}_2 = (-2, 1, 1)^T$ a

pro \mathbf{x}_3 např. $t = -1$ a $u = 1$ a tedy $\mathbf{x}_3 = (0, -1, 1)^T$.

PŘÍKLAD

Co by se stalo, kdybychom odstranili souřadnici x_1 ?

Protože body y_1 a y_2 leží na vrcholech krychlí s hranami o délce 1, resp. 2 protilehlých k počátku, je jejich vzdálenost od počátku a tím i souřadnice ve směru x_1 rovna délce prostorové úhlopříčky, tj. $d_1 = \sqrt{3}$ v případě vektoru y_1 , resp. $d_2 = \sqrt{12}$ v případě vektoru y_2 .



Protože je nová souřadnicová soustava ortogonální, promítaly by se oba obrazové vektory při odstranění osy x_1 do počátku. A konečně, vzhledem k tomu, že chybu popisu obrazových vektorů ε^2 vyjadřujeme pomocí střední kvadratické odchylky, je tato chyba rovna

$$\varepsilon^2 = \frac{(d_1^2 + d_2^2)}{2} = \frac{1}{2}(3 + 12) = 7,5$$

což je právě λ_1 .

VLASTNOSTI

- ☑ při daném počtu n členů rozvoje poskytuje ze všech možných aproximací nejmenší střední kvadratickou odchylku;
- ☑ při použití disperzní matice jsou transformované souřadnice nekorelované; pokud se výskyt obrazů řídí normálním rozložením zajišťuje nekorelovanost i jejich nezávislost;
- ☑ vliv každého členu uspořádaného rozvoje se zmenšuje s jeho pořadím;
- ☑ změna požadavků na velikost střední kvadratické odchylky nevyžaduje přepočítávat celý rozvoj, nýbrž jen změnit počet jeho členů.

ROZDĚLENÍ DO TŘÍD

Jak se změní podmínky, když obrazy \mathbf{y} budou vymezeny jako části spojitého obrazového prostoru γ^m ?

- ✓ Výskyt obrazů v jednotlivých klasifikačních třídách bude popsán podmíněnými hustotami pravděpodobnosti $p(\mathbf{y}|\omega_r)$, $r=1,2,\dots,R$ a apriorní pravděpodobnost klasifikačních tříd bude $P(\omega_r)$.

V tom případě autokorelační matice bude

$$\kappa(\mathbf{y}) = \sum_{r=1}^R P(\omega_r) \cdot \int_{\gamma^m} \mathbf{y} \cdot^T \mathbf{y} \cdot p(\mathbf{y} | \omega_r) \cdot d\mathbf{y} = \int_{\gamma^m} \mathbf{y} \cdot^T \mathbf{y} \cdot p(\mathbf{y}) \cdot d\mathbf{y}$$

ROZDĚLENÍ DO TŘÍD

☑ disperzní matice

$$D^1(\mathbf{y}) = \sum_{r=1}^R P(\omega_r) \cdot \int_{\mathcal{Y}^m} (\mathbf{y} - \boldsymbol{\mu}_r) \cdot^T (\mathbf{y} - \boldsymbol{\mu}_r) \cdot p(\mathbf{y} | \omega_r) \cdot d\mathbf{y}$$

kde

$$\boldsymbol{\mu}_r = \int_{\mathcal{Y}^m} \mathbf{y} \cdot p(\mathbf{y} | \omega_r) \cdot d\mathbf{y}$$

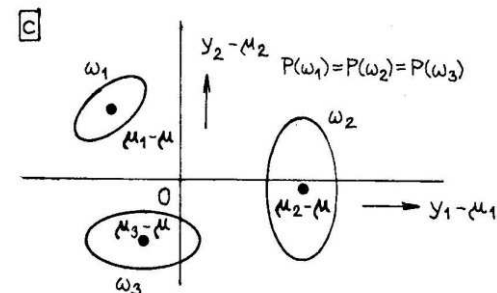
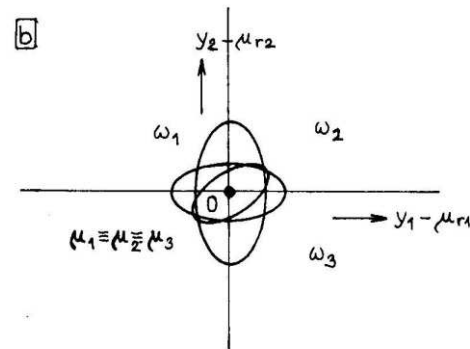
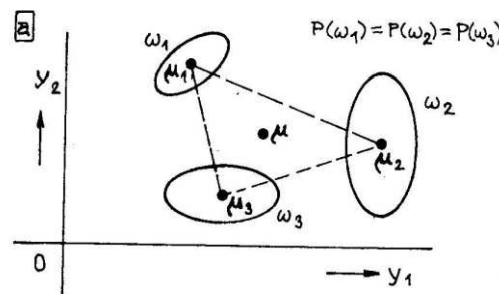
nebo vztahem

$$\begin{aligned} D^0(\mathbf{y}) &= \sum_{r=1}^R P(\omega_r) \cdot \int_{\mathcal{Y}^m} (\mathbf{y} - \boldsymbol{\mu}) \cdot^T (\mathbf{y} - \boldsymbol{\mu}) \cdot p(\mathbf{y} | \omega_r) \cdot d\mathbf{y} = \\ &= \int_{\mathcal{Y}^m} (\mathbf{y} - \boldsymbol{\mu}) \cdot^T (\mathbf{y} - \boldsymbol{\mu}) \cdot p(\mathbf{y}) \cdot d\mathbf{y} \end{aligned}$$

ROZDĚLENÍ DO TŘÍD

kde střední hodnota $\boldsymbol{\mu}$ je vážený průměr středních hodnot všech tříd, tj.

$$\boldsymbol{\mu} = \sum_{r=1}^R P(\omega_r) \cdot \int_{\mathcal{Y}^m} \mathbf{y} \cdot p(\mathbf{y} | \omega_r) \cdot d\mathbf{y} = \int_{\mathcal{Y}^m} \mathbf{y} \cdot p(\mathbf{y}) \cdot d\mathbf{y}$$



Příprava nových učebních materiálů
oboru Matematická biologie

je podporována projektem ESF

č. CZ.1.07/2.2.00/07.0318

„VÍCEBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE“



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ