

Pokročilé metody analýzy dat v neurovědách



RNDr. Eva Janoušová
doc. RNDr. Ladislav Dušek, Dr.

Jaro 2015

Blok 7

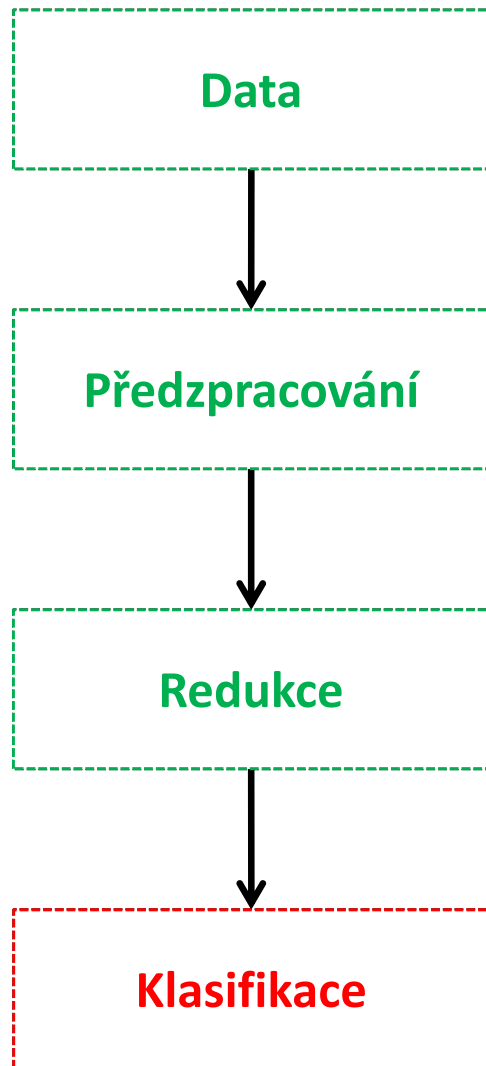
Klasifikace dat I

Osnova

1. Úvod do klasifikace dat
2. Klasifikace pomocí diskriminačních funkcí:
 - lineární diskriminační funkce
 - Bayesův klasifikátor
3. Klasifikace pomocí minimální vzdálenosti
4. Klasifikace pomocí hranic:
 - Fisherova lineární diskriminační analýza

Úvod do klasifikace dat

Schéma analýzy a klasifikace dat



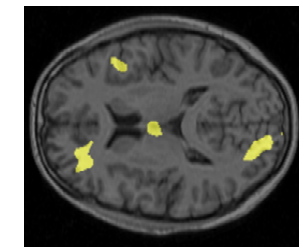
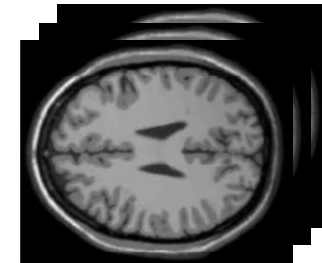
Ukázka - kognitivní data apod.

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M		90
4	3	26	Z	178	70

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

Ukázka - obrazová data



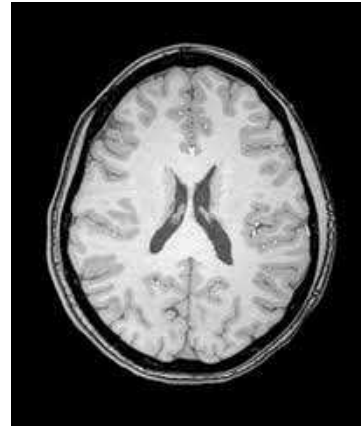
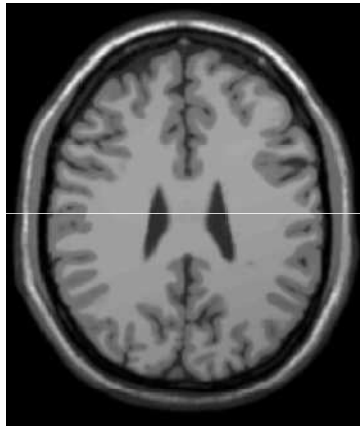
nebo



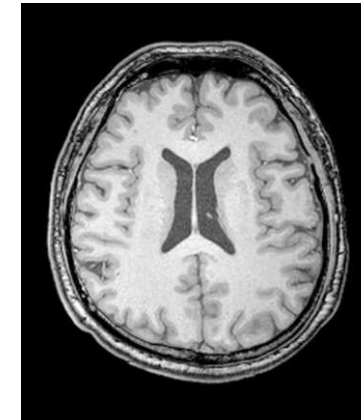
Proč používat klasifikaci dat?

1. Podpora diagnostiky onemocnění mozku (Alzheimerova choroba, schizofrenie atd.):

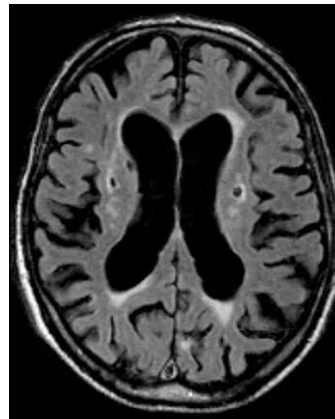
Zdravé subjekty



Nový subjekt



Pacienti

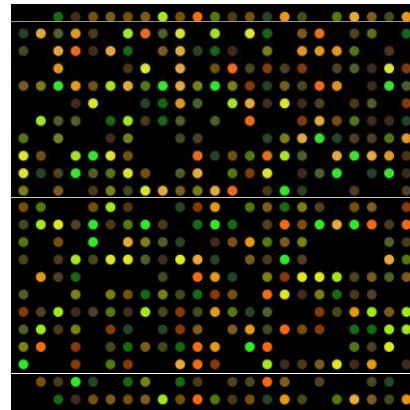
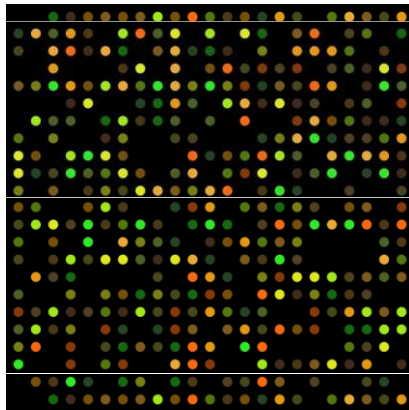


Pacient? x Zdravý?

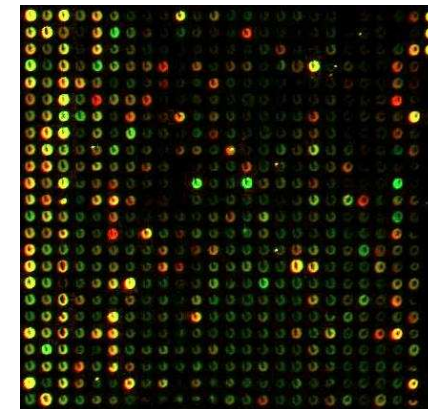
Proč používat klasifikaci dat?

2. Odhalení genetického onemocnění na základě dat s microarray experimentů:

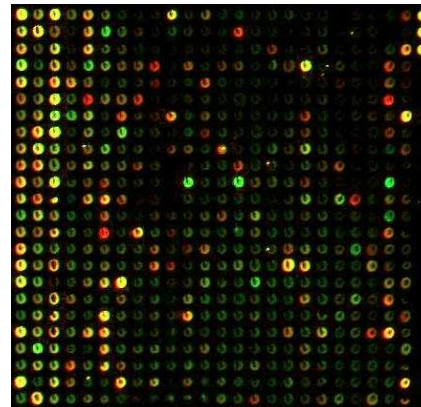
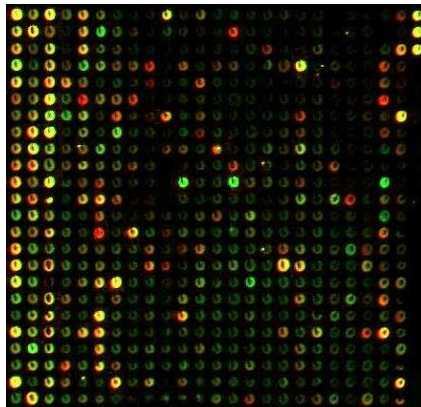
Zdravé subjekty



Nový subjekt



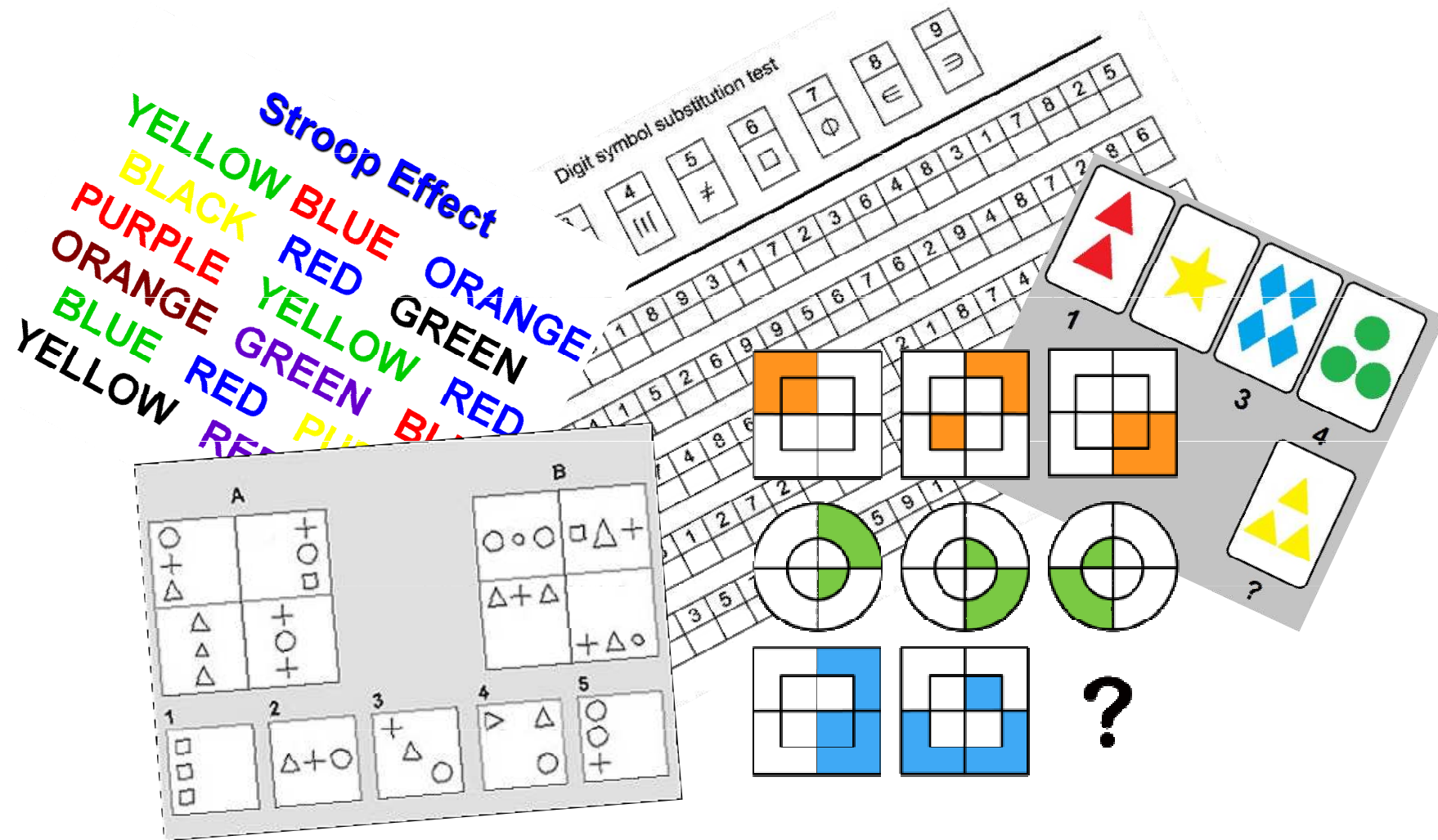
Pacienti



Pacient? x Zdravý?

Proč používat klasifikaci dat?

3. Zjištění demence a dalších onemocnění na základě kognitivních testů:



Demence ano? x Demence ne?

Proč používat klasifikaci dat?

4. Rozpoznání hmyzu:

Nejedovaté housenky



Jedovaté housenky



?



Jedovatá nebo nejedovatá housenka?

Proč používat klasifikaci dat?

5. Rozpoznání vadných výrobků:

Matičky bez vady



Matičky s vnitřní prasklinou



?



Matička bez vady nebo s vnitřní prasklinou?

Proč používat klasifikaci dat?

6. Rozpoznání tváře při vstupu do zabezpečené budovy:

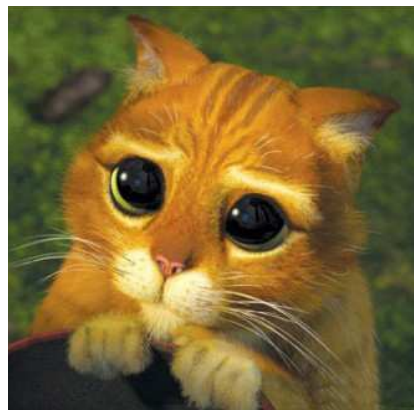
Nemá
přístup do
budovy



?



Má přístup
do budovy



Dostane se do
budovy: ano? x
ne?

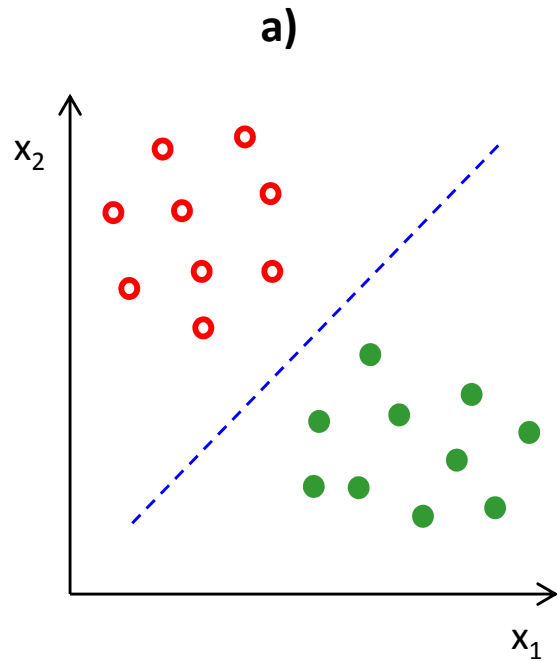
Cíle klasifikace dat - shrnutí

- **rozhodnutí o typu či charakteru objektu** – např. že daný člověk může vstoupit do budovy či nikoliv, že zvíře je medvěd hnědý nebo medvěd lední apod. – **klasifikační**, resp. **rozpoznávací úloha**;
- **posouzení kvality stavu analyzovaného objektu** – např. zda je pacient v pořádku, nebo má infarkt myokardu, cirhózu jater, apod. – opět **klasifikační**, resp. **rozpoznávací úloha**;
- **rozhodnutí o budoucnosti objektu** – např. zda lze pacienta léčit a vyléčit, zda les po 20 letech odumře, jaké bude sociální složení obyvatelstva na daném území a v daném čase – **klasifikační**, resp. **predikční úloha**
- poznámka: v některých oblastech se pojem predikce a klasifikace rozlišuje:
 - pojem **klasifikace** je používán, použije-li se klasifikačního algoritmu pro známá data; pokud jsou data nová, pro která apriori neznáme klasifikační třídu, pak hovoříme o **predikci** klasifikační třídy
 - pojem **klasifikace** používáme, pokud vybíráme identifikátor klasifikační třídy z určitého diskrétního konečného počtu možných identifikátorů; pokud určíme (predikujeme) spojitou hodnotu, např. pomocí regrese, pak hovoříme o **predikci**, i když tento pojem nemá časovou dimenzi

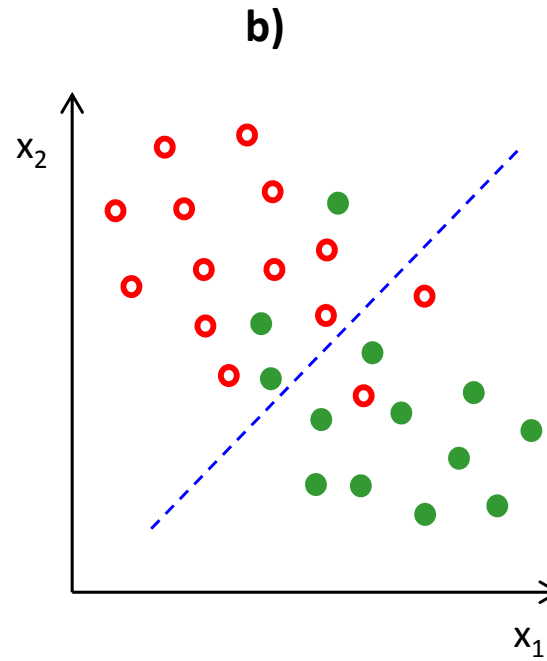
Klasifikace a další používané termíny

- termín **diskriminační analýza** často používán jako synonymum klasifikace, diskriminační analýza je však spíše podskupina klasifikačních metod
- analýza a klasifikace dat občas nazývána souhrnně i jako:
 - „rozpoznávání obrazů“ (*pattern recognition*) – obraz nejen ve smyslu obraz mozku či obraz sítnice oka, ale ve smyslu popis (tzn. „obraz“) reálného objektu
 - „dolování z dat“ (*data mining*)
 - „strojové učení“ (*machine learning*)

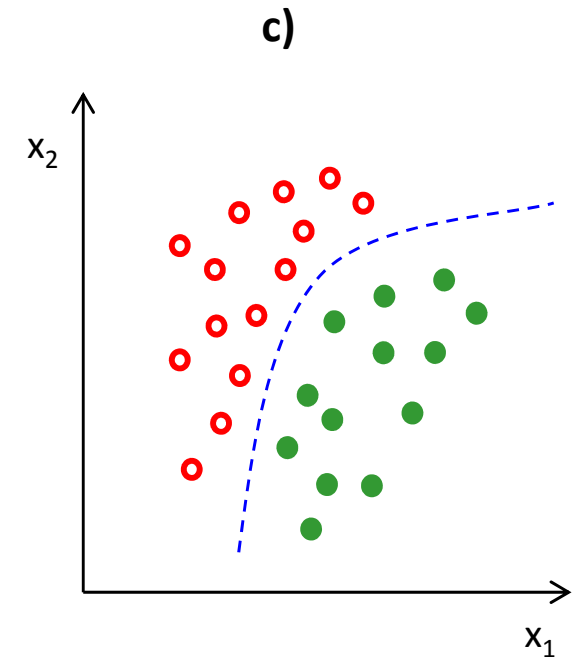
Lineární separabilita



lineárně separabilní
úloha



lineárně neseparabilní
úloha
lineárně separované
klasifikační třídy

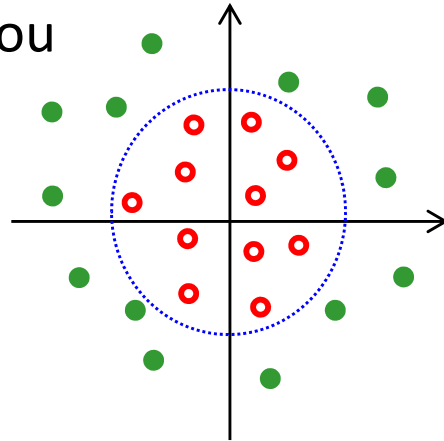


nelineárně
separabilní úloha

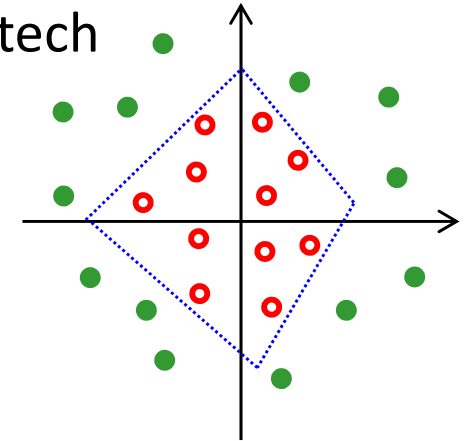
Lineárně neseparabilní třídy – způsoby řešení

1. zachováme původní obrazový prostor a zvolíme nelineární hranici:

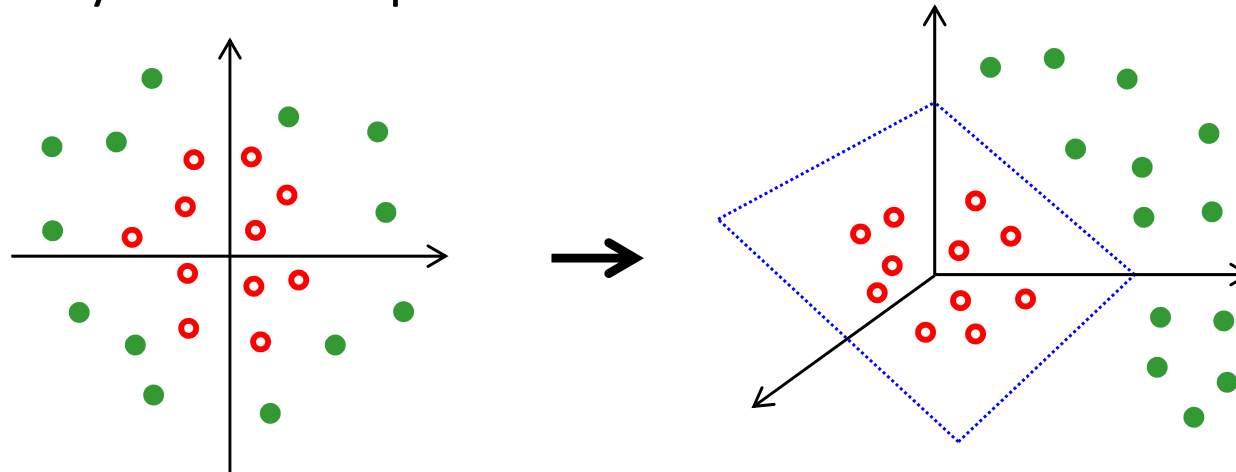
a) definovanou
obecně



b) složenou po částech
z lineárních úseků

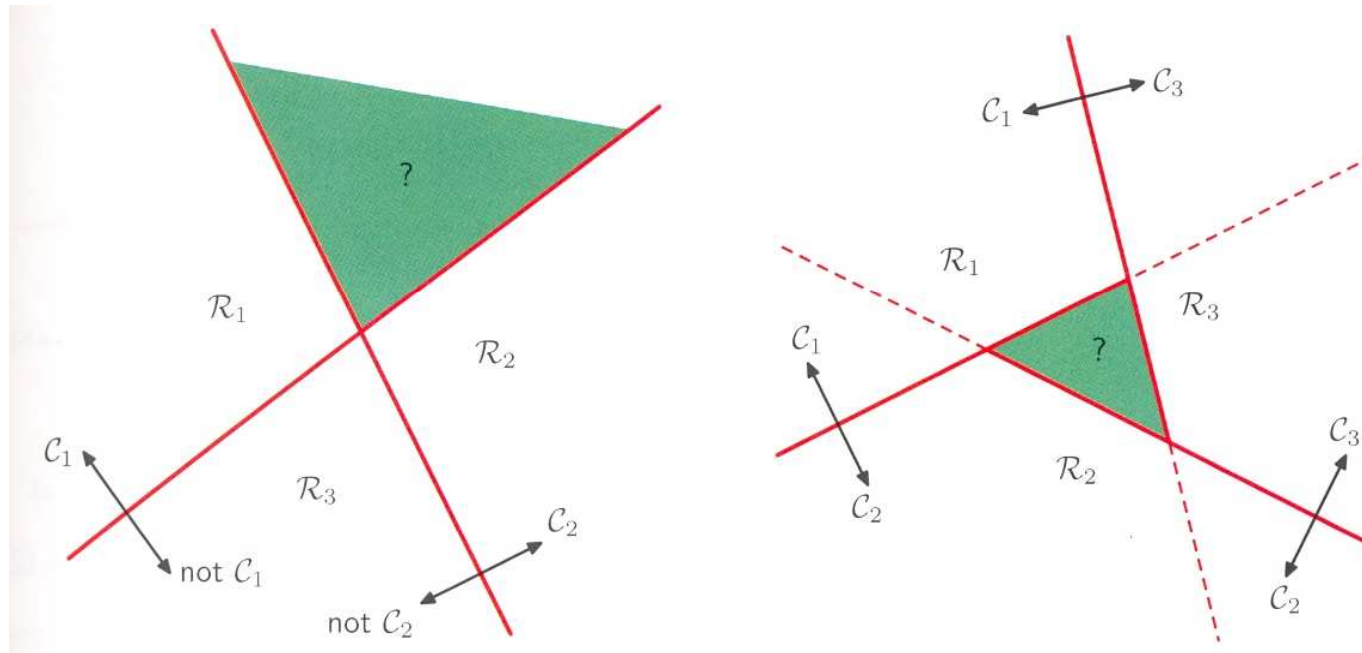


2. zobrazíme původní p -rozměrný obrazový prostor nelineární transformací do nového m -rozměrného prostoru tak, aby v novém prostoru byly klasifikační třídy lineárně separabilní



Klasifikace s více třídami

1. klasifikace „jedna versus zbytek“
R-1 hranice oddělí jednu klasifikační třídu od všech dalších
2. klasifikace „jedna versus jedna“
 $R(R-1)/2$ binárních hranic mezi každými dvěma třídami



- problematickým úsekům se můžeme vyhnout použitím diskriminačních funkcí (do r -té třídy ω_r zařadíme obraz \mathbf{x} za předpokladu, že $g_r(\mathbf{x}) > g_s(\mathbf{x})$ pro $\forall r \neq s$) \rightarrow klasifikační hranice je průmět průsečíku $g_r(\mathbf{x}) = g_s(\mathbf{x})$ do obrazového prostoru – takto definovaný klasifikační prostor je vždy spojitý a konvexní

Typy klasifikátorů – podle reprezentace vstupních dat

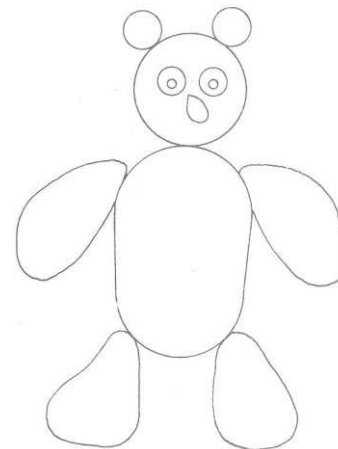
1. **Podle reprezentace vstupních dat:**
 - příznakové klasifikátory: paralelní x sekvenční
 - strukturální (syntaktické) klasifikátory
 - kombinované klasifikátory
2. **Podle jednoznačnosti zařazení do skupin:**
 - deterministické klasifikátory
 - pravděpodobnostní klasifikátory
3. **Podle typů klasifikačních a učících algoritmů:**
 - parametrické klasifikátory
 - neparametrické klasifikátory
4. **Podle způsobu učení:**
 - učení s učitelem: dokonalým x nedokonalým
 - učení bez učitele
5. **Podle principu klasifikace:**
 - klasifikace pomocí diskriminačních funkcí
 - klasifikace pomocí vzdálenosti od etalonů klasifikačních tříd
 - klasifikace pomocí hranic v obrazovém prostoru

Typy klasifikátorů – podle reprezentace vstupních dat

- **příznakové** – vstupní data vyjádřena vektorem hodnot jednotlivých proměnných (příznaků):
 - **paralelní** – zpracování vektoru jako celku (např. Bayesův klasifikátor)
 - **sekvenční** – zpracování (občas i měření) proměnných postupně (např. klasifikační stromy)

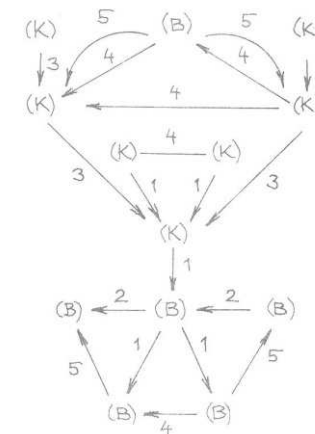
	A	B	C	D	E
1	id	vek	pohlavi	vyska	vaha
2	1	38	Z	164	45
3	2	36	M	167	90
4	3	26	Z	178	70

- **strukturální (syntaktické)** – vstupní data popsána relačními strukturami



PRIMITIVA:
(K) – KOLEČKO
(B) – BRAMBORA

RELACE:
(1) – DOTÝKÁ SE SHORA
(2) – DOTÝKÁ SE ZLEVA
(3) – LEŽÍ UVNITŘ
(4) – LEŽÍ VLEVO OD
(5) – LEŽÍ POD



- **kombinované** – jednotlivá primitiva doplněna příznakovým popisem

Typy klasifikátorů – dle jednoznačnosti zařazení do skupin

- **deterministické klasifikátory:**

- každý objekt musí patřit do nějaké třídy a nemůže být současně ve více třídách
- pozn. použití termínu „**deterministický klasifikátor**“ v případě, že klasifikátor daná data zpracuje vždy se stejným výsledkem (např. Bayesův klasifikátor) x „**nedeterministický klasifikátor**“, který může při opakovaném zpracování daných dat klasifikovat různě (např. neuronové sítě – záleží na tom, jaká bude inicializace)

- **pravděpodobnostní klasifikátory:**

- stanoví pravděpodobnost zařazení obrazů do daných klasifikačních tříd
- např. člověk má s pravděpodobností 0,6 infarkt, s pstí 0,3 má atrofii srdeční komory a s pstí 0,1 je zdravý

Typy klasifikátorů – dle typů klasifikačních a učících algoritmů

- **parametrické klasifikátory:**
 - potřeba nastavit či určit parametry
 - např. prahová klasifikace (potřeba stanovit práh), metoda podpůrných vektorů (potřeba stanovit parametr „C“) atd.
- **neparametrické klasifikátory:**
 - není potřeba nastavovat žádné parametry
 - např. klasifikace podle vzdáleností od reprezentativního objektu (tzv. „etalonu“) skupin
- pozn. z tohoto pohledu jsou klasifikační stromy parametrické klasifikátory, pokud to však hodnotíme ze statistického pohledu, jsou to neparametrické metody, protože nemají předpoklad normálního rozdělení

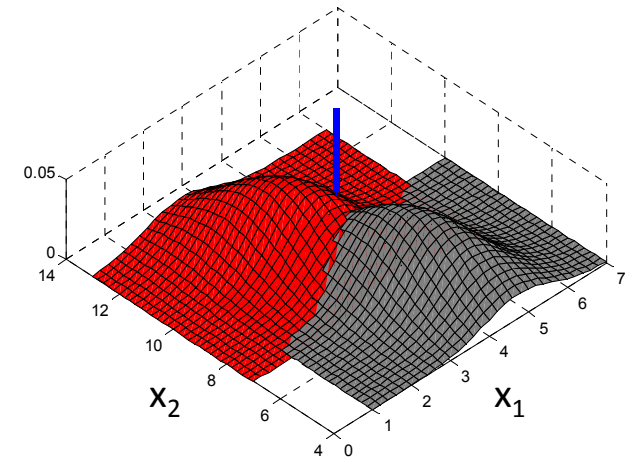
Typy klasifikátorů – podle způsobu učení

- **učení s učitelem** – k dispozici trénovací množina, u níž známe zařazení každého objektu do jednotlivých klasifikačních tříd
 - **učení s dokonalým učitelem** – učitel se nemůže splést (tzn. předpokládáme, že všechny trénovací objekty jsou správně označené, že patří do dané třídy)
 - **učení s nedokonalým učitelem** – připouštíme, že v trénovací množině mohou být nesprávně označené subjekty (např. u některých duševních onemocnění se lékař může splést a označit pacienta za schizofrenika, i když trpí bipolární poruchou, což se však prokáže až za několik let, takže v naší trénovací množině je takto špatně zařazený subjekt)
- **učení bez učitele:**
 - trénovací množina není k dispozici a často ani předem neznáme, jaké třídy (skupiny) se v datech budou vyskytovat
 - typickým příkladem je shlukování

Typy klasifikátorů – podle principu klasifikace

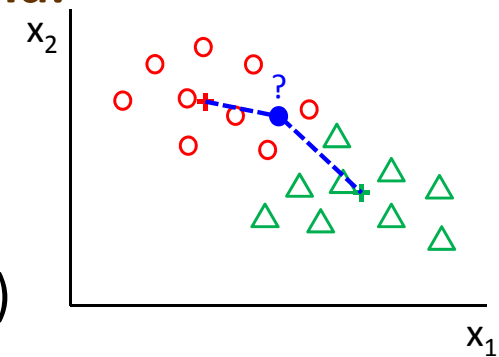
- **klasifikace pomocí diskriminačních funkcí:**

- diskriminační funkce určují míru příslušnosti k dané klasifikační třídě
- pro danou třídu má daná diskriminační funkce nejvyšší hodnotu



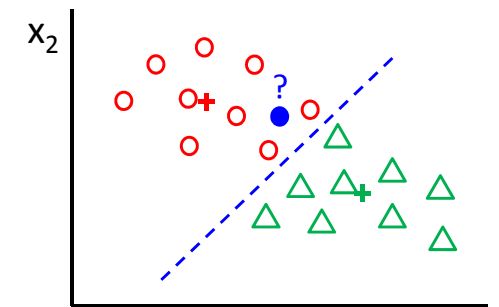
- **klasifikace pomocí min. vzdálenosti od etalonů klasif. tříd:**

- etalon = reprezentativní objekt(y) klasifikační třídy
- počet etalonů klasif. třídy různý – od jednoho vzorku (např. centroidu) po úplný výčet všech objektů dané třídy (např. u klasif. pomocí metody průměrné vazby)



- **klasifikace pomocí hranic v obrazovém prostoru:**

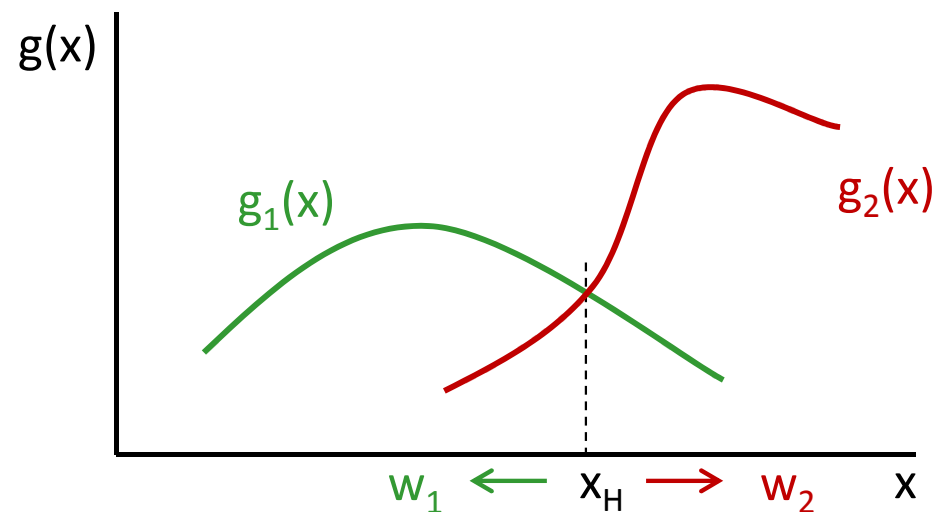
- stanovení hranic (hraničních ploch) oddělujících klasifikační třídy



Klasifikace pomocí diskriminačních funkcí

Klasifikace pomocí diskriminačních funkcí

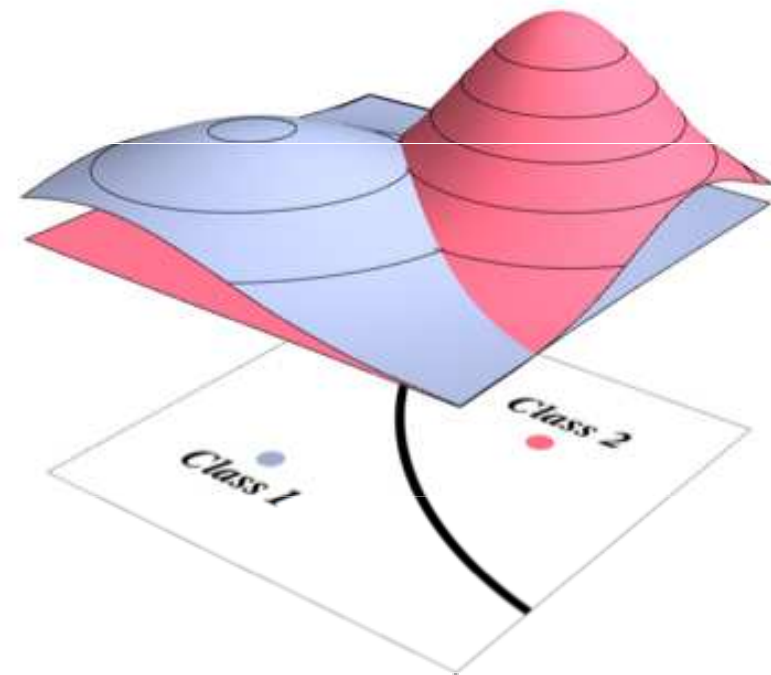
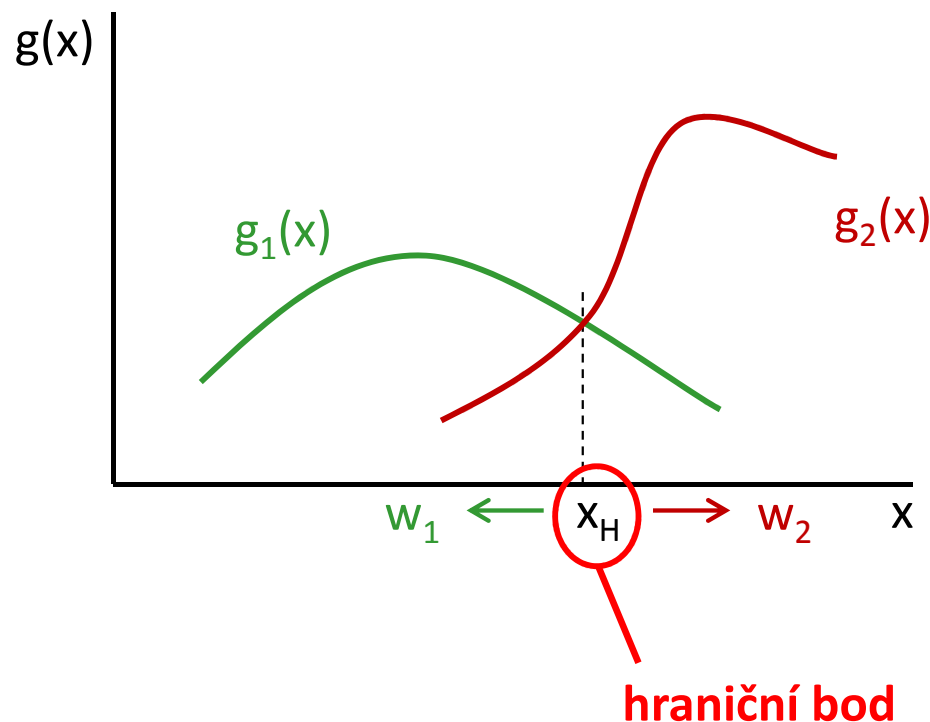
- **diskriminační funkce $g_i(\mathbf{x})$** – vyjadřují míru příslušnosti objektu \mathbf{x} do jednotlivých klasifikačních tříd
- zařadíme \mathbf{x} do takové třídy ω_i , pro kterou je $g_i(\mathbf{x})$ maximální
- matematicky: pro objekt \mathbf{x} z třídy ω_r platí, že $g_r(\mathbf{x}) > g_s(\mathbf{x})$ pro $s = 1, 2, \dots, R$ a $r \neq s$



- pro klasifikaci do dvou tříd lze rozhodovací pravidlo klasifikátoru zapsat jako:
$$\omega_k = d(\mathbf{x}) = \text{sign}(g_1(\mathbf{x}) - g_2(\mathbf{x}))$$
- pokud $d(\mathbf{x}) \geq 0 \rightarrow$ zařazení \mathbf{x} do třídy ω_1
- pokud $d(\mathbf{x}) < 0 \rightarrow$ zařazení \mathbf{x} do třídy ω_2

Souvislost klasifikace pomocí diskriminačních funkcí s klasifikací pomocí hranic

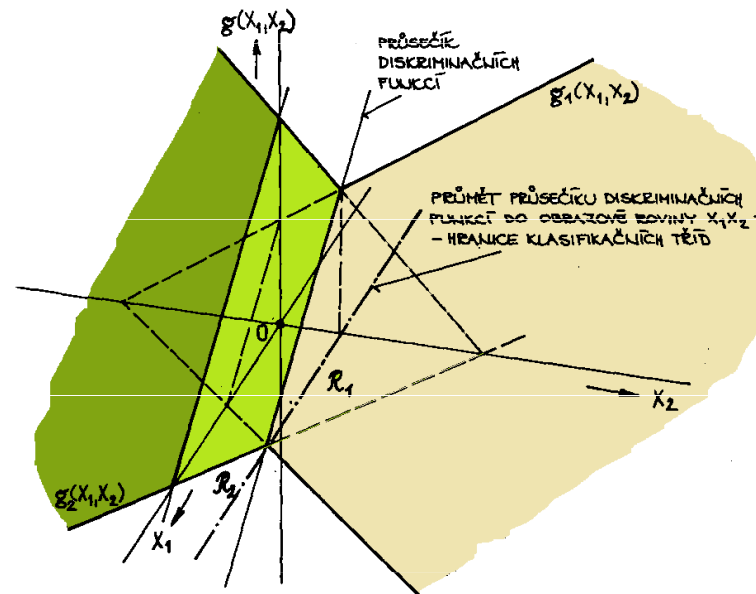
Hranice mezi dvěma sousedními třídami ω_1 a ω_2 je určena průmětem průsečíku funkcí $g_r(\mathbf{x})$ a $g_s(\mathbf{x})$, definovaného rovnicí $g_r(\mathbf{x}) = g_s(\mathbf{x})$, do obrazového prostoru.



Příklady diskriminačních funkcí

- nejjednodušším tvarem diskriminační funkce je lineární funkce:

$$g_r(\mathbf{x}) = a_{r0} + a_{r1}x_1 + a_{r2}x_2 + \dots + a_{rp}x_p$$



- diskriminační funkce na základě statistických vlastností množiny objektů:

$$g_r(\mathbf{x}) = P(\omega_r | \mathbf{x})$$

kde $P(\omega_r | \mathbf{x})$ je pravděpodobnost zatřídění \mathbf{x} do třídy ω_r

→ **Bayesův klasifikátor**

Bayesův klasifikátor

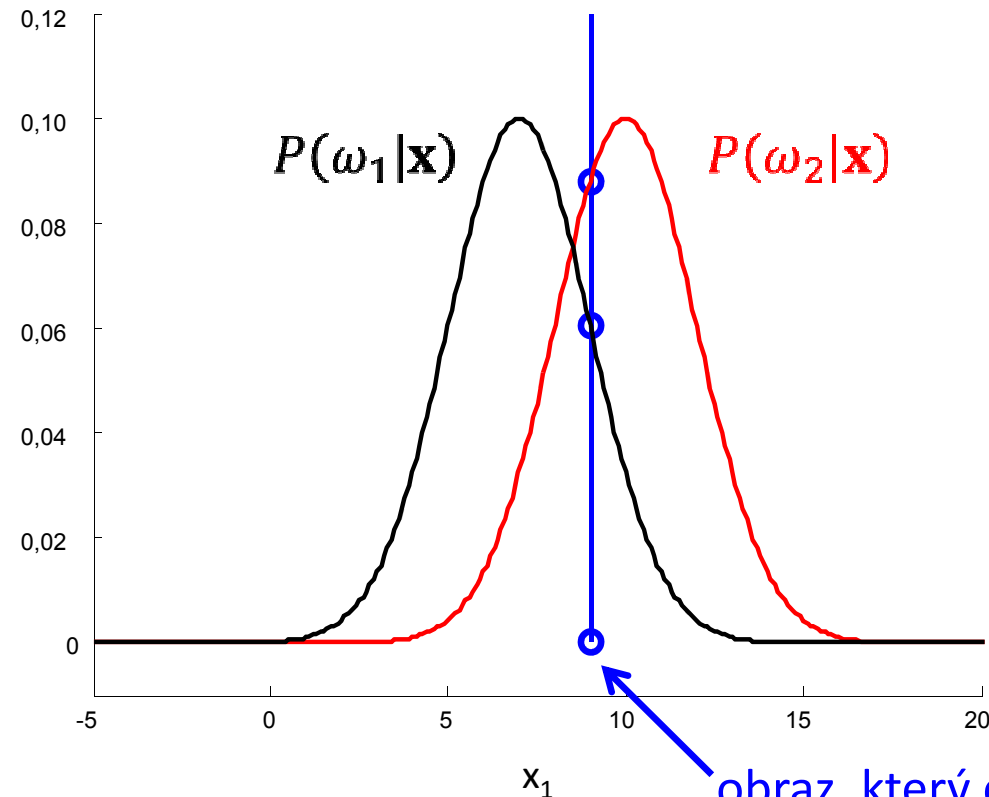
- diskriminační funkce určeny na základě statistických vlastností množiny obrazů
- vyjdeme z **Bayesova vzorce**: $P(\omega_k|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_k) \cdot P(\omega_k)}{p(\mathbf{x})}$, kde
 - $P(\omega_k|\mathbf{x})$ je aposteriorní podmíněná pravděpodobnost zatřídění obrazu \mathbf{x} do třídy ω_k
 - $p(\mathbf{x}|\omega_k)$ je podmíněná hustota pravděpodobnosti výskytu obrazu \mathbf{x} ve třídě ω_k , $k = 1, 2$
 - $P(\omega_k)$ je apriorní pravděpodobnost třídy ω_k
 - $p(\mathbf{x})$ je celková hustota pravděpodobnosti rozložení obrazu \mathbf{x} v celém obrazovém prostoru

Bayesův klasifikátor – kritéria

- Kritérium maximální a posteriorní pravděpodobnosti
- Kritérium minimální střední ztráty
- kritérií existuje více, ale tyto dvě jsou základní a ostatní z nich lze zpravidla odvodit – např.:
 - kritérium minimální pravděpodobnosti chybného rozhodnutí
 - kritérium maximální pravděpodobnosti

Bayesův kl. – kritérium maximální aposteriori psti

- zatřídění obrazu \mathbf{x} do třídy s větší aposteriori pravděpodobností, tedy:
když $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ → zařazení \mathbf{x} do třídy ω_1
když $P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x})$ → zařazení \mathbf{x} do třídy ω_2



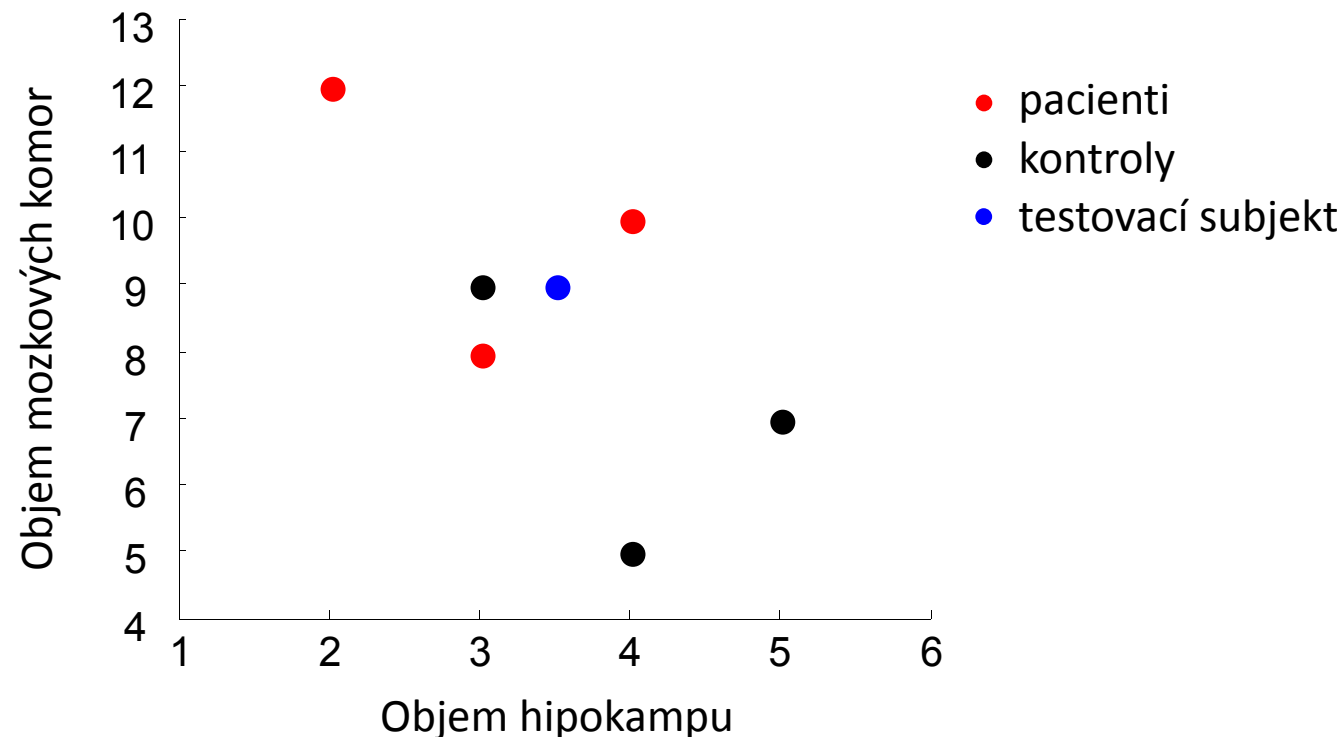
obraz, který chceme zatřídít

Bayesův kl. – kritérium maximální aposteriorní psti

Příklad: Bylo provedeno měření objemu hipokampu a mozkových komor

(v cm^3) u 3 pacientů se schizofrenií a 3 kontrol: $\mathbf{X}_D = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix}$, $\mathbf{X}_H = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}$.

Určete, zda testovací subjekt $\mathbf{x} = [3,5 \quad 9]$ patří do skupiny pacientů či kontrolních subjektů pomocí Bayesova klasifikátoru.



Bayesův kl. – kritérium maximální aposteriori psti

Příklad: Bylo provedeno měření objemu hipokampu a mozkových komor

(v cm^3) u 3 pacientů se schizofrenií a 3 kontrol: $\mathbf{X}_D = \begin{bmatrix} 2 & 12 \\ 4 & 10 \\ 3 & 8 \end{bmatrix}$, $\mathbf{X}_H = \begin{bmatrix} 5 & 7 \\ 3 & 9 \\ 4 & 5 \end{bmatrix}$.

Určete, zda testovací subjekt $\mathbf{x} = [3,5 \ 9]$ patří do skupiny pacientů či kontrolních subjektů pomocí Bayesova klasifikátoru.

$$P(\omega_k|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_k) \cdot P(\omega_k)}{p(\mathbf{x})}$$

Označení a pomocné výpočty:

$$n_D = 3; \quad n_H = 3; \quad n = 6$$

Apriorní psti:

$$P(\omega_D) = \frac{n_D}{n} = \frac{3}{6} = 0,5$$

$$P(\omega_H) = \frac{n_H}{n} = \frac{3}{6} = 0,5$$

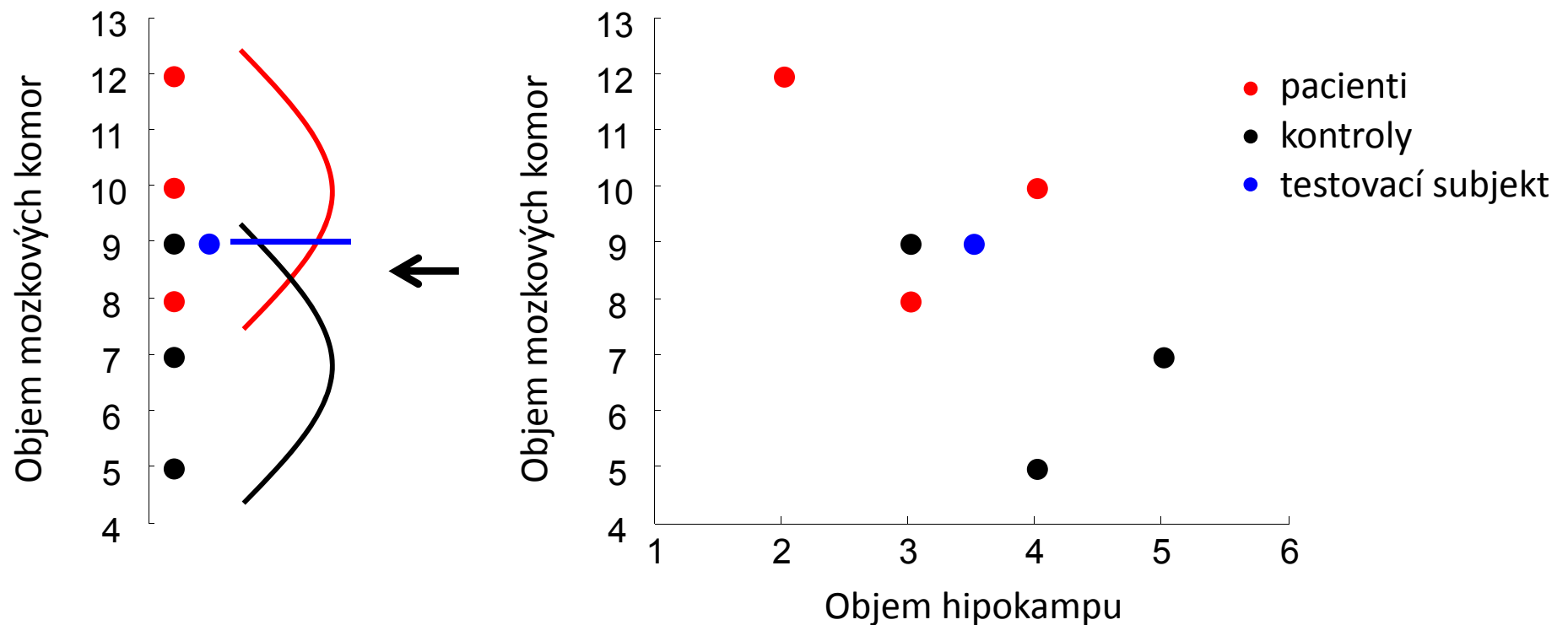
Podmíněné hustoty psti:

$$p(\mathbf{x}|\omega_k) = \frac{1}{\sqrt{(2\pi)^2 |\mathbf{S}_k|}} \cdot \exp\left(-\frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})^T \mathbf{S}_k^{-1}(\mathbf{x} - \bar{\mathbf{x}})\right)$$

Bayesův kl. – kritérium maximální aposteriorní psti

Příklad:

1. Klasifikace podle objemu mozkových komor:



$$P(\omega_D|x_2) = \frac{0,176 \cdot 0,5}{0,1485} = 0,593$$

$$P(\omega_H|x_2) = \frac{0,121 \cdot 0,5}{0,1485} = 0,407$$

→ subjekt zařazen do třídy pacientů

Bayesův kl. – kritérium maximální aposteriorní psti

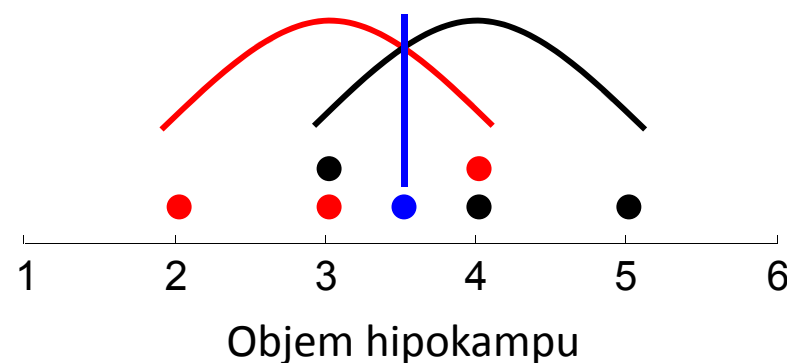
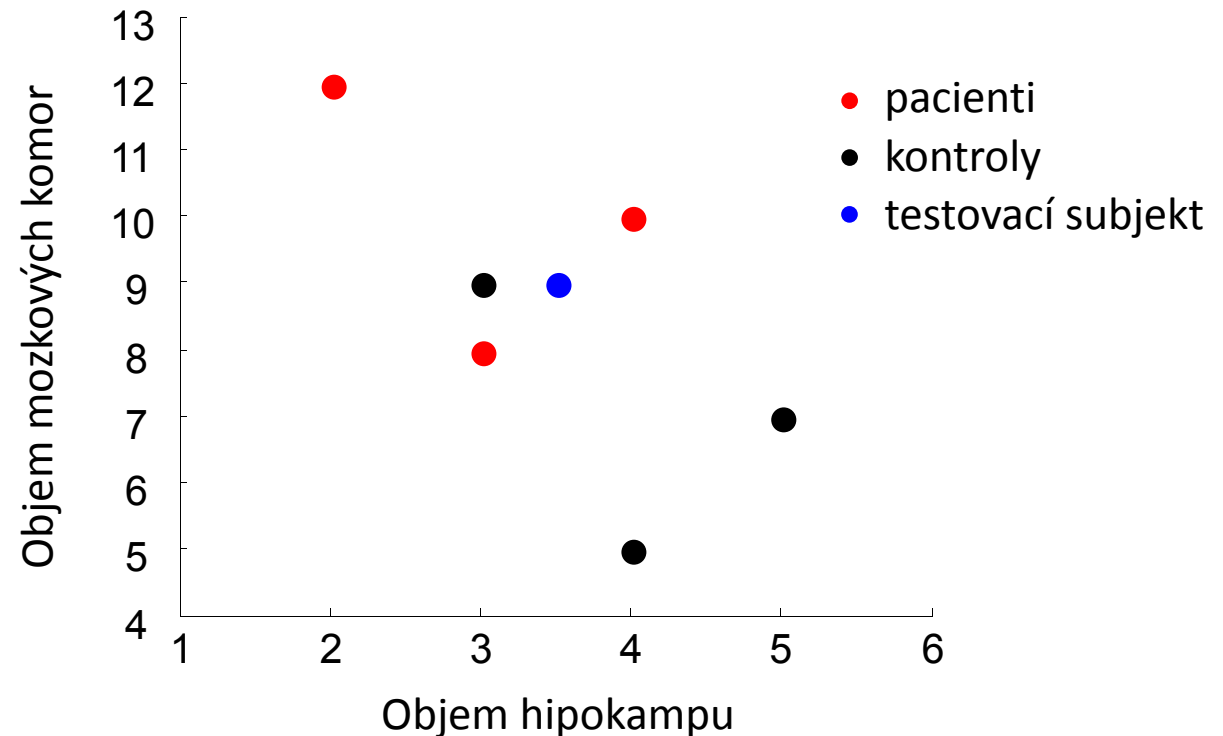
Příklad:

2. Klasifikace podle
objemu hipokampu:

$$P(\omega_D|x_1) = \frac{0,352 \cdot 0,5}{0,352} = 0,5$$

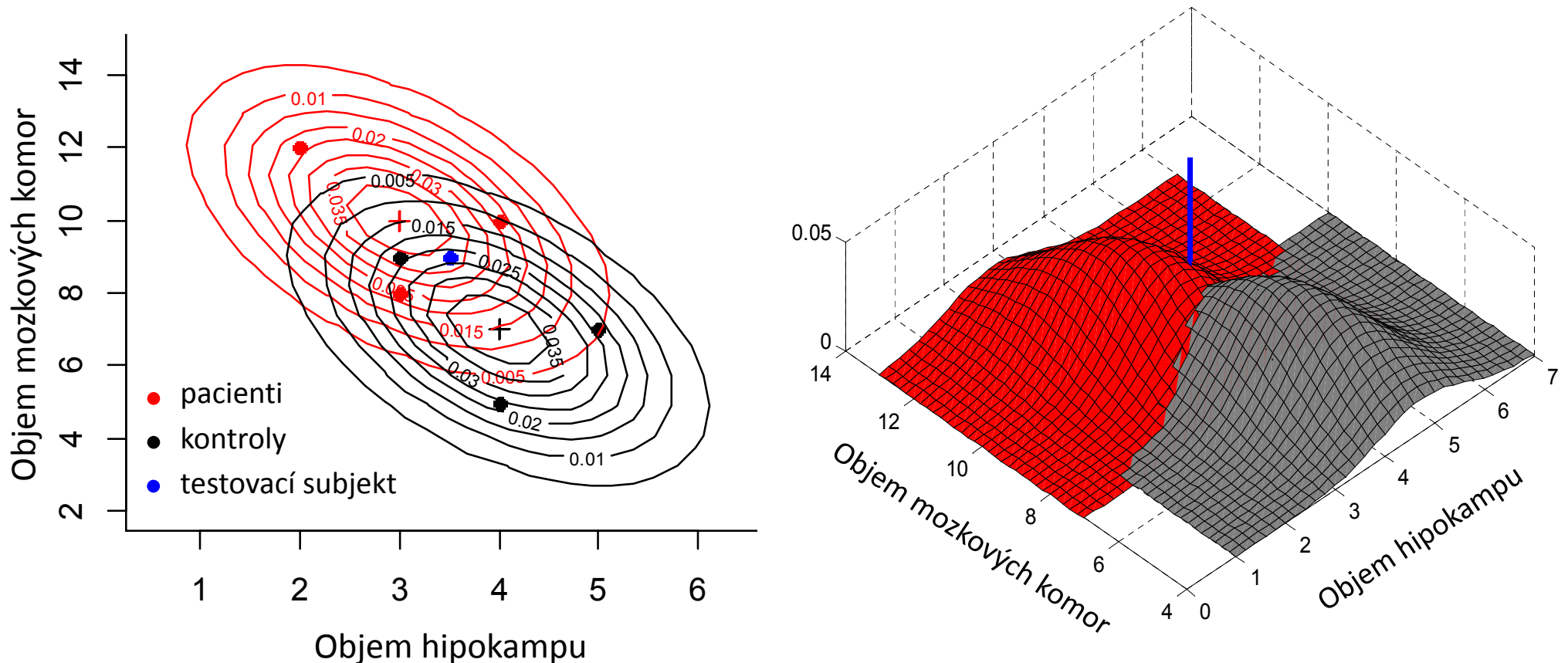
$$P(\omega_H|x_1) = \frac{0,352 \cdot 0,5}{0,352} = 0,5$$

→ nelze jednoznačně určit,
kam subjekt zařadíme



Bayesův kl. – kritérium maximální aposteriorní psti

Příklad – klasifikace podle obou proměnných:



$$P(\omega_D|\mathbf{x}) = \frac{0,078 \cdot 0,5}{0,067} = 0,582$$

$$P(\omega_H|\mathbf{x}) = \frac{0,056 \cdot 0,5}{0,067} = 0,418$$

→ subjekt zařazen do třídy pacientů

Bayesův kl. – kritérium minimální střední ztráty

- pokud rozepíšeme $P(\omega_1|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_1) \cdot P(\omega_1)}{p(\mathbf{x})}$ a $P(\omega_2|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_2) \cdot P(\omega_2)}{p(\mathbf{x})}$, pak kritérium maximální aposteriorní pravděpodobnosti:
 - když $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x}) \rightarrow$ zařazení \mathbf{x} do třídy ω_1
 - když $P(\omega_1|\mathbf{x}) < P(\omega_2|\mathbf{x}) \rightarrow$ zařazení \mathbf{x} do třídy ω_2
- můžeme přepsat jako:
 - když $p(\mathbf{x}|\omega_1) \cdot P(\omega_1) \geq p(\mathbf{x}|\omega_2) \cdot P(\omega_2) \rightarrow$ zařazení \mathbf{x} do třídy ω_1
 - když $p(\mathbf{x}|\omega_1) \cdot P(\omega_1) < p(\mathbf{x}|\omega_2) \cdot P(\omega_2) \rightarrow$ zařazení \mathbf{x} do třídy ω_2
- přičemž $p(\mathbf{x})$ můžeme vypustit, protože je v obou zlomcích stejné
- pokud chceme do výpočtů zahrnout ztrátu při chybné klasifikaci obrazu ze třídy ω_s do třídy ω_r (ztráta definována pomocí **ztrátové funkce** $\lambda(\omega_r|\omega_s)$), dostáváme:
 - když $p(\mathbf{x}|\omega_1) \cdot P(\omega_1) \cdot (\lambda(\omega_2|\omega_1) - \lambda(\omega_1|\omega_1)) \geq p(\mathbf{x}|\omega_2) \cdot P(\omega_2) \cdot (\lambda(\omega_1|\omega_2) - \lambda(\omega_2|\omega_2)) \rightarrow$ zař. \mathbf{x} do ω_1
 - když $p(\mathbf{x}|\omega_1) \cdot P(\omega_1) \cdot (\lambda(\omega_2|\omega_1) - \lambda(\omega_1|\omega_1)) < p(\mathbf{x}|\omega_2) \cdot P(\omega_2) \cdot (\lambda(\omega_1|\omega_2) - \lambda(\omega_2|\omega_2)) \rightarrow$ zař. \mathbf{x} do ω_2

Bayesův kl. – kritérium minimální střední ztráty

- ztrátové funkce $\lambda(\omega_r|\omega_s)$ se obvykle zapisují do **matice ztrátových funkcí**:

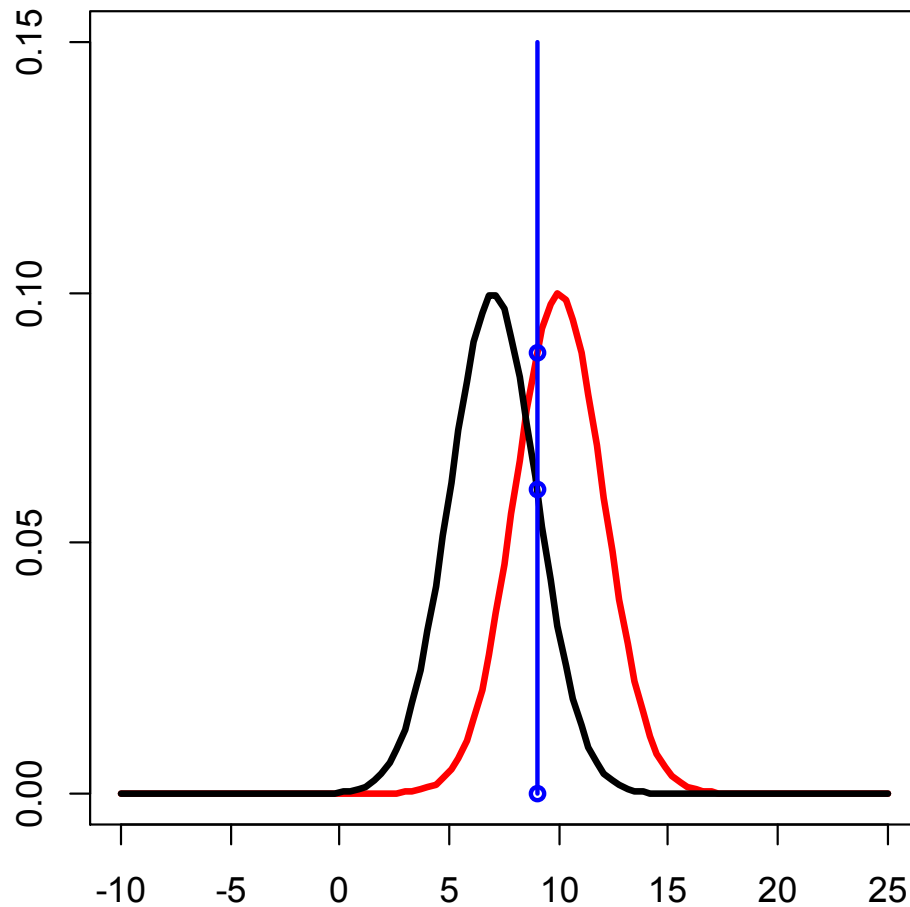
$$\mathbf{\lambda} = \begin{bmatrix} \lambda(\omega_1|\omega_1) & \lambda(\omega_1|\omega_2) & \cdots & \lambda(\omega_1|\omega_R) \\ \lambda(\omega_2|\omega_1) & \lambda(\omega_2|\omega_2) & \cdots & \lambda(\omega_2|\omega_R) \\ \vdots & \vdots & \ddots & \vdots \\ \lambda(\omega_R|\omega_1) & \lambda(\omega_R|\omega_2) & \cdots & \lambda(\omega_R|\omega_R) \end{bmatrix}$$

- prvky na diagonále $\lambda(\omega_1|\omega_1)$ bývají zpravidla nulové, protože při správném zařazení objektu ze třídy ω_1 do třídy ω_1 nevzniká žádná ztráta
- např. $\lambda = \begin{bmatrix} 0 & 1 \\ 2 & 0 \end{bmatrix}$ \rightarrow víc penalizují, když je pacient nesprávně zařazen do třídy kontrolních subjektů (ω_2), než když je kontrolní subjekt nesprávně zařazen do třídy pacientů (ω_1)
- např. $\lambda = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}$ \rightarrow víc penalizují, když je kontrolní subjekt nesprávně zařazen do třídy pacientů (ω_1), než když je pacient nesprávně zařazen do třídy kontrolních subjektů (ω_2)

Bayesův klasifikátor – poznámka

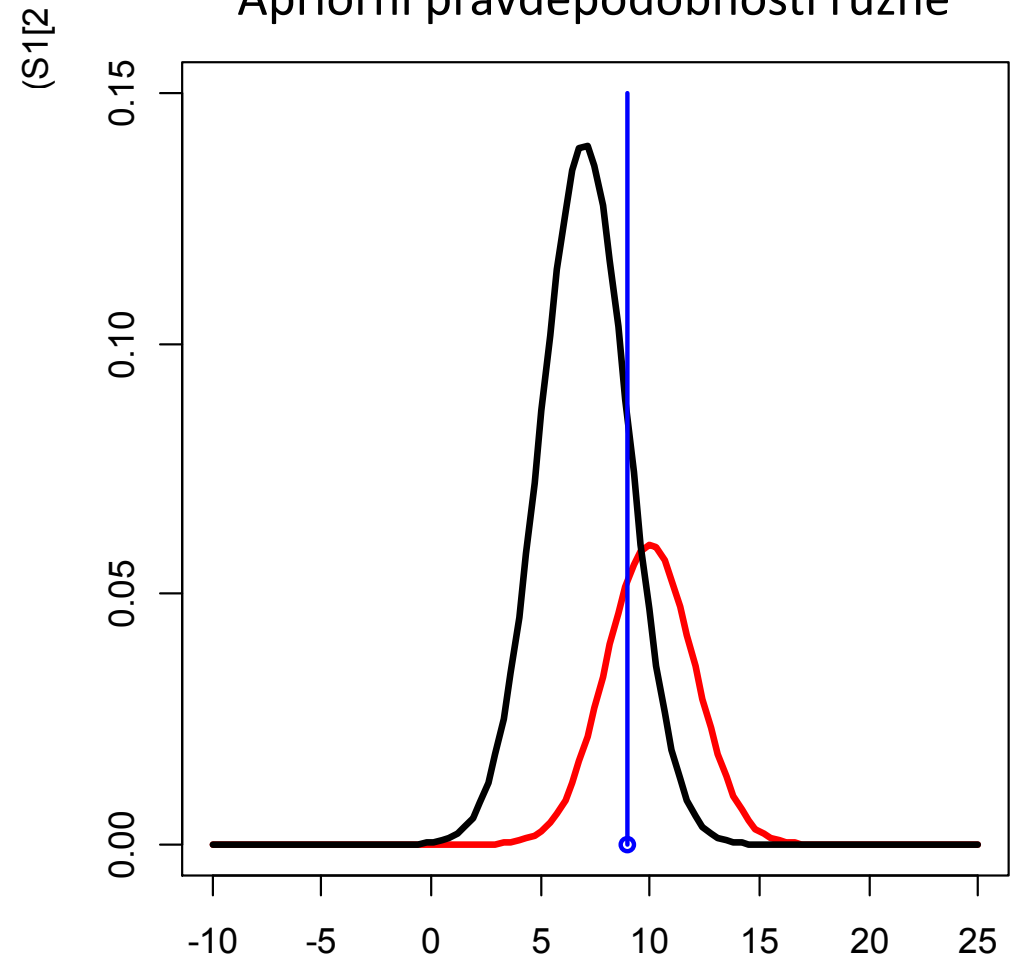
- kromě nastavování ztrát je možné nastavovat i apriorní pravděpodobnosti

Apriorní pravděpodobnosti stejné



→ zařazení objektu do červené třídy

Apriorní pravděpodobnosti různé



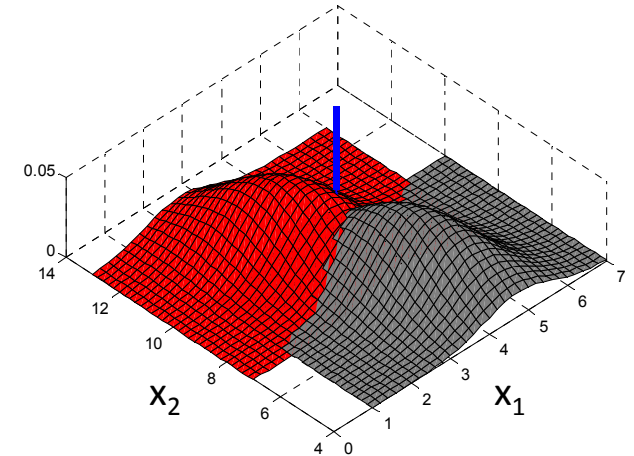
→ zařazení objektu do černé třídy

Klasifikace pomocí minimální vzdálenosti

Typy klasifikátorů – podle principu klasifikace

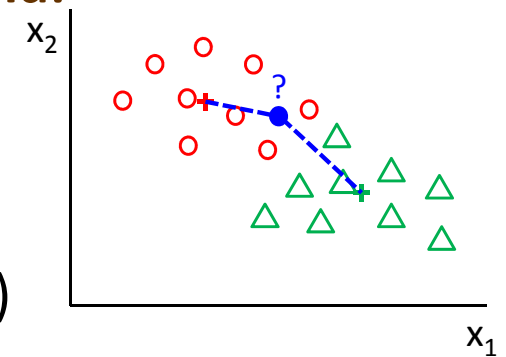
- **klasifikace pomocí diskriminačních funkcí:**

- diskriminační funkce určují míru příslušnosti k dané klasifikační třídě
- pro danou třídu má daná diskriminační funkce nejvyšší hodnotu



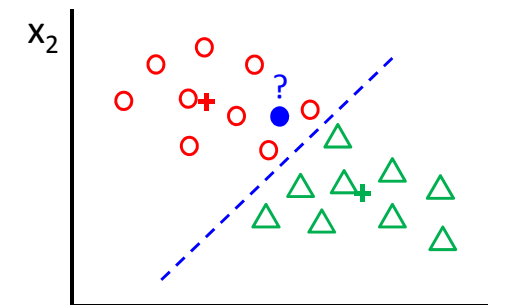
- **klasifikace pomocí min. vzdálenosti od etalonů klasif. tříd:**

- etalon = reprezentativní objekt(y) klasifikační třídy
- počet etalonů klasif. třídy různý – od jednoho vzorku (např. centroidu) po úplný výčet všech objektů dané třídy (např. u klasif. pomocí metody průměrné vazby)

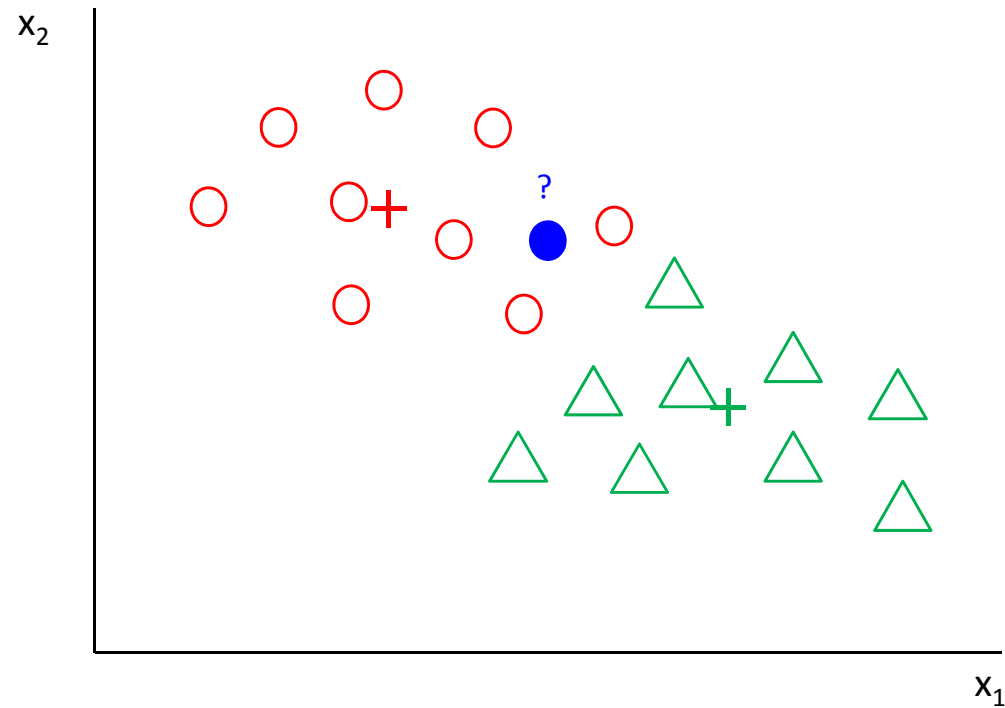


- **klasifikace pomocí hranic v obrazovém prostoru:**

- stanovení hranic (hraničních ploch) oddělujících klasifikační třídy



Klasifikace pomocí minimální vzdálenosti



- nutno zvolit metriku vzdálenosti či podobnosti:
 1. mezi jednotlivými objekty
 2. mezi množinami objektů

Typy metrik a konkrétní příklady – opakování

MEZI DVĚMA OBJEKTY

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvantitativními** proměnnými

Euklidova m., Hammingova (manhattanská) m., Minkovského m., Čebyševova m., Mahalanobisova m., Canberrská m.

Metriky pro určení **podobnosti** 2 objektů s **kvantitativními** proměnnými

Skalární součin, m. kosinové podobnosti, Pearsonův korelační koeficient, Tanimotova m.

Metriky pro určení **vzdálenosti** mezi 2 objekty s **kvalitativními** proměnnými

Hammingova m.

Metriky pro určení **podobnosti** 2 objektů s **kvalitativními** proměnnými

Tanimotova m., Jaccardův-Tanimotův a.k., Russelův-Raovův a.k., Sokalův-Michenerův a.k., Dicův k., Rogersův-Tanimotův k., Hamanův k.

MEZI DVĚMA MNOŽINAMI OBJEKTŮ

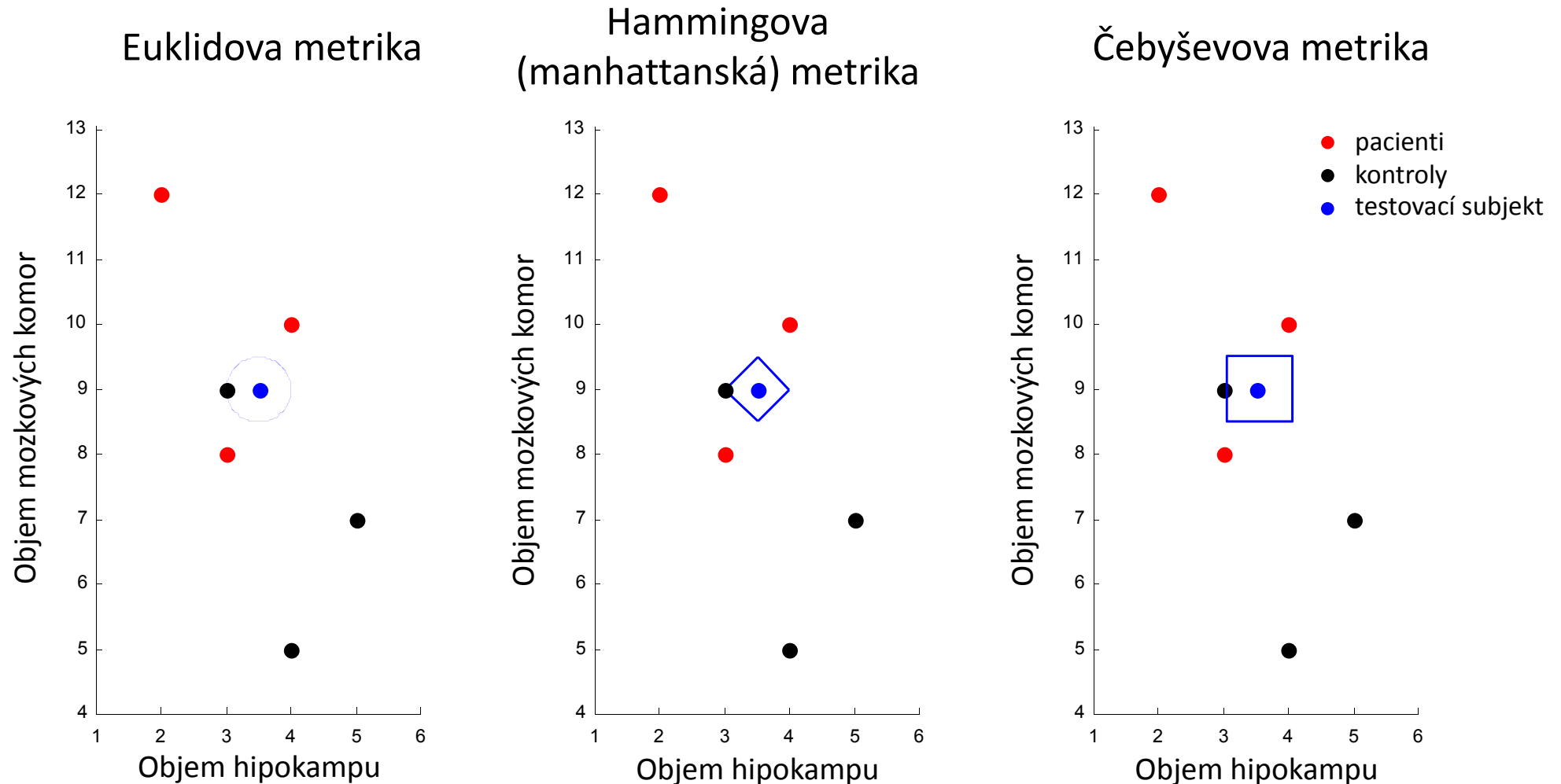
Deterministické metriky pro určení vzdálenosti mezi 2 množinami objektů

Metoda nejbližšího souseda, k nejbližších sousedů, nejvzdálenějšího souseda, průměrné vazby, Wardova metoda

Metriky pro určení vzdálenosti mezi 2 množinami objektů používající jejich pravděpodobnostní charakteristiky

Chernoffova m., Bhattacharyyova m. atd.

Euklidova, Hammingova (manhattanská), Čebyševova metrika – opakování



- zobecnění těchto 3 metrik: **Minkovského metrika**
- začleněním inverze kovarianční matice získáváme **Mahalanobisovu metriku**

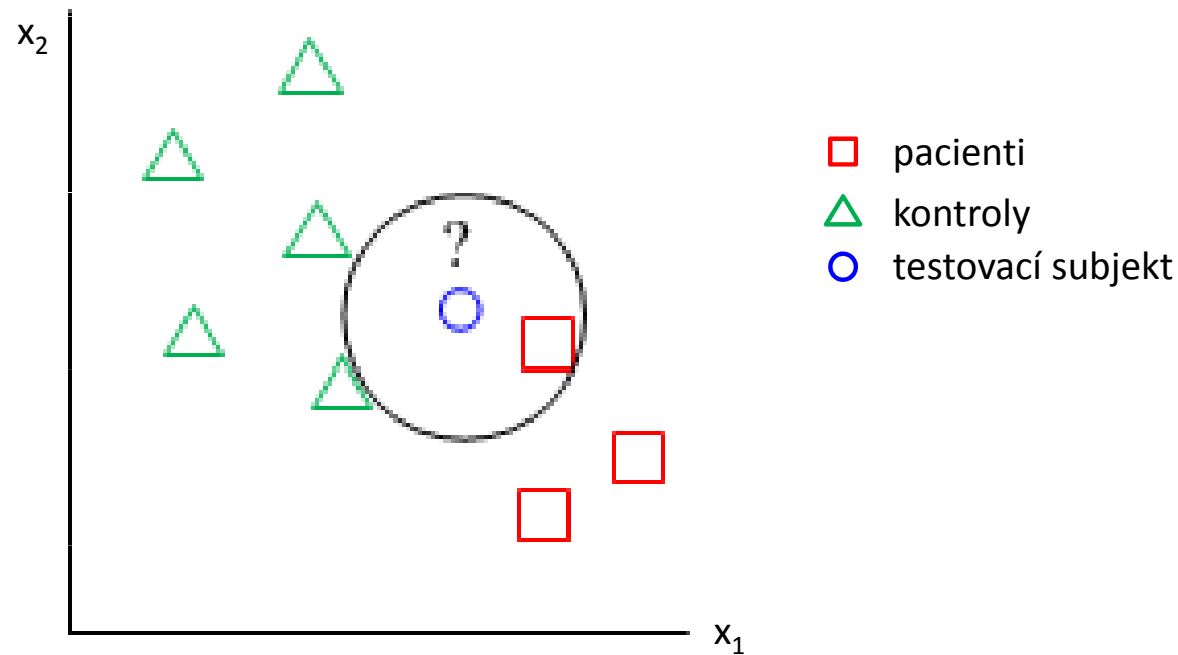
Nejpoužívanější metriky pro určení vzdálenosti mezi dvěma množinami obrazů – opakování

- Metoda nejbližšího souseda
- Metoda k nejbližších sousedů
- Metoda nejvzdálenějšího souseda – obtížně použitelná pro klasifikaci
- Centroidová metoda
- Metoda průměrné vazby
- Wardova metoda – zřídka používaná pro klasifikaci

- poznámka: podobnost (resp. vzdálenost) mezi třídami dána:
 - „podobností“ jednoho obrazu s jedním či více obrazy jedné třídy (skupin, shluků) – použitelné při klasifikaci
 - „podobností“ skupin obrazů či „podobností“ jednoho obrazu z každé skupiny – použitelné při shlukování

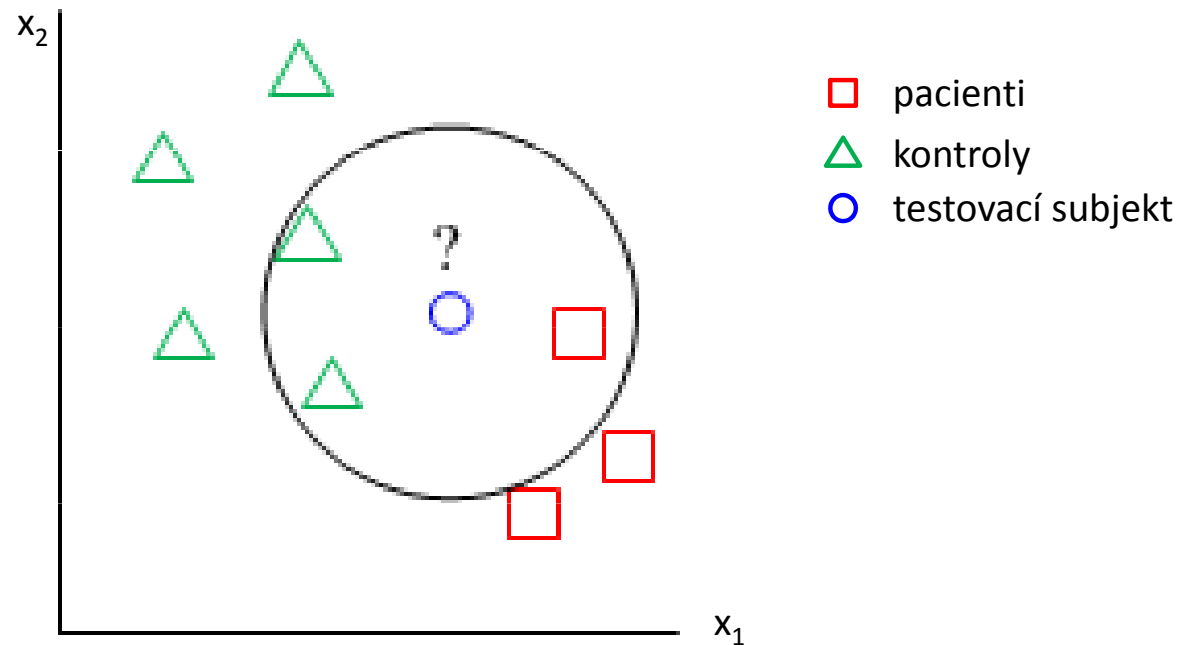
Metoda nejbližšího souseda

- je-li d libovolná míra nepodobnosti (vzdálenosti) dvou obrazů a ω_i a ω_j jsou libovolné skupiny obrazů, potom metoda nejbližšího souseda definuje mezi skupinami ω_i a ω_j vzdálenost $D_{NN}(\omega_i, \omega_j) = \min_{\substack{x_p \in \omega_i \\ x_q \in \omega_j}} d(x_p, x_q)$



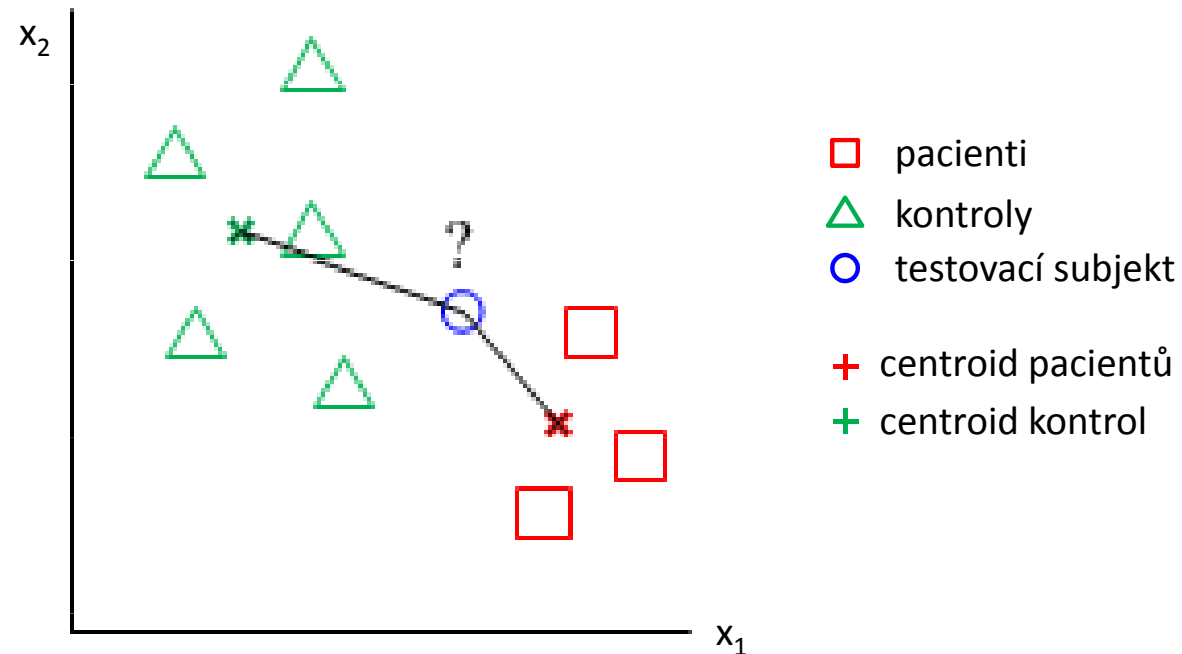
Metoda k nejbližších sousedů

- zobecněním metody nejbližšího souseda
- definována vztahem $D_{NNk}(\omega_i, \omega_j) = \min_{\substack{x_p \in \omega_i \\ x_q \in \omega_j}} \sum^k d(x_p, x_q)$, tzn. vzdálenost dvou shluků je definována součtem nejkratších vzdáleností mezi obrazy obou skupin



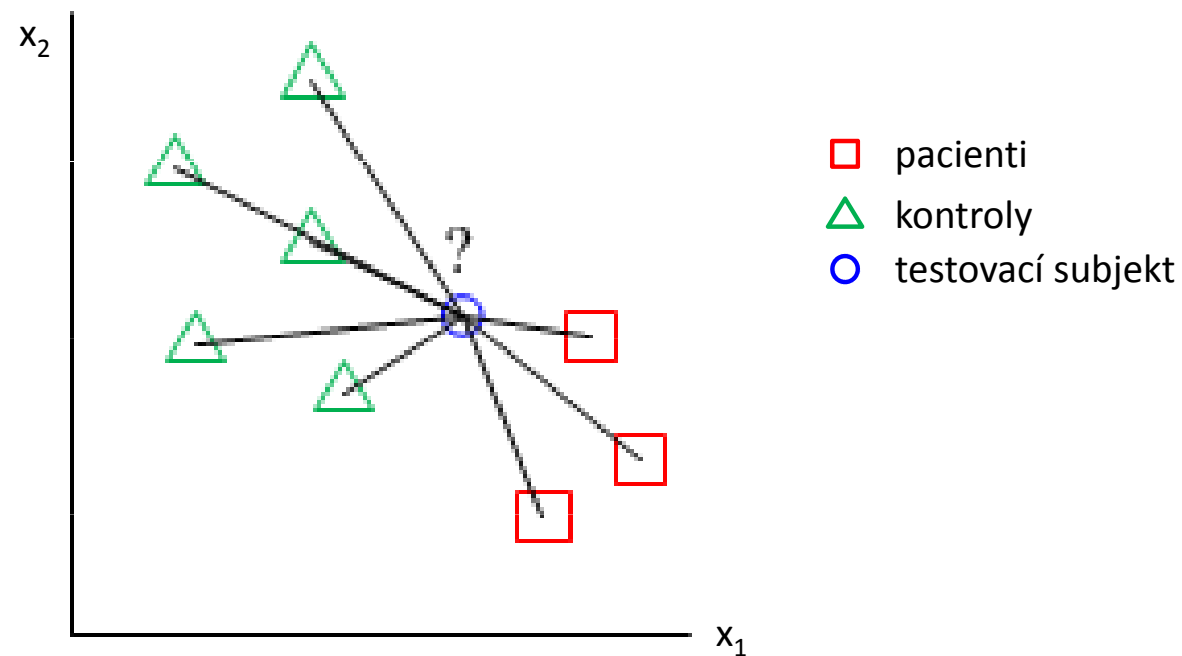
Centroidová metoda

- vychází z výpočtu centroidů pro jednotlivé třídy ω_i a ω_j
- při klasifikaci: zařazení subjektu do třídy s nejbližším centroidem



Metoda průměrné vazby

- vzdálenost dvou tříd ω_i a ω_j je průměrná vzdálenost mezi všemi obrazy tříd ω_i a ω_j
- při klasifikaci: zařazení subjektu do skupiny s nejmenší průměrnou vzdáleností od všech obrazů dané skupiny

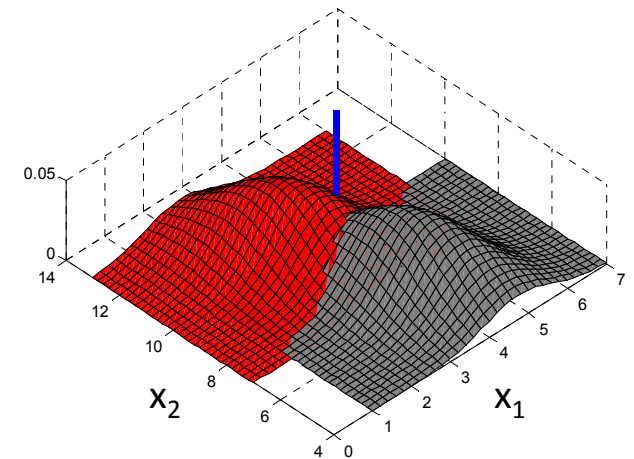


Klasifikace pomocí hranic

Typy klasifikátorů – podle principu klasifikace

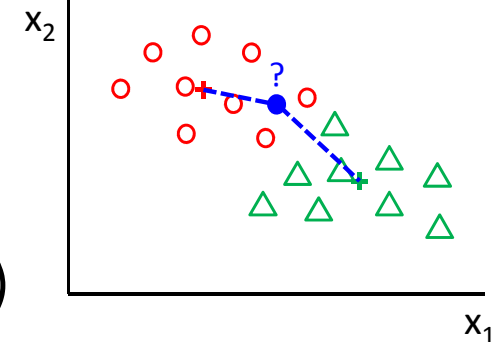
- **klasifikace pomocí diskriminačních funkcí:**

- diskriminační funkce určují míru příslušnosti k dané klasifikační třídě
- pro danou třídu má daná diskriminační funkce nejvyšší hodnotu



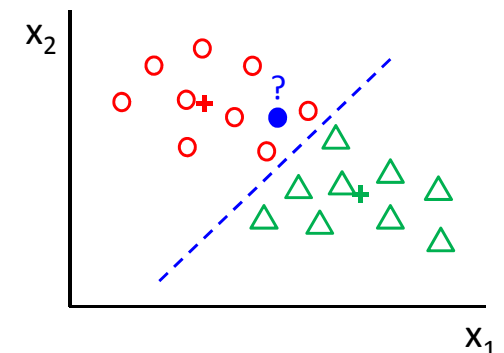
- **klasifikace pomocí min. vzdálenosti od etalonů klasif. tříd:**

- etalon = reprezentativní objekt(y) klasifikační třídy
- počet etalonů klasif. třídy různý – od jednoho vzorku (např. centroidu) po úplný výčet všech objektů dané třídy (např. u klasif. pomocí metody průměrné vazby)



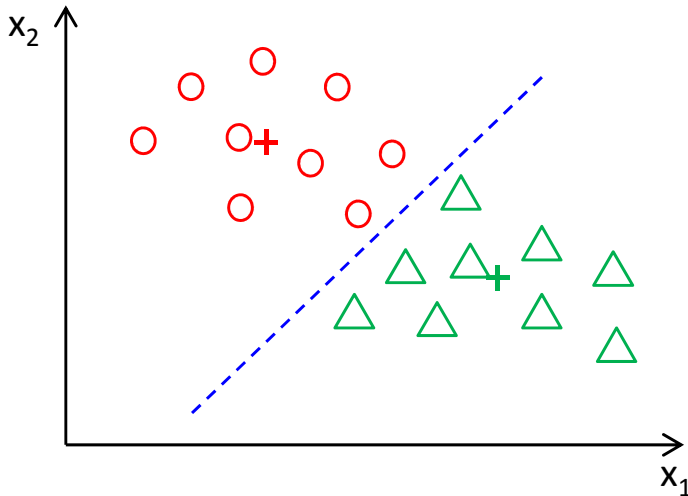
- **klasifikace pomocí hranic v obrazovém prostoru:**

- stanovení hranic (hraničních ploch) oddělujících klasifikační třídy

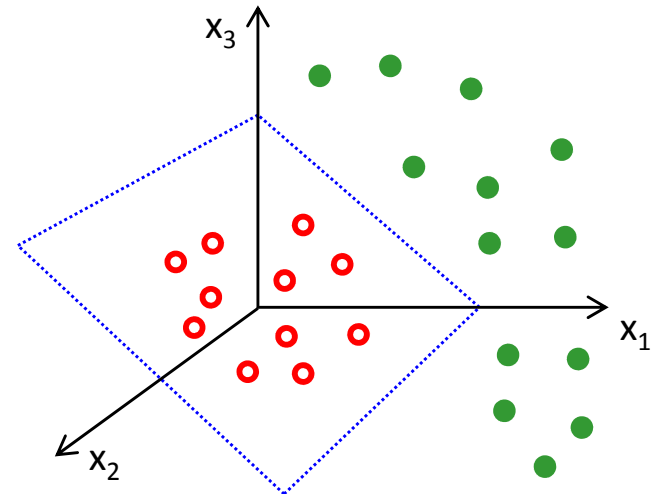


Motivace

2-rozměrný prostor



3-rozměrný prostor



Hranice je nadplocha o rozměru o jedna menší než je rozměr prostoru

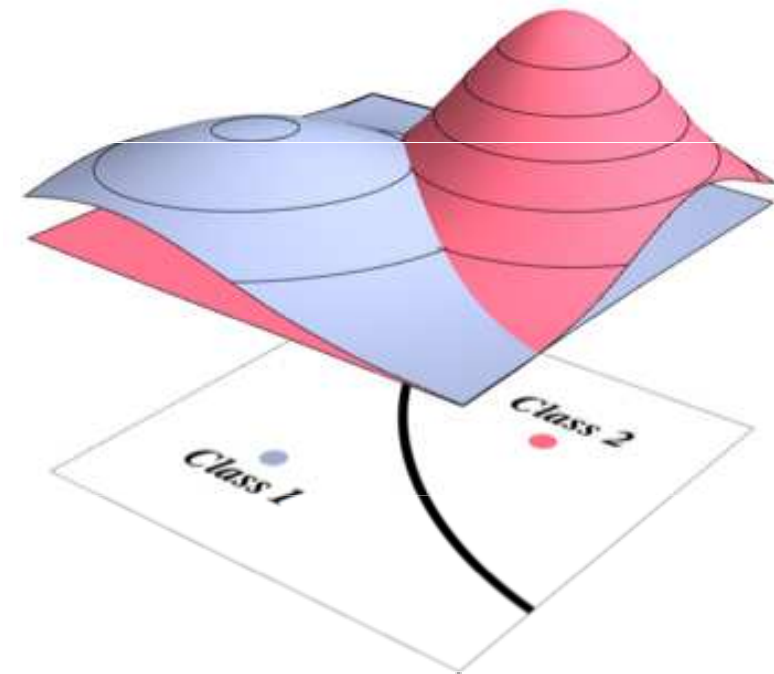
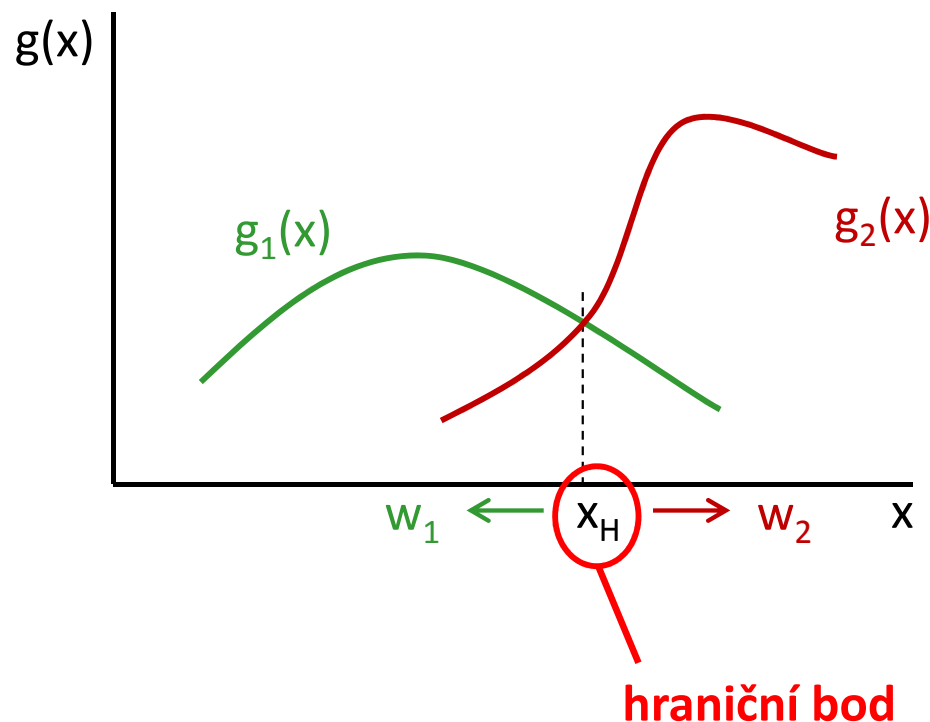
- ve 2-rozměrném prostoru je hranicí křivka (v lineárním případě přímka)
- v 3-rozměrném prostoru plocha (v lineárním případě rovina)

Hranice je tedy dána rovnicí: $\mathbf{h}(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0 = 0$

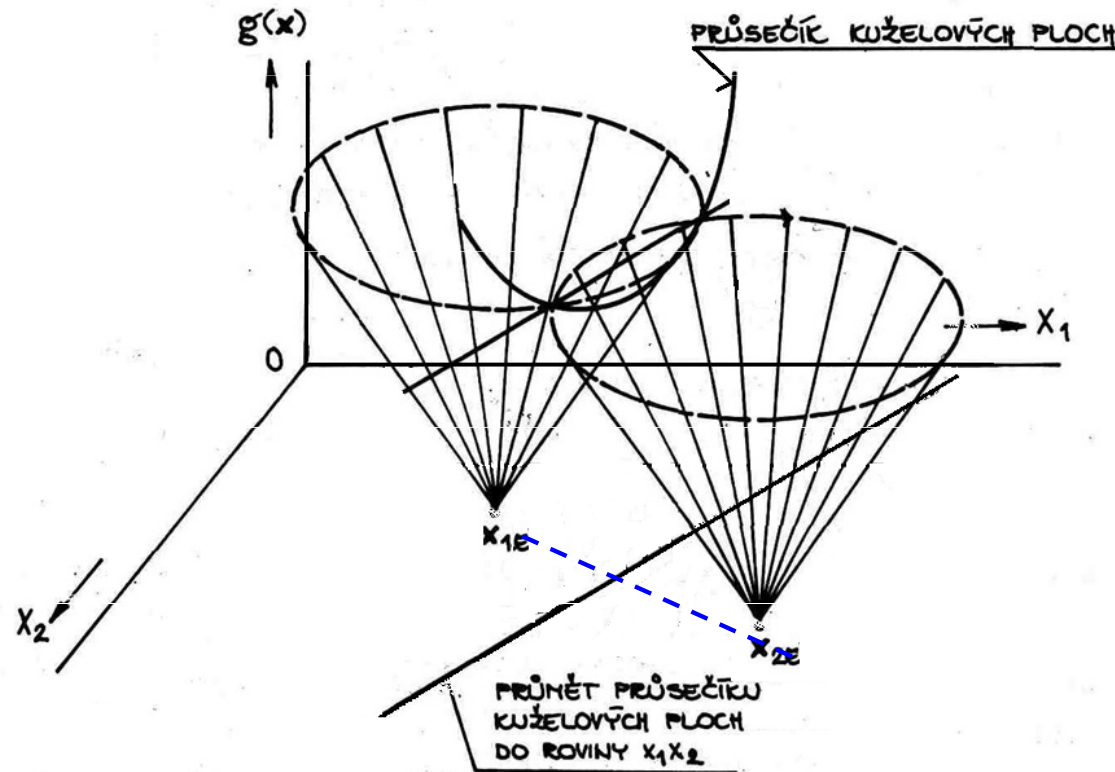
Výpočet hranice různými metodami (např. Fisherova LDA, SVM, perceptron, metoda nejmenších čtverců apod.)

Souvislost klasifikace pomocí diskriminačních funkcí s klasifikací pomocí hranic

Hranice mezi dvěma sousedními třídami ω_1 a ω_2 je určena průmětem průsečíku funkcí $g_r(\mathbf{x})$ a $g_s(\mathbf{x})$, definovaného rovnicí $g_r(\mathbf{x}) = g_s(\mathbf{x})$, do obrazového prostoru, tzn. $h(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) = 0$



Souvislost klasifikace podle minimální vzdálenosti s klasifikací pomocí hranic



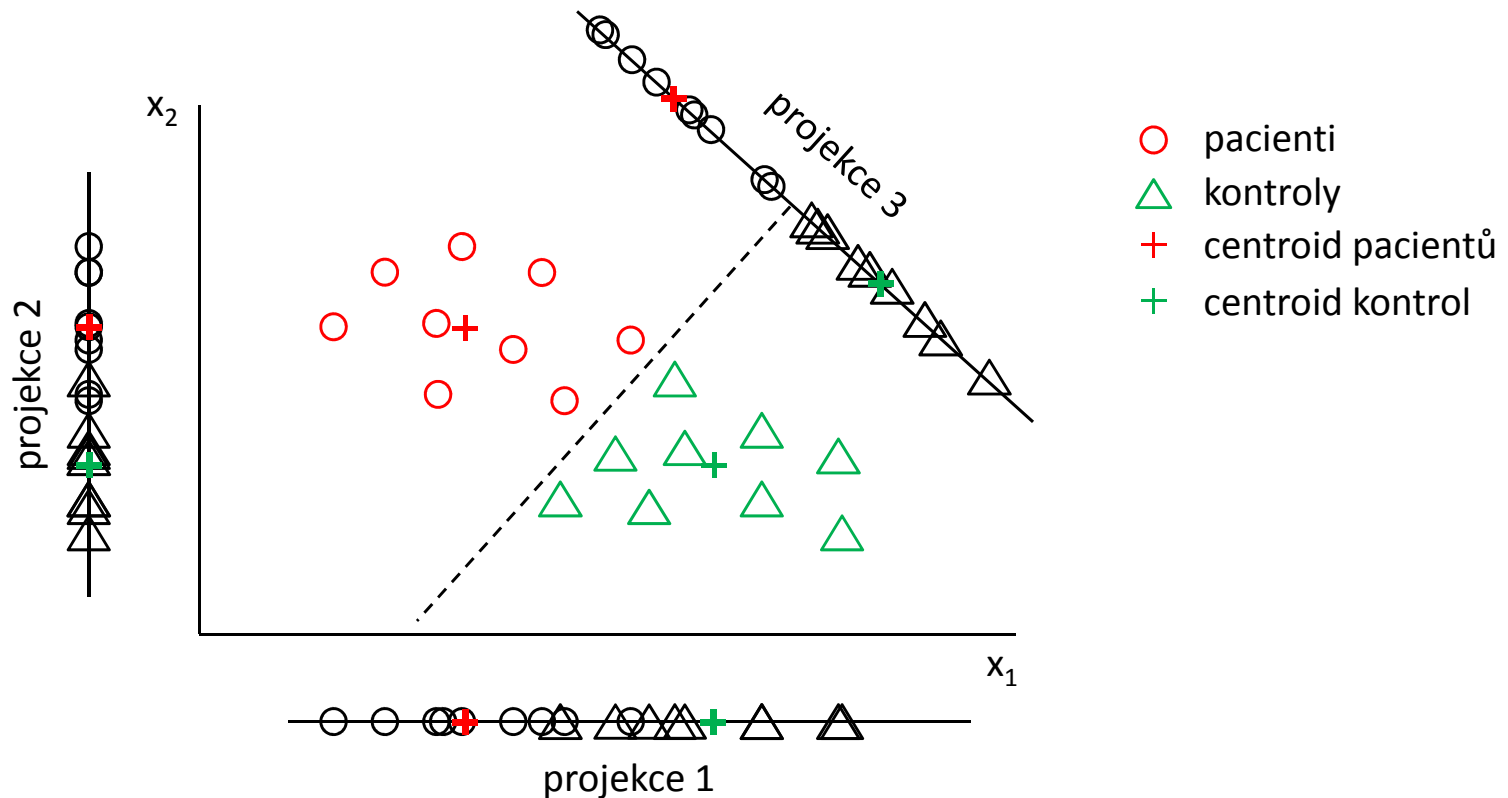
- body se stejnou vzdáleností od etalonů leží na kuželových plochách, které se protínají v parabole, jejíž průmět do obrazové roviny je přímka
- tato hraniční přímka mezi klasifikačními třídami je vždy **kolmá** na spojnici obou etalonů a tuto spojnici **půlí**

Souvislost jednotlivých principů klasifikace - shrnutí

- Hranice mezi klasifikačními třídami jsou dány průmětem diskriminačních funkcí do obrazového prostoru.
- Klasifikace podle minimální vzdálenosti definuje hranici, která je kolmá na spojnici etalonů klasifikačních tříd a půlí ji.
- Princip klasifikace dle minimální vzdálenosti vede buď přímo, nebo prostřednictvím využití metrik podobnosti k definici diskriminačních funkcí a ty dle prvního ze zde uvedených pravidel k určení hranic mezi klasifikačními třídami.

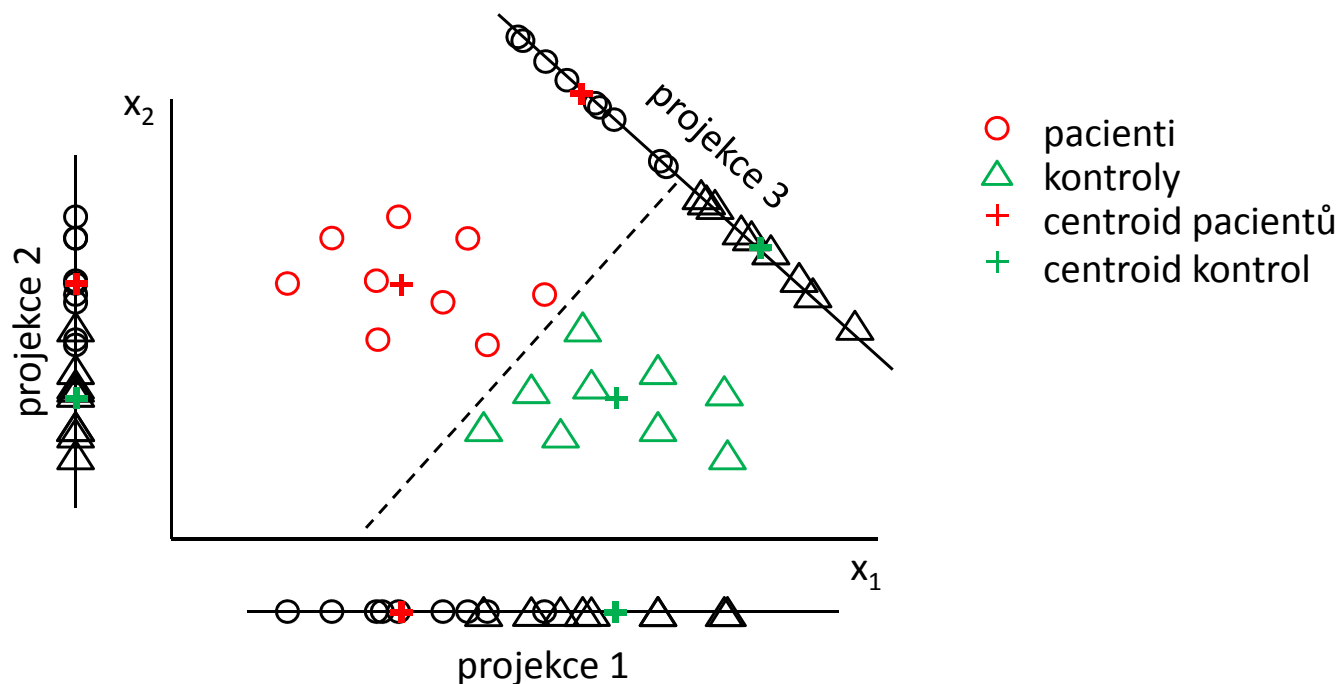
Fisherova lineární diskriminace

- jiný název: Fisherova lineární diskriminační analýza (FLDA)
- použití pro lineární klasifikaci
- princip: transformace do jednorozměrného prostoru tak, aby se třídy od sebe maximálně oddělily



- předpoklad: vícerozměrné normální rozdělení u jednotlivých skupin

Fisherova lineární diskriminace – princip



- podstatou FLDA tedy projekce do 1-D prostoru tak, že chceme:

- maximalizovat vzdálenost skupin
- minimalizovat variabilitu uvnitř skupin

- Fisherovo diskriminační kritérium je tedy ve tvaru: $J(\mathbf{w}) = \frac{(\bar{y}_D - \bar{y}_H)^2}{s_D^2 + s_H^2}$

kde s_D^2 a s_H^2 jsou rozptyly uvnitř třídy pacientů resp. kontrol po projekci do 1-D prostoru a \bar{y}_D a \bar{y}_H jsou projekce centroidu třídy pacientů resp. kontrol

Fisherovo diskriminační kritérium – úpravy, výpočet

- Fisherovo diskriminační kritérium: $J(\mathbf{w}) = \frac{(\bar{y}_D - \bar{y}_H)^2}{s_D^2 + s_H^2}$
- Fisher. disk. kritérium lze rovněž vyjádřit jako: $J(\mathbf{w}) = \frac{(\bar{y}_D - \bar{y}_H)^2}{s_D^2 + s_H^2} = \frac{\mathbf{w}^T \mathbf{S}_B \mathbf{w}}{\mathbf{w}^T \mathbf{S}_W \mathbf{w}}$, kde:
 - \mathbf{S}_B je suma čtverců variability mezi skupinami
 - \mathbf{S}_W je suma čtverců variability uvnitř skupin
 - \mathbf{w} je váhový vektor udávající směr 1-D prostoru, do něhož promítáme
- z čehož po úpravách vypočteme váhový vektor \mathbf{w} jako: $\mathbf{w} \sim \mathbf{S}_W^{-1}(\bar{\mathbf{x}}_D - \bar{\mathbf{x}}_H)$
- hranice je pak dána: $\mathbf{w}^T \mathbf{x} - \tilde{y} = 0$, kde \tilde{y} je průmět hraničního bodu v 1-D prostoru a lze ho vypočítat jako: $\tilde{y} = \frac{\bar{y}_D + \bar{y}_H}{2}$
- pokud chceme zařadit nový subjekt \mathbf{x}_0 do jedné z daných tříd, jeho průmět do 1-D prostoru ($y_0 = \mathbf{w}^T \mathbf{x}_0$) srovnáme s průmětem hraničního bodu \tilde{y} :
 - Pokud $y_0 < \tilde{y}$ (příčemž $\bar{y}_H < \tilde{y}$), subjekt zařadíme do skupiny kontrolních subjektů
 - Pokud $y_0 > \tilde{y}$ (příčemž $\bar{y}_H < \tilde{y}$), subjekt zařadíme do skupiny pacientů

Souvislost lineární diskriminační analýzy s logistickou regresí

- stejně jako lineární diskriminační analýzu lze i logistickou regresí použít pro zařazení objektů/subjektů do hodnocených skupin
- hlavním cílem logistické regrese je ale identifikace vztahů mezi spojitými či binárními prediktory a binárním endpointem (výskyt onemocnění, úmrtí, komplikace atd.) a jejich popis pomocí poměru šancí (odds ratio)
- logistická regrese patří do skupiny zobecněných lineárních modelů
- výstupy logistické regrese:

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	64,211 ^a	,525	,700

a. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

Hosmer and Lemeshow Test

Step	Chi-square	df	Sig.
1	6,832	8	,555

Classification Table^a

		Predicted		
		VERSICOL		Percentage Correct
Observed	,00000000	1,00000000		
Step 1	VERSICOL ,00000000	45	5	90,0
	1,00000000	6	44	88,0
Overall Percentage				89,0

a. The cut value is .500

Poděkování

Příprava výukových materiálů předmětu „DSAN02 Pokročilé metody analýzy dat v neurovědách“ byla finančně podporována prostředky projektu FRMU č. MUNI/FR/0260/2014 „Pokročilé metody analýzy dat v neurovědách jako nový předmět na LF MU“

