Analýza a klasifikace dat – přednáška 8

RNDr. Eva Janoušová

Podzim 2014

Volba a výběr proměnných – úvod

- kolik a jaké proměnné (příznaky)?
  - málo příznaků – možná chyba klasifikace;
  - moc příznaků – možná nepřiměřená pracnost, vysoké náklady;

⇓

KOMPROMIS

(potřebujeme kritérium)

- přípustná míra spolehlivosti klasifikace (např. pravděpodobnost chybné klasifikace, odchylka obrazu vytvořeného z vybraných příznaků vůči určitému referenčnímu)
- určit ty příznakové proměnné, jejichž hodnoty nesou nejvíce informace z hlediska řešené úlohy, tj. ty proměnné, kterou jsou nejefektivnější pro vytvoření co nejoddělenějších klasifikačních tříd

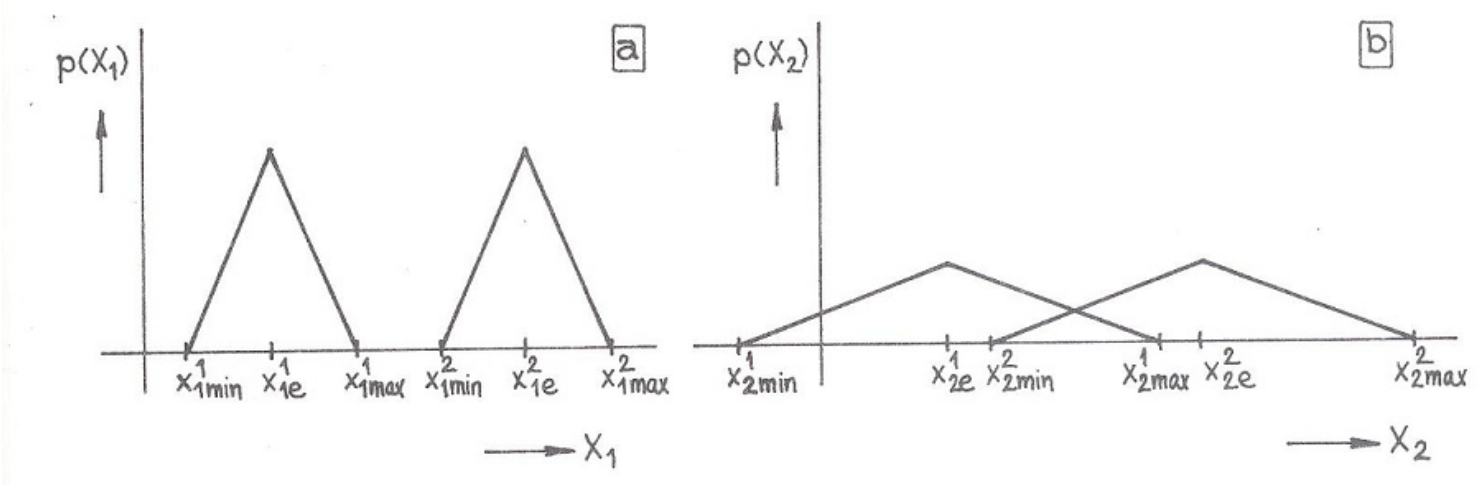
Volba a výběr proměnných

- algoritmus pro určení příznakových veličin nesoucích nejvíce informace pro klasifikátor není dosud teoreticky formalizován - pouze dílčí suboptimální řešení spočívající:
  - ve výběru nezbytného množství veličin z předem zvolené množiny – **selekce**
  - vyjádření původních veličin pomocí menšího počtu skrytých nezávislých veličin, které zpravidla nelze přímo měřit, ale mohou nebo také nemusí mít určitou věcnou interpretaci – **extrakce**

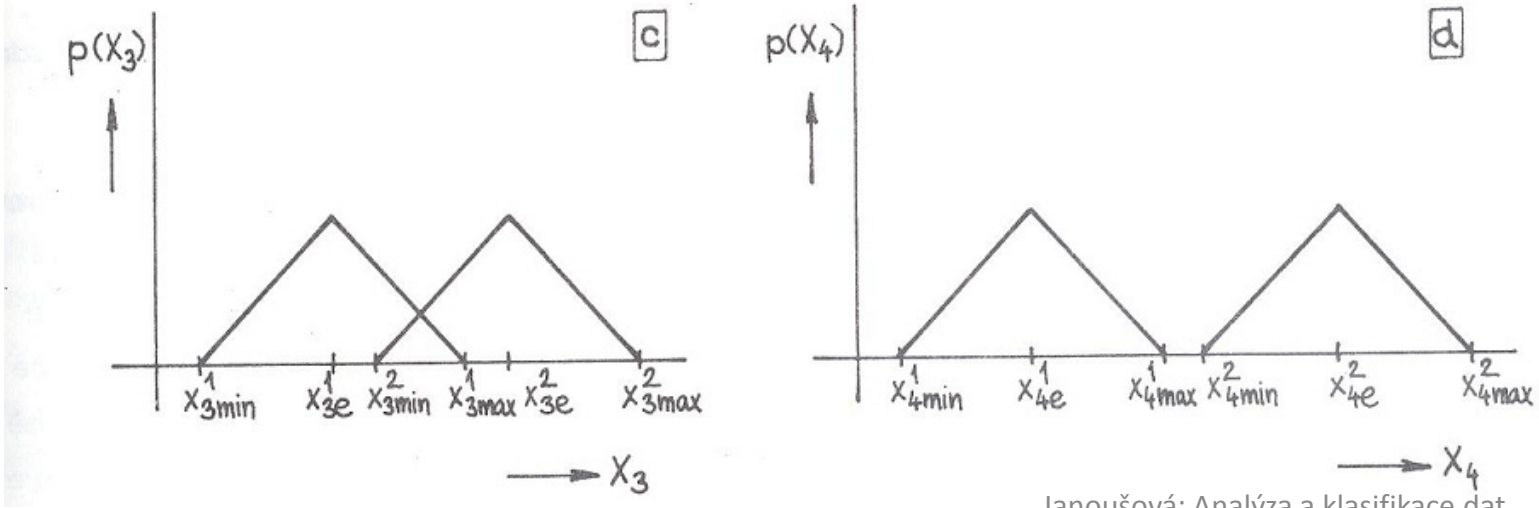
- počáteční volba příznakových veličin je z velké části empirická, vychází ze zkušeností získaných při empirické klasifikaci člověkem a závisí kromě rozboru podstaty problému i na technických (ekonomických) možnostech a schopnostech hodnoty veličin určit

Zásady pro volbu příznaků I

- výběr veličin s minimálním rozptylem uvnitř tříd



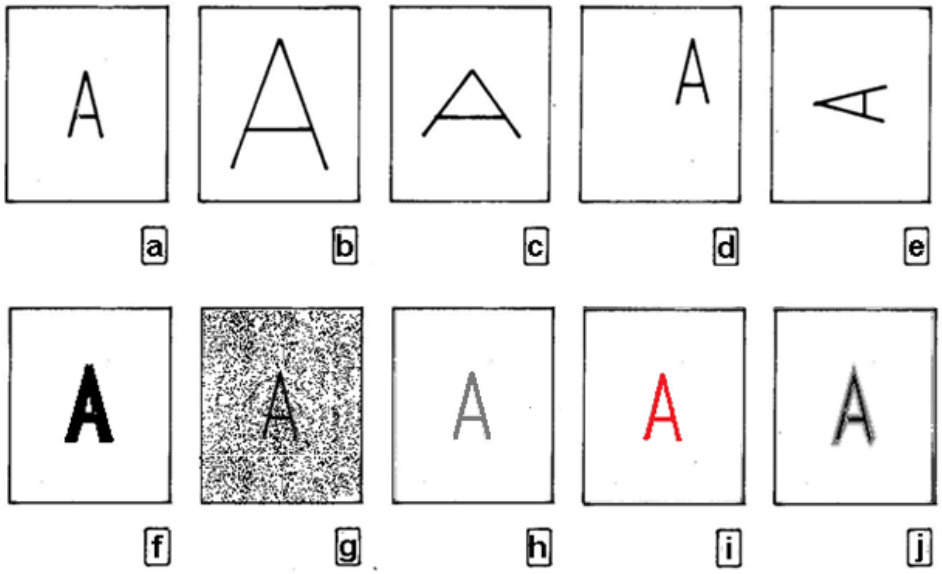
- výběr veličin s maximální vzdáleností mezi třídami



Zásady pro volbu příznaků II

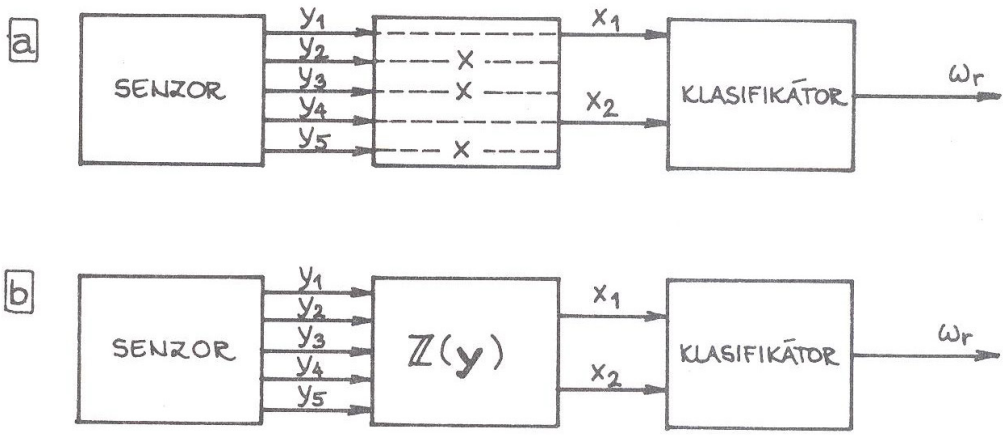
- výběr vzájemně nekorelovaných veličin
  - pokud jsou hodnoty jedné příznakové veličiny závislé na příznacích druhé veličiny, pak použití obou těchto veličin nepřináší žádnou další informaci pro správnou klasifikaci – stačí jedna z nich, jedno která

- výběr veličin invariantních vůči deformacím
  - volba elementů formálního popisu závisí na vlastnostech původních i předzpracovaných dat a může ovlivňovat způsob předzpracování



Výběr příznaků

- formální popis objektu původně reprezentovaný *m* rozměrným vektorem se snažíme vyjádřit vektorem *n* rozměrným tak, aby množství diskriminační informace obsažené v původním vektoru bylo v co největší míře zachováno
- dva principiálně různé způsoby:
  1. **selekce** – nalezení a odstranění těch příznakových funkcí, které přispívají k separabilitě klasifikačních tříd nejméně
  2. **extrakce** – transformace původních příznakových proměnných na menší počet jiných příznakových proměnných

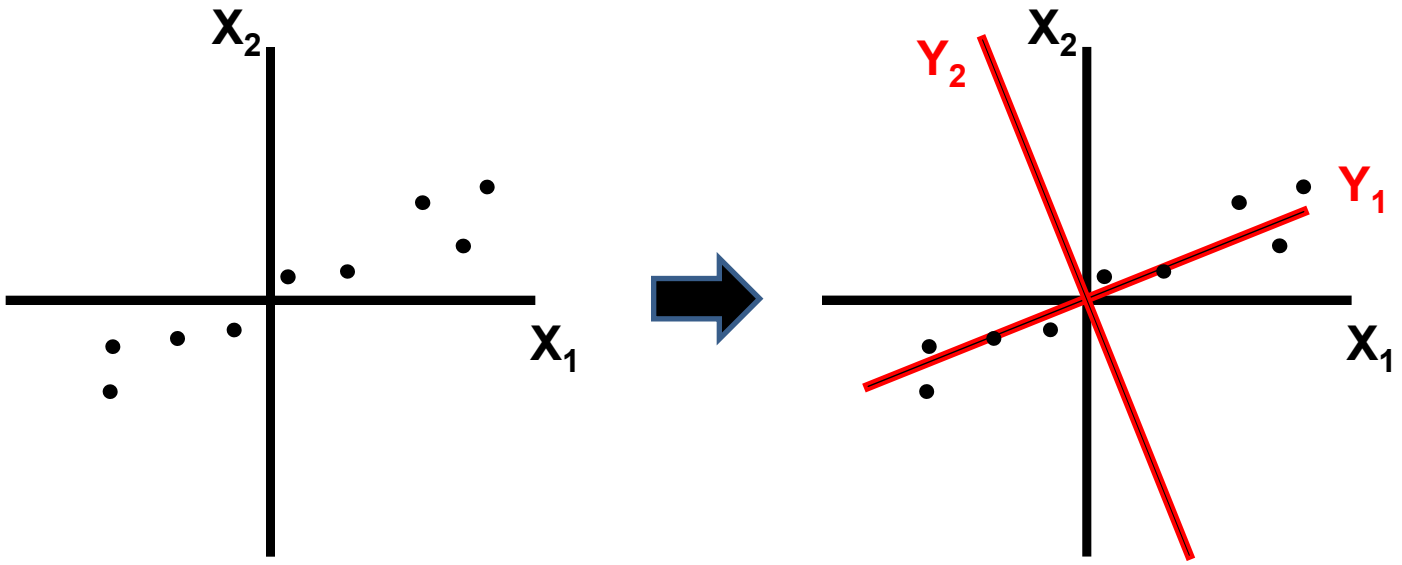


Extrakce příznaků

- jedním z principů výběru příznaků
- transformace původních příznakových proměnných na menší počet jiných příznakových proměnných $\Rightarrow$ tzn. hledání (optimálního) zobrazení Z, které transformuje původní *m*-rozměrný prostor (obraz) na prostor (obraz) *n*-rozměrný ($m \geq n$)
- pro snadnější řešitelnost hledáme zobrazení Z v oboru lineárních zobrazení
- 3 kritéria pro nalezení optimálního zobrazení Z:
  - **obrazy v novém prostoru budou aproximovat původní obrazy ve smyslu minimální střední kvadratické odchylky**
  - rozložení pravděpodobnosti veličin v novém prostoru budou splňovat podmínky kladené na jejich pravděpodobnostní charakteristiky
  - obrazy v novém prostoru budou minimalizovat odhad pravděpodobnosti chyby

Analýza hlavních komponent – opakování

- anglicky Principal component analysis (PCA)
- snaha redukovat počet proměnných nalezením nových latentních proměnných (hlavních komponent) vysvětlujících co nejvíce variability původních proměnných
- nové proměnné (Y_1, Y_2) lineární kombinací původních proměnných (X_1, X_2)



PCA – opakování

- vstup do PCA:
  - kovarianční matice
  - matice korelačních koeficientů
- hlavní komponenty odpovídají vlastním vektorům kovarianční matice (či matice korelačních koef.)
- variabilita vysvětlená příslušnou komponentou odpovídá vlastním číslům
- vlastní vektory seřazeny podle vlastních hodnot (sestupně) $\Rightarrow$ vybráno prvních *n* komponent vyčerpávajících nejvíce variability původních dat
- předpoklady: kvantitativní proměnné s normálním rozdělením

PCA – obecněji

- dáno *K* obrazů charakterizovaných *m* příznakovými proměnnými (nerozdělenými do klasifikačních tříd)

příznaky

obrazy	$\mathbf{p}_1$ $\mathbf{p}_2$ … $\mathbf{p}_m$
$\mathbf{x}_1$	
$\mathbf{x}_2$	
…	
$\mathbf{x}_K$	

- aproximujme nyní kterýkoliv obraz $\mathbf{x}_k$ lineární kombinací *n* ortonormálních vektorů $\mathbf{e}_i$ ($n \leq m$)

$$\mathbf{y}_k = \sum_{i=1}^{n} c_{ki} e_i$$

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

$$c_{ki} = \mathbf{x}_k^T e_i$$

PCA – kritérium minimální střední kvadratické odchylky

- nalezení optimálního zobrazení pomocí **kritéria minimální střední kvadratické odchylky**:

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

- vztah lze pomocí dříve uvedených vztahů upravit na:

$$\varepsilon_k^2 = \left\| \mathbf{x}_k \right\|^2 - \sum_{i=1}^{n} c_{ki}^2$$

- střední kvadratická odchylka pro všechny obrazy $\mathbf{x}_k$, k=1,…,K je

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

PCA – kritérium minimální střední kvadratické odchylky

- musíme zvolit bázový systém $\mathbf{e}_i$ tak, aby střední kvadratická odchylka ε^2 byla minimální

- diskrétní konečný rozvoj podle vztahu $\mathbf{y}_k = \sum_{i=1}^{n} c_{ki} e_i$
  s bázovým systémem $\mathbf{e}_i$, optimálním podle kritéria minimální střední kvadratické chyby, nazýváme diskrétní Karhunenův – Loevův rozvoj

- střední kvadratická odchylka

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

  je minimální, když je maximální výraz

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

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

PCA – kritérium minimální střední kvadratické odchylky

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

  $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_m \geq 0$

  a podle toho očíslujeme i odpovídající vlastní vektory, lze dokázat, výše uvedený výraz dosahuje maxima, jestliže platí

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

  a pro velikost maxima je

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

- pak pro minimální střední kvadratickou platí

$$\varepsilon_{\min}^2 = \frac{1}{K} \sum_{k=1}^{K} \|\mathbf{x}_k\|^2 - \sum_{i=1}^{n} \lambda_i = tr(\kappa(\mathbf{x})) - \sum_{i=1}^{n} \lambda_i = \sum_{i=n+1}^{m} \lambda_i$$

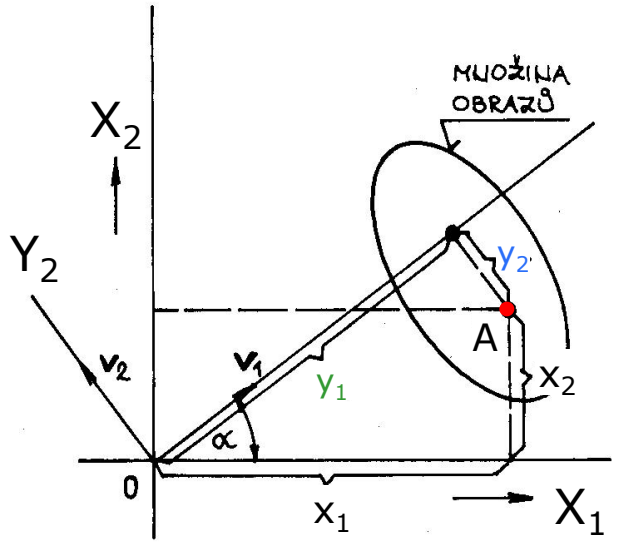
PCA – vstupní matice

- **autokorelační matice** – data nejsou nijak upravena (zohledňována průměrná hodnota i rozptyl původních dat)
- **kovarianční (disperzní) matice** – data centrována (od každé příznakové proměnné odečtena její střední hodnota) – zohledňován rozptyl původních dat
- **matice korelačních koeficientů** – data standardizována (odečtení středních hodnot a podělení směrodatnými odchylkami) – použití pokud mají proměnné různá měřítka
- **každou úpravou původních dat ale přicházíme o určitou informaci !!!**

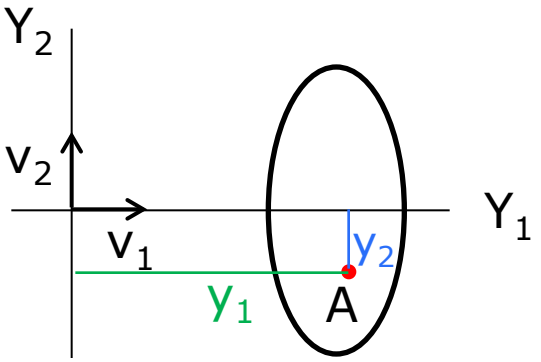
PCA – vlastnosti Karhunenova-Loevova rozvoje

- 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

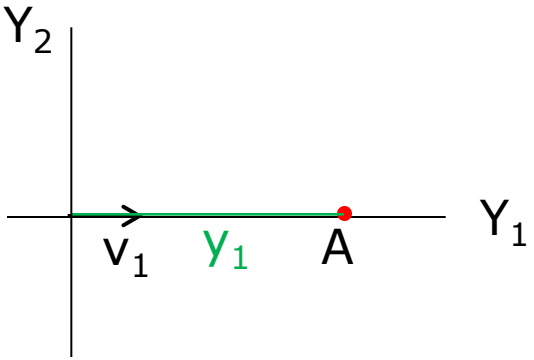
PCA – geometrická interpretace



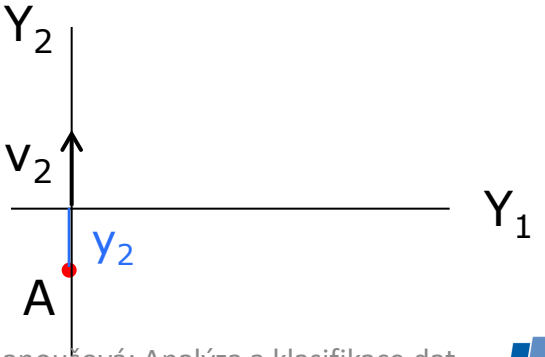
použití obou hlavních komponent



použití 1. hlavní komponenty



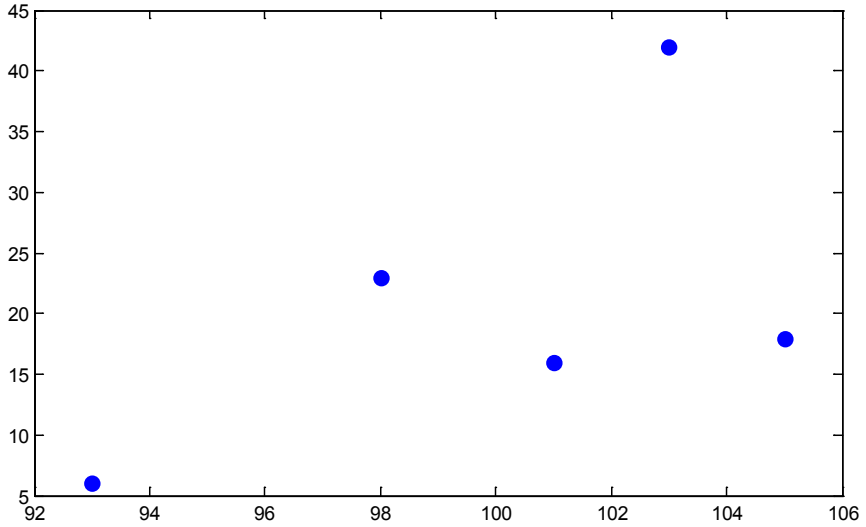
použití 2. hlavní komponenty



PCA – příklad

- data:

A	101	16
B	105	18
C	103	42
D	98	23
E	93	6



PCA – příklad

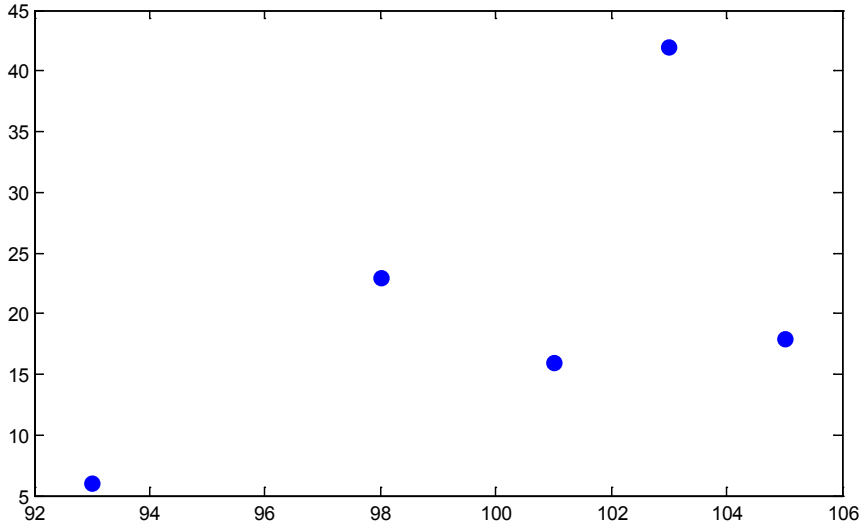
- data:

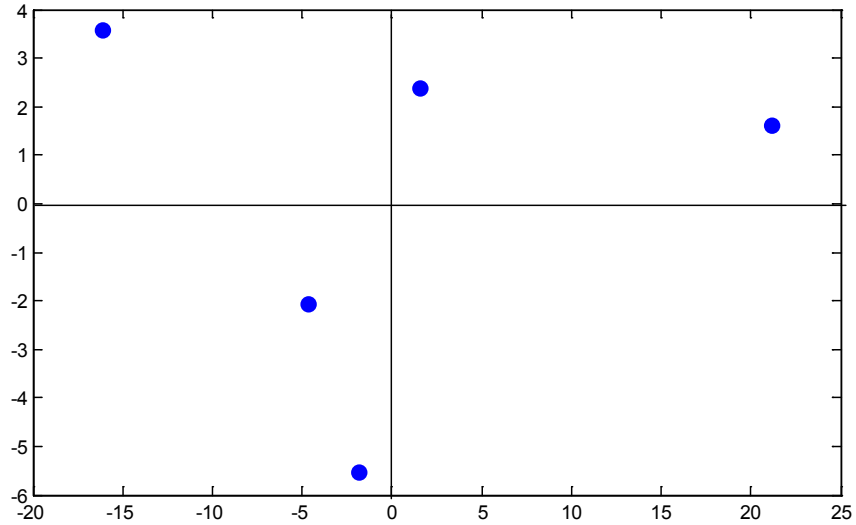
A	101	16
B	105	18
C	103	42
D	98	23
E	93	6

PCA – příklad

- data:

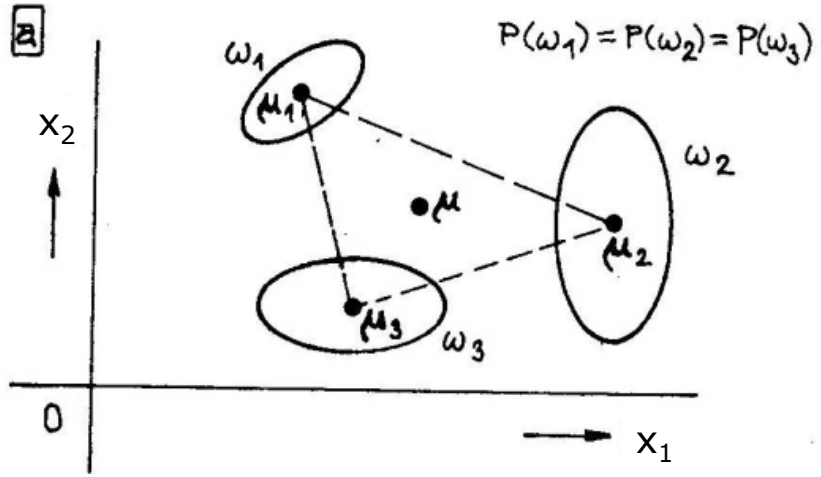
A	101	16
B	105	18
C	103	42
D	98	23
E	93	6





PCA – rozdělení do tříd

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



- V tom případě autokorelační matice bude

$$\kappa(\mathbf{x}) = \sum_{r=1}^{R} P(\omega_r).\int_{\Upsilon^m} x.x^T.p(\mathbf{x}\,|\,\omega_r).d\mathbf{x} = \int_{\Upsilon^m} x.^T x.p(\mathbf{x}).d\mathbf{x}$$

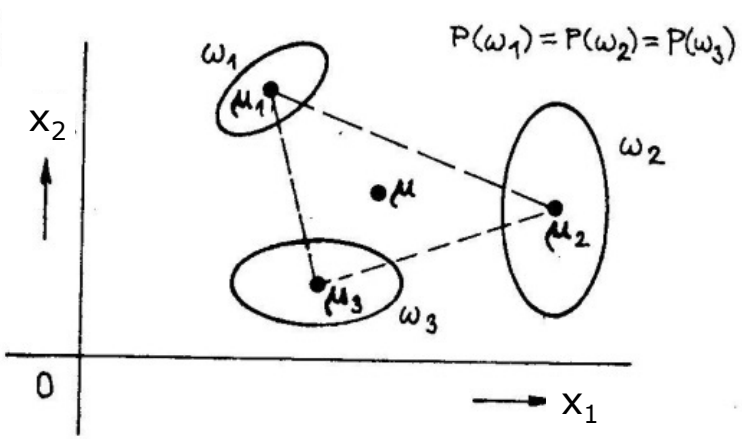
PCA – rozdělení do tříd

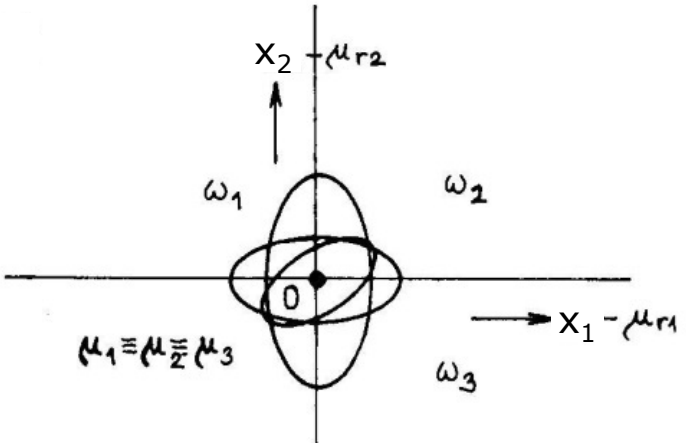
- disperzní matice – vztah 1:

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

- kde $\boldsymbol{\mu}_r = \int_{\mathcal{Y}^m} \mathbf{x}.p(\mathbf{x}\,|\,\omega_r)d\mathbf{x}$

- rozlišení klasifikačních tříd jen podle disperze
- transformované příznak. proměnné nekorelované





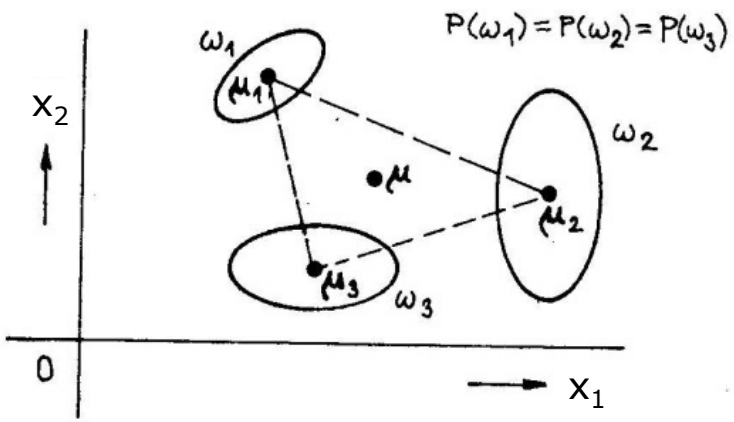
PCA – rozdělení do tříd

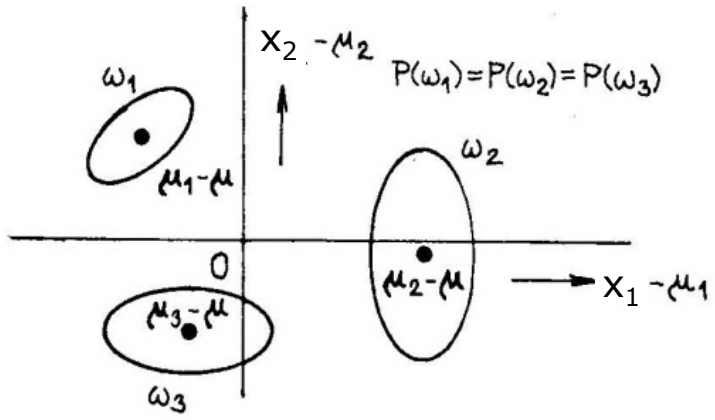
- disperzní matice – vztah 2:

$$D^0(\mathbf{x}) = \sum_{r=1}^{R} P(\omega_r).\int_{\Upsilon^m} (\mathbf{x}-\boldsymbol{\mu}).(\mathbf{x}-\boldsymbol{\mu})^T.p(\mathbf{x}\,|\,\omega_r).d\mathbf{x} = \int_{\Upsilon^m} (\mathbf{x}-\boldsymbol{\mu}).(\mathbf{x}-\boldsymbol{\mu})^T.p(\mathbf{x}).d\mathbf{x}$$

kde $\boldsymbol{\mu} = \sum_{r=1}^{R} P(\omega_r).\int_{\Upsilon^m} \mathbf{x}.p(\mathbf{x}\,|\,\omega_r)d\mathbf{y} = \int_{\Upsilon^m} \mathbf{x}.p(\mathbf{x})d\mathbf{x}$

- neodstraňuje vliv středních hodnot obrazů v jednotlivých třídách – použití pokud jsou stř. h. výrazně odlišné a nesou velké množství informace





PCA – rozšiřující poznatky

- výpočet PCA, když je m >> K
- souvislost se singulárním rozkladem (SVD – Singular Value Decomposition)

Příprava nových učebních materiálů pro obor Matematická biologie

je podporována projektem OPVK

č. CZ.1.07/2.2.00/28.0043

„Interdisciplinární rozvoj studijního oboru Matematická biologie“

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ