

Klasifikace, rozpoznávání a shlukování

18. 3. – 25. 3. 2015

4. Klasifikace, rozpoznávání a shlukování

Motivace

Aplikace

4. Klasifikace, rozpoznávání a shlukování

Klasifikace, rozpoznávání a shlukování

- předmětů
- jevů
- situací
- . . .

Obecně hovoříme o klasifikaci, rozpoznávání a shlukování **objektů**.

Reprezentace objektů



prostřednictvím signálů

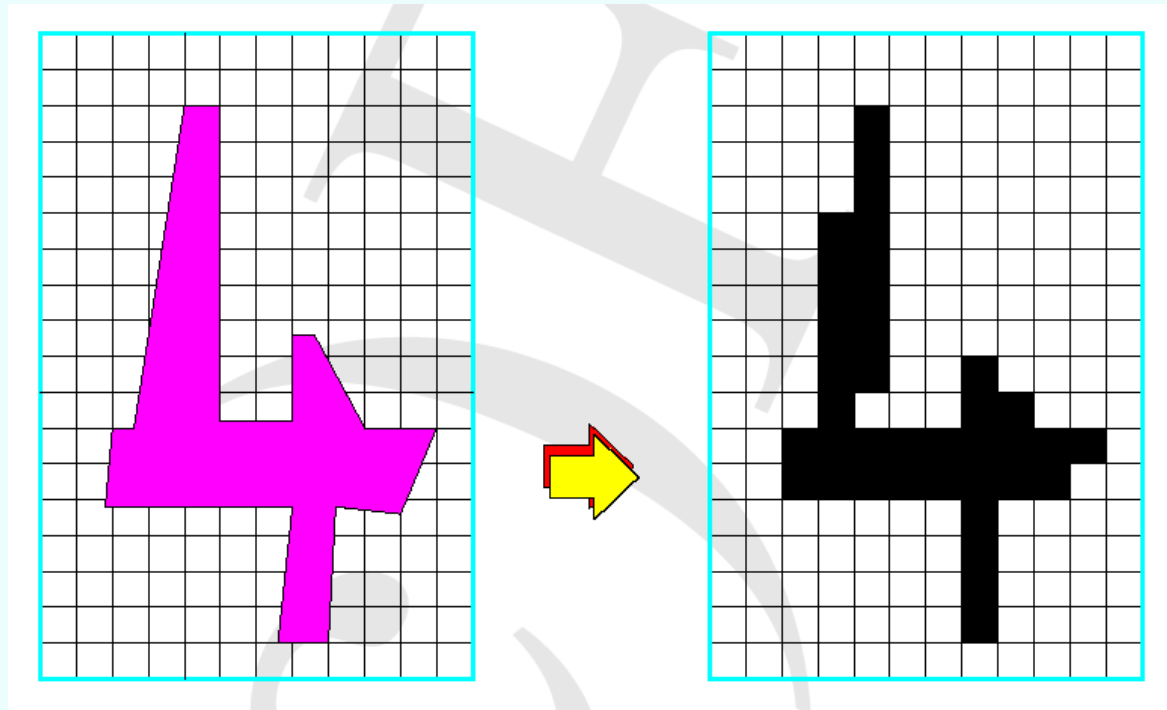
- jednorozměrných (zvuky, hudba, řeč, ...)
- dvourozměrných (snímky scén, ...)
- vícerozměrných (obecné)

Podle charakteru reprezentace objektů signály rozlišujeme objekty (a jejich rozpoznávání)

- vizuální (reálné, upravené, symbolické, stylizované aj.)
- akustické (zvuky, hluky, hudba, řeč, ..., vždy jednorozm.)
- taktilní (obecně reprezentované n -rozměrnými signály)

4. Klasifikace, rozpoznávání a shlukování

Příklad reprezentace číslice '4' v pravoúhlém rastru:



Reprezentace tzv. jasovou funkcí:

$$f(x, y) = \left\{ \begin{array}{l} 0, \text{ reprezentuje-li obrazový bod bílou barvu podkladu} \\ 1, \text{ reprezentuje-li obrazový bod černou barvu písma} \end{array} \right\}$$

Rozpoznávání, klasifikace a shlukování

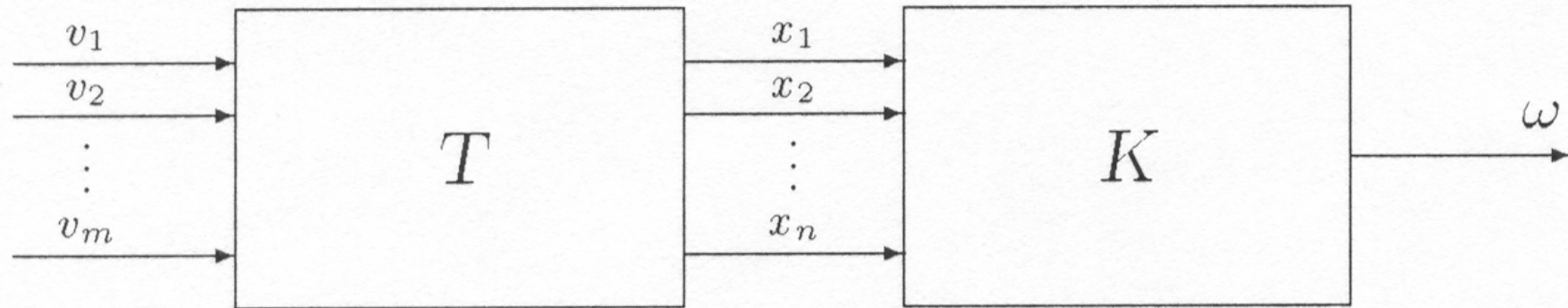
Rozpoznávání chápeme jako úlohu, při které objekty zařazujeme do tříd podle jejich společných vlastností tak, že objekty vzájemně si podobné zařazujeme do stejné třídy.

Rozlišujeme:

- **klasifikaci** – zařazujeme do předem známého, pevného počtu tříd (například rozpoznávání znaků)
- **rozpoznávání** – počet tříd není předem znám a třídy identifikujeme až během vlastního rozpoznávání (například rozpoznávání plynulé řeči)
- **shlukování (klastrování)** – zařazujeme objekty do tříd na základě jejich podobnosti, příslušné třídy nejsou známy

4. Klasifikace, rozpoznávání a shlukování

Obecná klasifikační úloha



T . . . transformace vstupních charakteristik – vytvoření obrazu

K . . . klasifikátor

\mathbf{v} . . . vektor vstupních charakteristik

\mathbf{x} . . . obraz (symbolický popis) objektu

ω . . . indikátor třídy

4. Klasifikace, rozpoznávání a shlukování

Rozhodovací pravidlo

Rozhodovací pravidlo, podle kterého klasifikátor přiřazuje obraz do klasifikační třídy, můžeme obecně definovat jako skalární funkci vektorového argumentu

$$\omega = d(\mathbf{x}) .$$

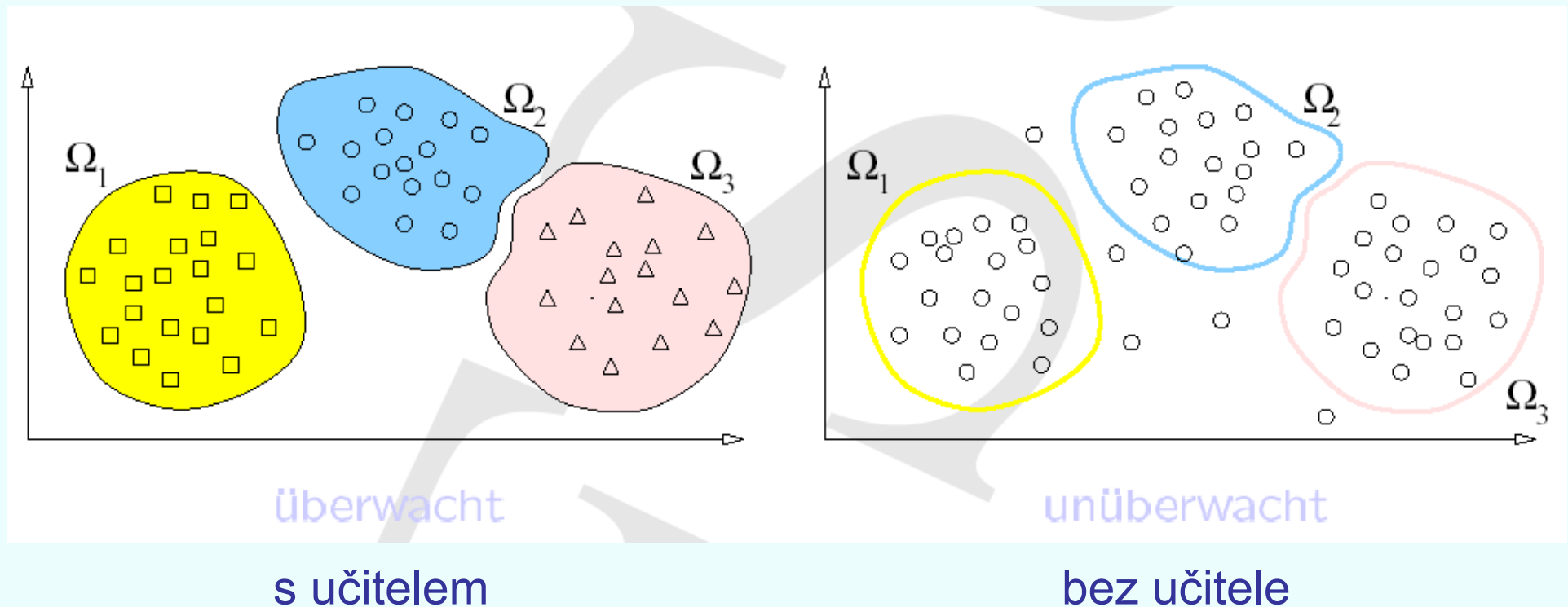
Přesnější vyjádření rozhodovacího pravidla, které zohledňuje i tzv. nastavení klasifikátoru \mathbf{q} , je

$$\omega = d(\mathbf{x}, \mathbf{q}) .$$

Nastavení klasifikátoru se provádí **trénováním** neboli **učením**. Rozlišujeme **učení s učitelem**, kdy klasifikátoru předkládáme obrazy, u nichž známe jejich příslušnost k třídě, a **učení bez učitele**, kdy správné zařazení do klasifikačních tříd neznáme.

4. Klasifikace, rozpoznávání a shlukování

Učení (trénování) klasifikátoru



Rozpoznávání a klasifikace objektů

➡ na základě vlastností objektů

Vlastnosti objektů

- měřitelné a kvantifikovatelné
- strukturní

Podle typu vlastností objektů hovoříme o

- vytváření příznakového nebo strukturního popisu objektů
- příznakových nebo strukturních metodách rozpoznávání

4. Klasifikace, rozpoznávání a shlukování

Příznakový nebo strukturní popis rozpoznávaných (klasifikovaných) objektů – obecně hovoříme o vytváření symbolických popisů objektů – nazýváme popisem objektů jejich **obrazy**.

Tedy: **Obraz objektu = symbolický popis objektu**

Stejně jako metody rozlišujeme i obrazy (symbolické popisy objektů)

- tvořené příznaky (vektory příznaků)
- strukturní

Obsahují-li obrazy objektů n složek, pak všechny rozpoznávané (klasifikované) obrazy tvoří **obrazový prostor** úlohy.

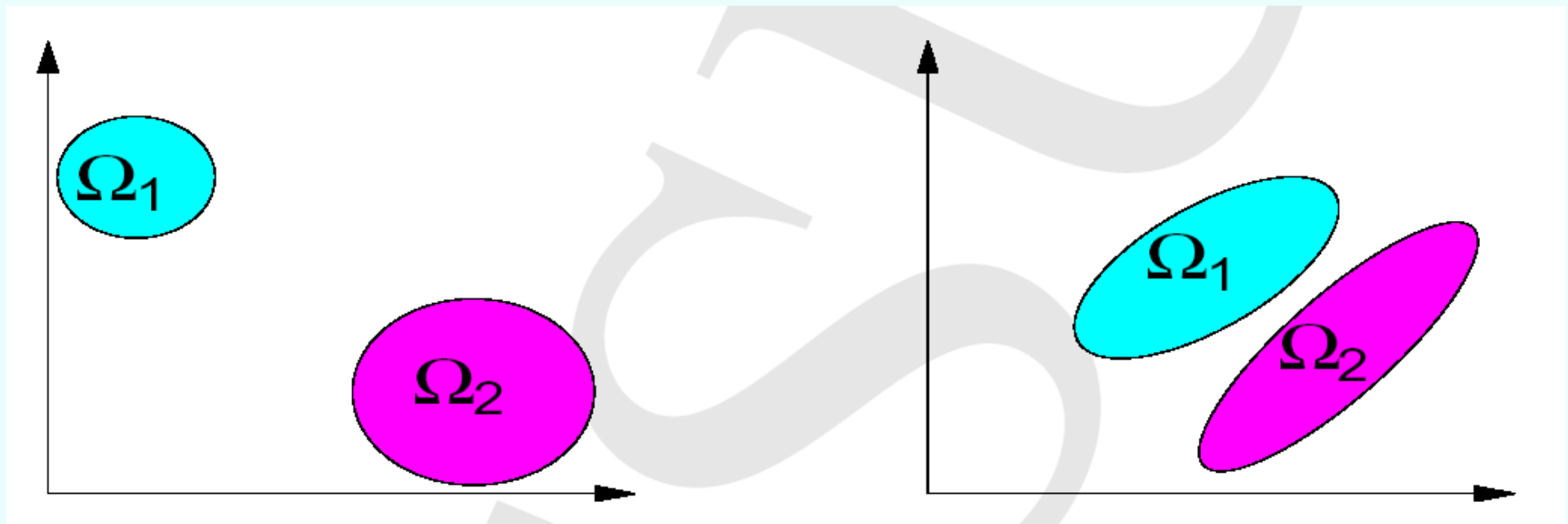
Jednotlivé klasifikační třídy objektů získáme **rozkladem** obrazového prostoru úlohy na R klasifikačních tříd.

4. Klasifikace, rozpoznávání a shlukování

Příklady rozkladu obrazového prostoru na třídy

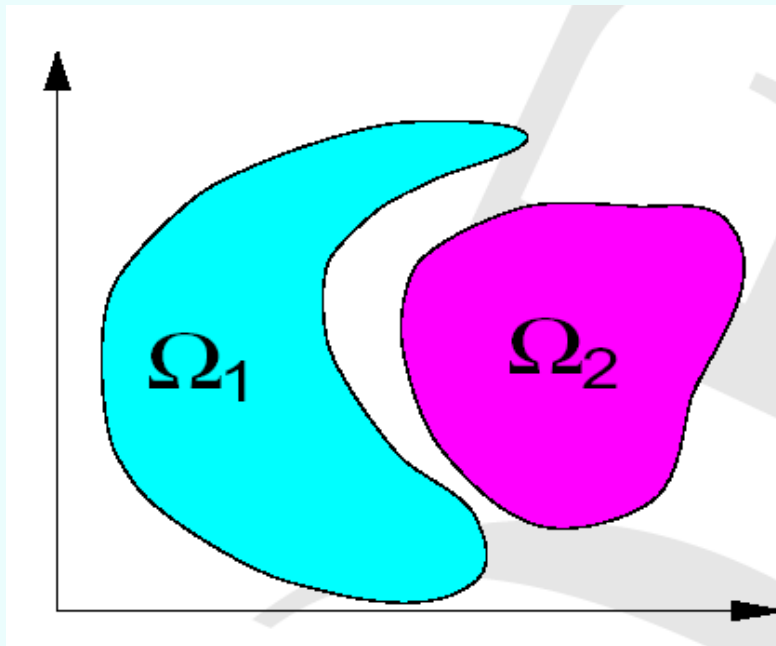
a) třídy disjunktní, s velkou mezi-
třídní vzdáleností – ideální případ

b) třídy disjunktní, kompaktní,
geometricky „blízke“

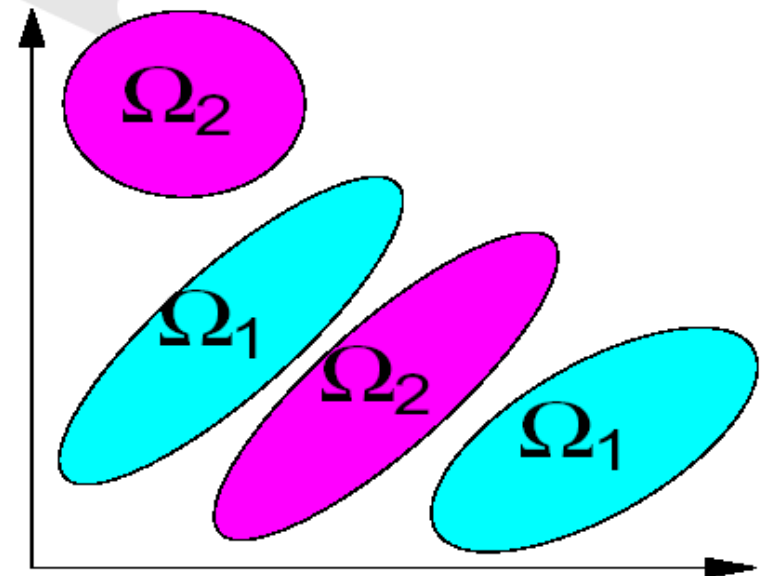


4. Klasifikace, rozpoznávání a shlukování

c) třídy disjunktní, kompaktní,
s nelineární oddělovací nad-
plochou

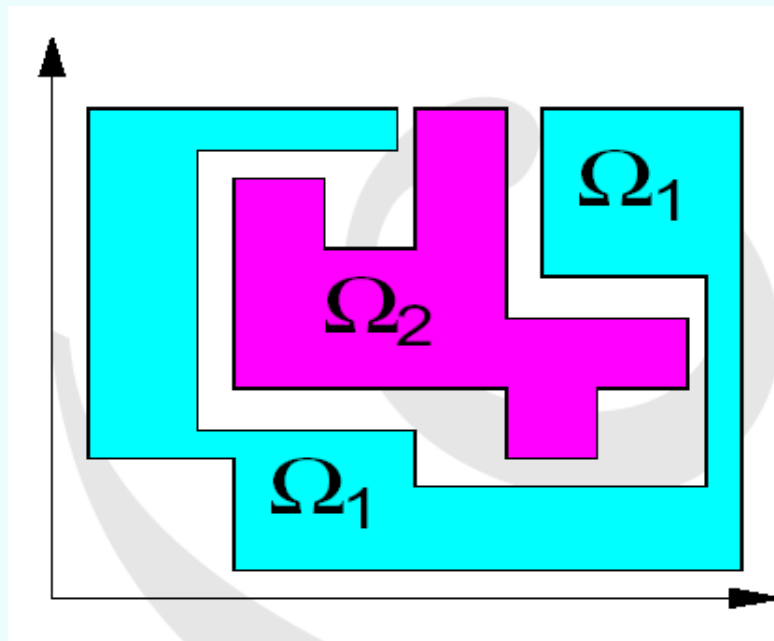


d) třídy disjunktní, nekompaktní,
avšak lineárně oddělitelné

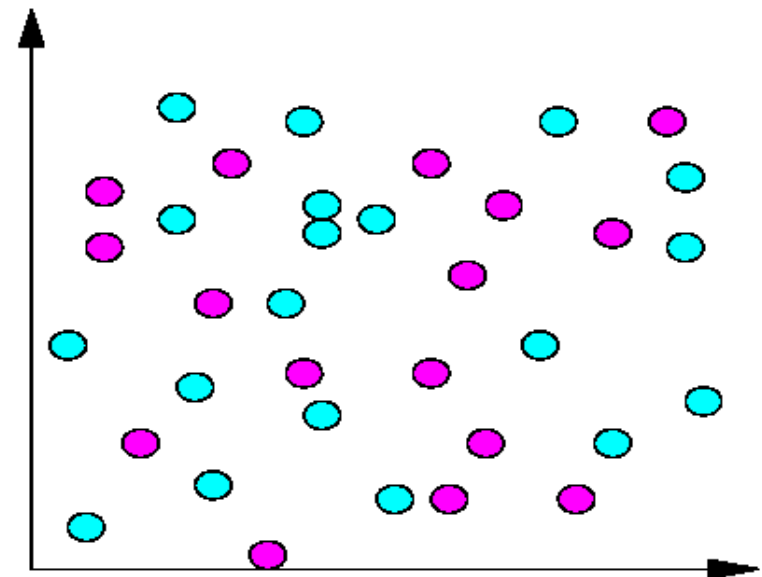


4. Klasifikace, rozpoznávání a shlukování

e) třídy disjunktní, kompaktní, vnořené, komplikovaně oddělitelné (s rozdělovací nadplochou složitého tvaru)

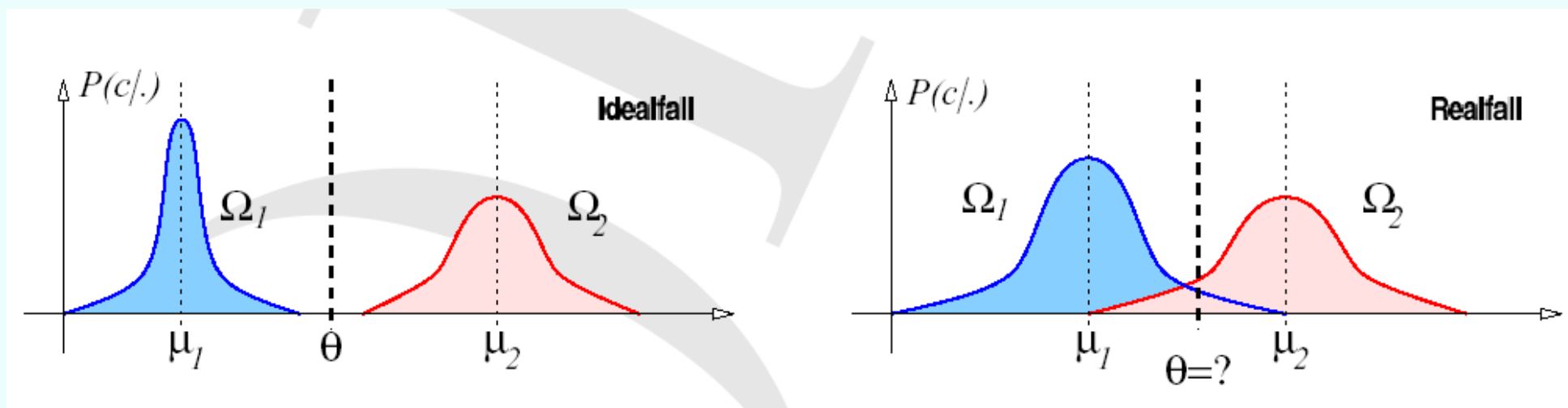


f) nekompaktní, prolínající se třídy s velmi obtížně určitelnou rozdělovací nadplochou



4. Klasifikace, rozpoznávání a shlukování

Pokud vezmeme v úvahu pravděpodobnostní rozložení obrazů uvnitř jednotlivých tříd:



ideální případ – klasifikační třídy se nepřekrývají – jsou disjunktní – o zařazení obrazu do třídy ω_1 nebo ω_2 lze rozhodnout s jistotou, čili s $P = 1$

reálný případ – klasifikační třídy se překrývají – v oblasti okolo rozdělovací nadplochy lze o zařazení obrazu do příslušné třídy jen s pravděpodobností $P < 1$

4. Klasifikace, rozpoznávání a shlukování

Shlukování

= sdružování objektů do tříd na základě podobnosti

Shluk

= skupina objektů, které si jsou **co nejvíce** podobné uvnitř shluku a **co nejméně** podobné mezi shluky

Metody

- rozdělovací metody
- hierarchické metody
- metody založené na hustotě prvků
- mřížkové metody
- další metody (založené např. na neuronových sítích)

4. Klasifikace, rozpoznávání a shlukování

Algoritmus k-means

— rozdělovací metoda

Princip: dělení prvků podle kritéria, podobnostní funkce (vzdálenost)

Shluk je reprezentován těžištěm objektů

Popis algoritmu

1. zadání počtu shluků **k** a množiny všech objektů
2. volba **k** výchozích „středů“ shluků
3. přiřazení všech objektů ke „středu“ shluku s **nejmenší vzdáleností** na základě podobnostní funkce (vzdálenosti)
4. výpočet nového „středu“ u každého shluku = „těžiště“ množiny objektů
5. návrat ke kroku 2

Konec = nezměněna pozice žádného objektu

4. Klasifikace, rozpoznávání a shlukování

Základní typy dat

Intervalové proměnné

Př: pozice, váha, výška, intenzita (jas), datum

Nutnost **normalizace** — obvykle interval $[0; 1]$

— snaha nastavit všem proměnným stejnou váhu

Binární proměnné

Pouze dva stavy: **1** — přítomnost vlastnosti

0 — nepřítomnost vlastnosti

Př: objekt složený z primitiv (čtverec, trojúhelník a kružnice)

Objekt č. 1 — 1 0 0

č. 2 — 1 0 1

Nejznámější metriky pro měření vzdálenosti

- Eukleidovská vzdálenost
- Manhattan vzdálenost
- Cosinová vzdálenost

4. Klasifikace, rozpoznávání a shlukování

Evaluační metriky

- vyhodnocení přesnosti klasifikace (rozpoznávání)
- udává se zpravidla v % nebo jako desetinné číslo

Dělení:

- | | |
|--------------------------|---------------------------|
| <i>1–1 (klasifikace)</i> | <i>1–n (rozpoznávání)</i> |
| • chybovost (Error Rate) | accuracy |
| • accuracy | přesnost, úplnost, F-míra |

Konfúzní matice

4. Klasifikace, rozpoznávání a shlukování

Chybovost (Error Rate)

$$ER = E / ALL$$

E = počet chyb

ALL = počet všech klasifikovaných vzorků

Přesnost (accuracy, klasifikace 1–1)

$$ACC = OK / ALL = 1 - ERR$$

OK = počet správně klasifikovaných vzorků

Př:

4. Klasifikace, rozpoznávání a shlukování

Přesnost (accuracy, rozpoznávání 1–n)

$$\text{ACC} = (N - D - S - I) / N$$

N = počet všech rozpoznávaných jednotek v sekvenci

D = počet vynechaných jednotek

S = počet zaměněných jednotek

I = počet vložených jednotek

Př:

4. Klasifikace, rozpoznávání a shlukování

Přesnost, úplnost, F-míra

odhad / reál	1	0
1	TP	FP
0	TN	FN

TP = správně odhadnuté pozitivní příklady

TN = správně odhadnuté negativní příklady

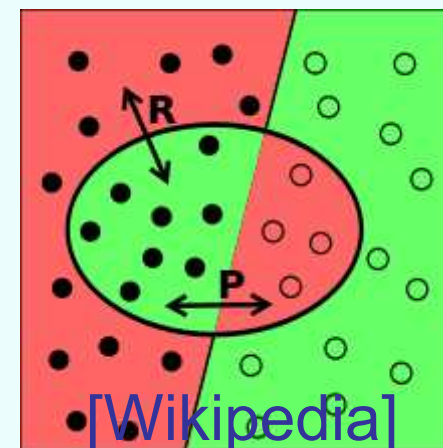
FP = špatně odhadnuté negativní příklady

FN = špatně odhadnuté pozitivní příklady

$Pr = TP / (TP + FP)$... přesnost

$Rec = TP / (TP + FN)$... úplnost

$F\text{-mes} = 2.Pr.Rec / (Pr + Rec)$



Př:

4. Klasifikace, rozpoznávání a shlukování

Konfúzní matice (matice záměn)

= matice, která obsahuje hodnoty přesnosti klasifikátoru s ohledem na všechny klasifikované prvky

Př:

	a	b	c
a	88	4	12
b	8	65	4
c	11	15	33

4. Klasifikace, rozpoznávání a shlukování

Praktické příklady

- klasifikace dokumentů
- rozpoznávání dialogových aktů