ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

III. PŘÍZNAKOVÁ KLASIFIKACE - ÚVOD

PŘÍZNAKOVÝ POPIS

Příznakový obraz **x** zpracovávaných dat je vyjádřen n-rozměrným (sloupcovým) vektorem hodnot x_i, i=1,2,…,n příznakových proměnných (veličin) charakterizujících vlastnosti těchto dat, tj. platí

$$\mathbf{x}=(x_1,x_2,\ldots,x_n)^T.$$

PŘÍZNAKOVÝ POPIS

Příznakové proměnné mohou popisovat kvantitativní i kvalitativní vlastnosti souboru dat. Jejich hodnoty nazýváme příznaky.

Podle definičního oboru rozlišujeme proměnné:

- spojité
- nespojité, diskrétní, vyjmenovatelné
- logické, binární, alternativní, dichotomické

PŘÍZNAKOVÝ POPIS

Vrchol každého příznakového vektoru (obrazu) představuje bod n-rozměrného prostoru $\mathcal{X}^n$, který nazýváme obrazovým prostorem.

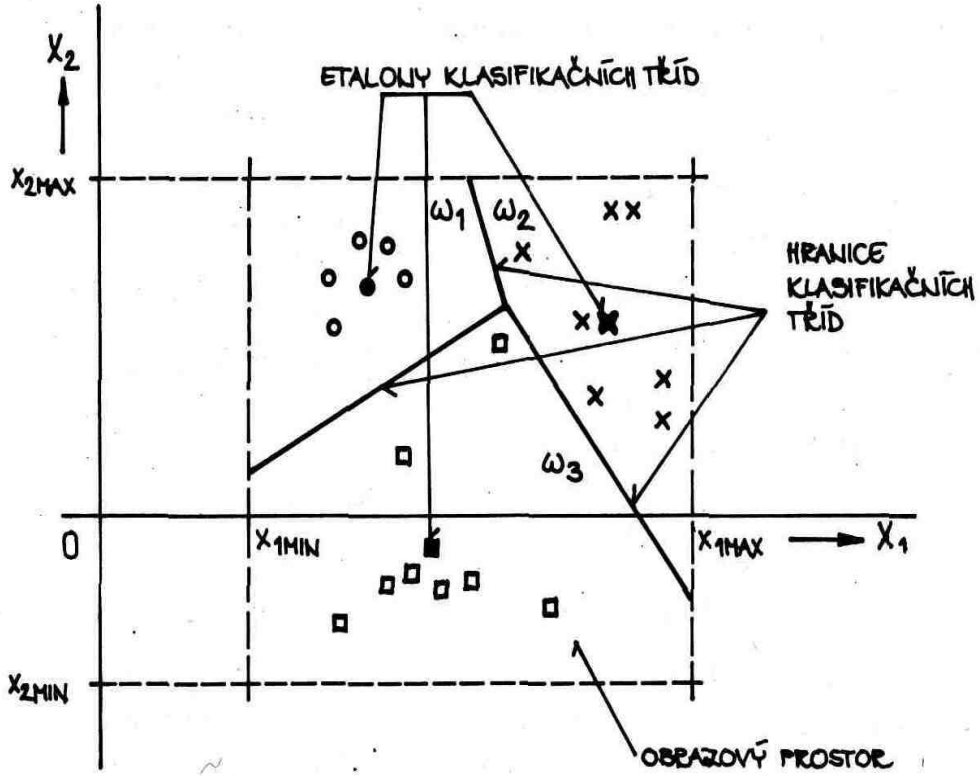
Obrazový prostor je definován kartézským součinem definičních oborů všech příznakovým proměnných, tzn. že jej tvoří všechny možné obrazy zpracovávaného souboru dat.

PŘÍZNAKOVÝ POPIS

Při vhodném výběru příznakových veličin je podobnost signálů jedné klasifikační třídy vyjádřena blízkostí jejich obrazů v obrazovém prostoru.

Vymezení klasifikační třídy:

- etalony - charakteristické reprezentativní obrazy
- hranice
- diskriminační funkce



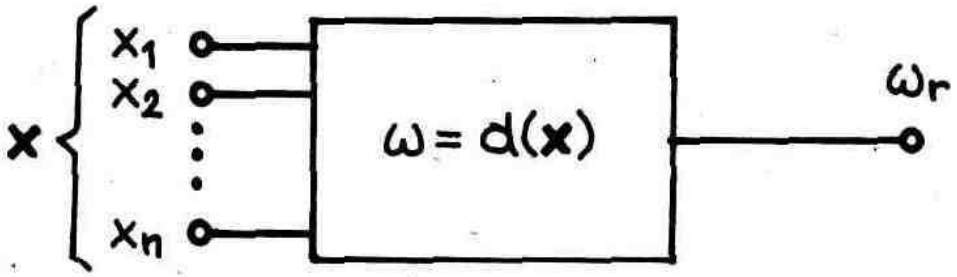
PŘÍZNAKOVÝ KLASIFIKÁTOR

Příznakový klasifikátor je stroj s tolika vstupy, kolik je příznaků a s jedním diskrétním výstupem, který udává třídu, do které klasifikátor zařadil rozpoznávaný obraz.

$$\omega_r = d(\mathbf{x})$$

$d(\mathbf{x})$ je skalární funkce vektorového argumentu $\mathbf{x}$, kterou nazýváme rozhodovací pravidlo klasifikátoru;

ω_r je identifikátor klasifikační třídy



PŘÍZNAKOVÝ KLASIFIKÁTOR

- deterministický a nedeterministický
- s pevným a proměnným počtem příznaků
- bez učení a s učením

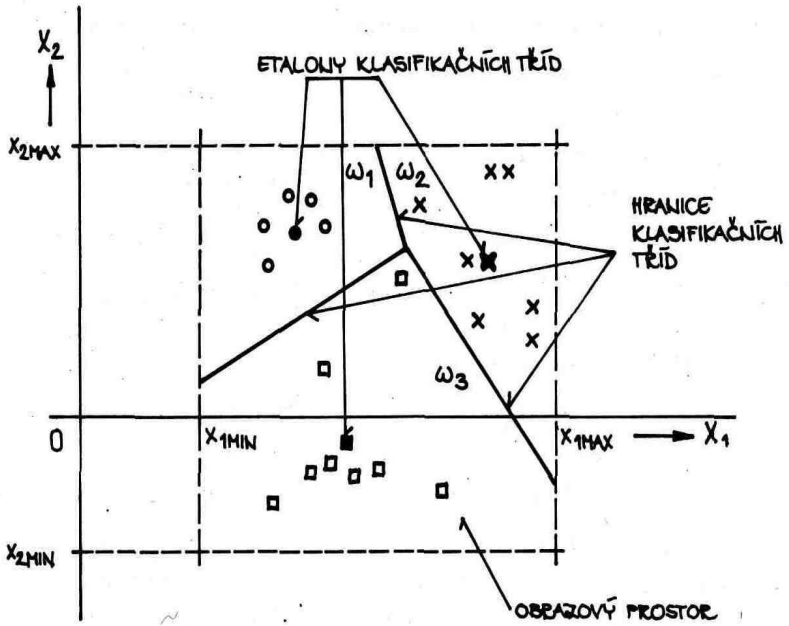
PŘÍZNAKOVÝ KLASIFIKÁTOR

- ☑ deterministický a nedeterministický
- ☑ s pevným a proměnným počtem příznaků
- ☑ bez učení a s učením

Nadále se nějaký čas věnujme deterministickým klasifikátorům s pevným počtem příznaků.

PŘÍZNAKOVÝ KLASIFIKÁTOR

- Obrazový prostor je rozhodovacím pravidlem rozdělen na R disjunktních prostorů $\mathcal{R}_r$, r=1,…,R, přičemž každá podmnožina $\mathcal{R}_r$ obsahuje ty obrazy **x**, pro které je $\omega_r = d(\mathbf{x})$.
- Návrh rozhodovacího pravidla je základním problémem návrhu klasifikátoru.



KLASIFIKACE PODLE DISKRIMINAČNÍCH FUNKCÍ

DISKRIMINAČNÍ ANALÝZA

týká se obecně vztahu mezi kategoriální proměnnou a množinou vzájemně vázaných příznakových proměnných.

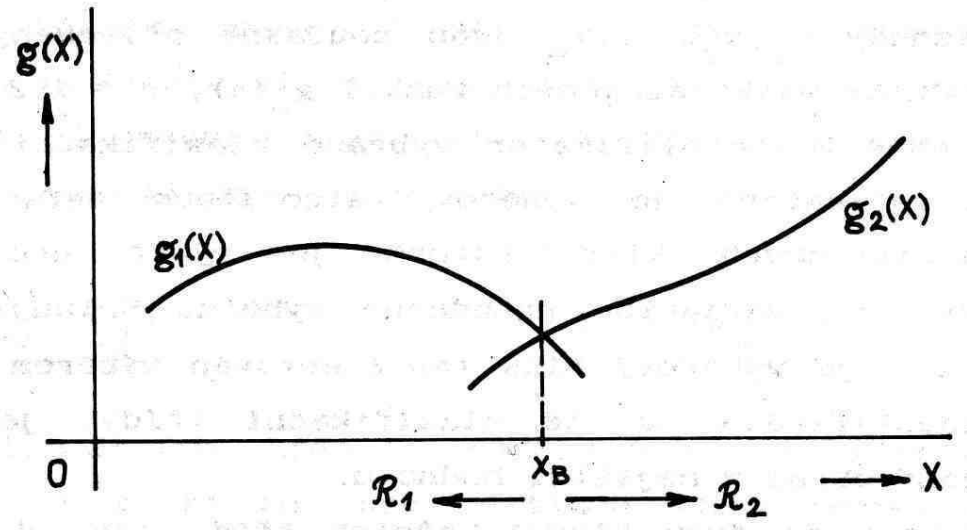
Konkrétně, předpokládejme že existuje konečný počet, řekněme R, různých a priori známých populací, kategorií, tříd nebo skupin, které označujeme ω_r, r=1,…,R a úkolem diskriminační analýzy je nalézt vztah, na základě kterého pro daný vektor příznaků popisujících konkrétní objekt tomuto vektoru přiřadíme hodnotu ω_r.

KLASIFIKACE PODLE DISKRIMINAČNÍCH FUNKCÍ

- hranice klasifikačních tříd definujeme pomocí R skalárních funkcí $g_1(\mathbf{x})$, $g_2(\mathbf{x})$,…, $g_R(\mathbf{x})$ takových , že pro obraz **x** z podmnožiny $\mathcal{R}_r$ pro všechna r platí

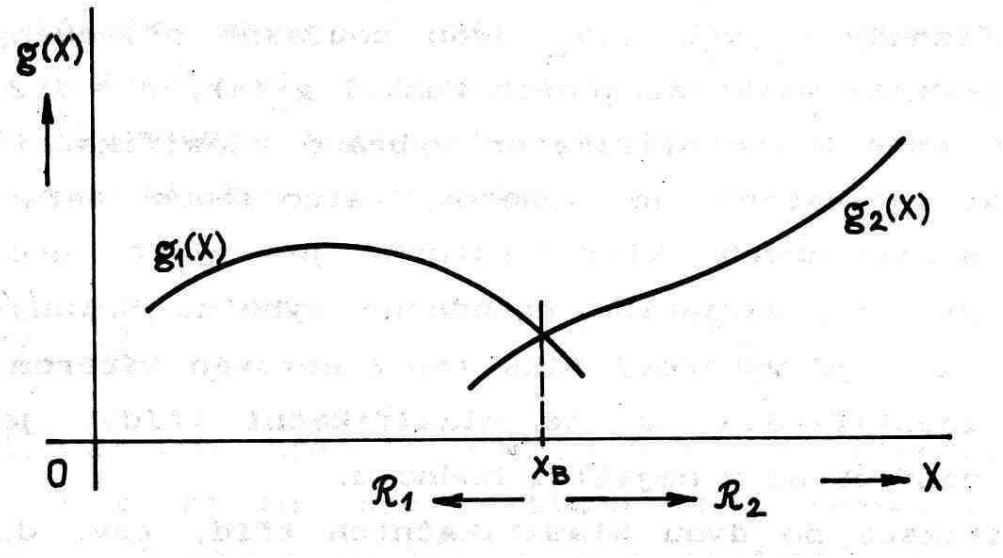
$$g_r(\mathbf{x}) > g_s(\mathbf{x}), \text{ pro } s = 1,2,\ldots,R \text{ a } r \neq s$$

- funkce $g_r(\mathbf{x})$ mohou vyjadřovat např. míru výskytu obrazu **x** patřícího do r-té klasifikační třídy v daném místě obrazového prostoru – nazýváme je **diskriminační funkce**

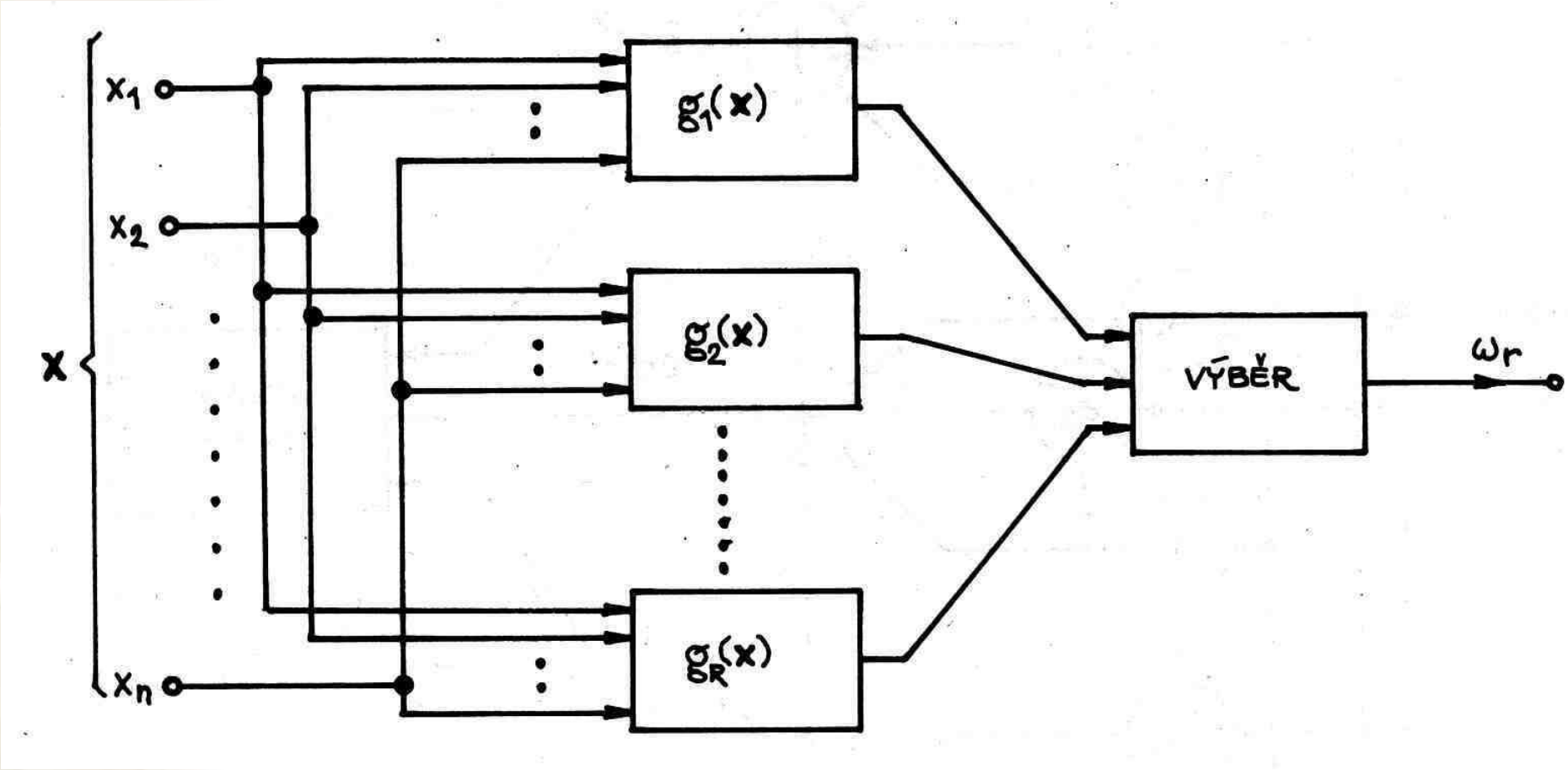


KLASIFIKACE PODLE DISKRIMINAČNÍCH FUNKCÍ

- hranice mezi dvěma sousedními podmnožinami $\mathcal{R}_r$ a $\mathcal{R}_s$ 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.



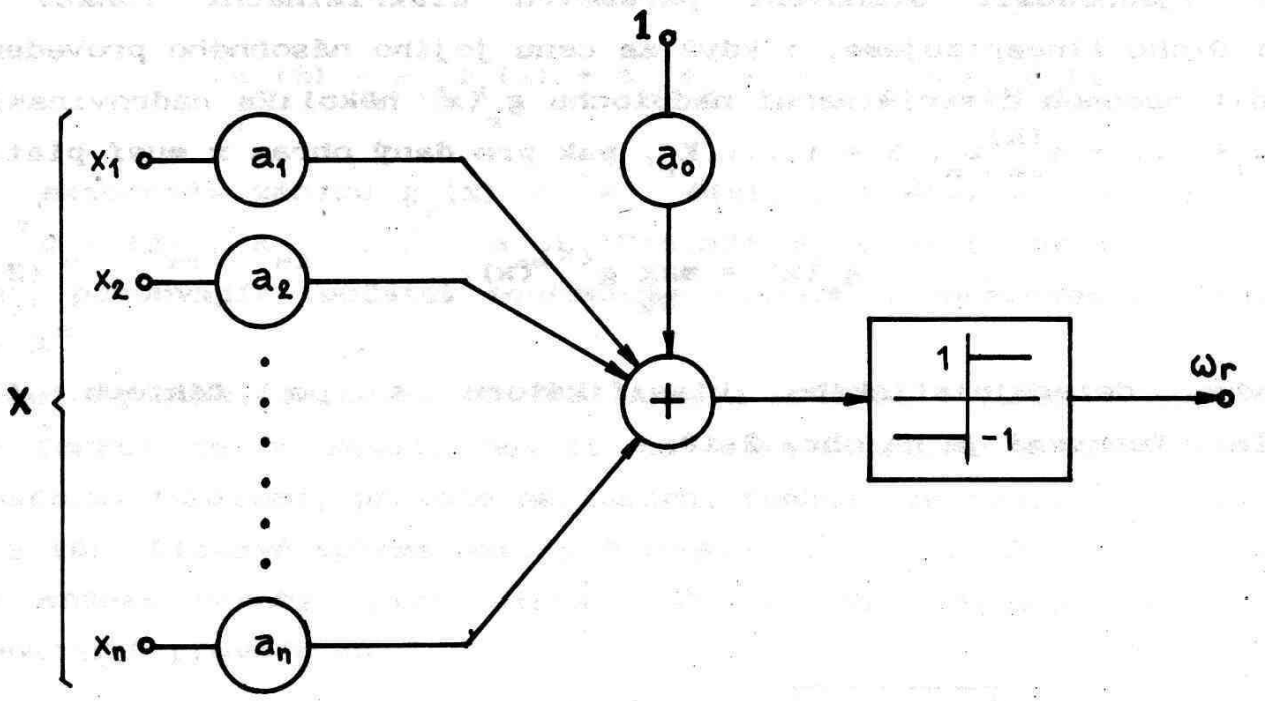
BLOKOVÉ SCHÉMA KLASIFIKÁTORU POMOCÍ DISKRIMINAČNÍCH FUNKCÍ



BLOKOVÉ SCHÉMA KLASIFIKÁTORU POMOCÍ DISKRIMINAČNÍCH FUNKCÍ

- u dichotomického klasifikátoru (dvě třídy) je

$$\omega = \text{sign}\,(g_1(\mathbf{x}) - g_2(\mathbf{x}))$$



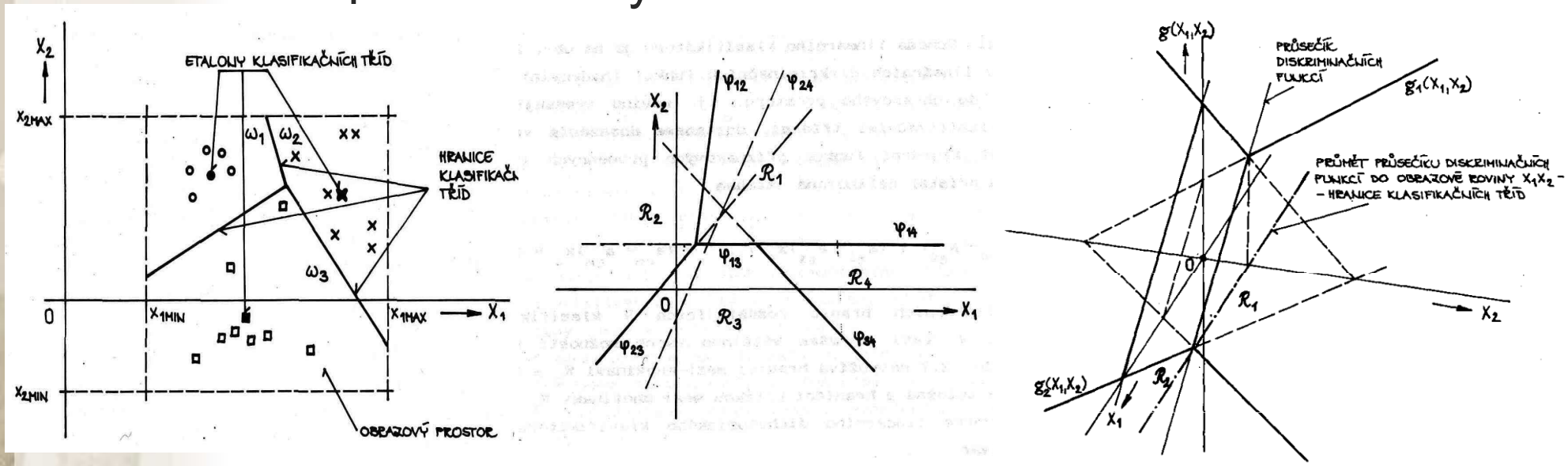
KLASIFIKACE PODLE DISKRIMINAČNÍCH FUNKCÍ

- nejjednodušším tvarem diskriminační funkce je funkce lineární, která má tvar

$$g_r(\mathbf{x}) = a_{r0} + a_{r1}x_1 + a_{r2}x_2 + \ldots + a_{rn}x_n$$

kde a_{r0} je práh diskriminační funkce posouvající počátek souřadného systému a a_{ri} jsou váhové koeficienty i-tého příznaku x_i

- lineárně separabilní třídy



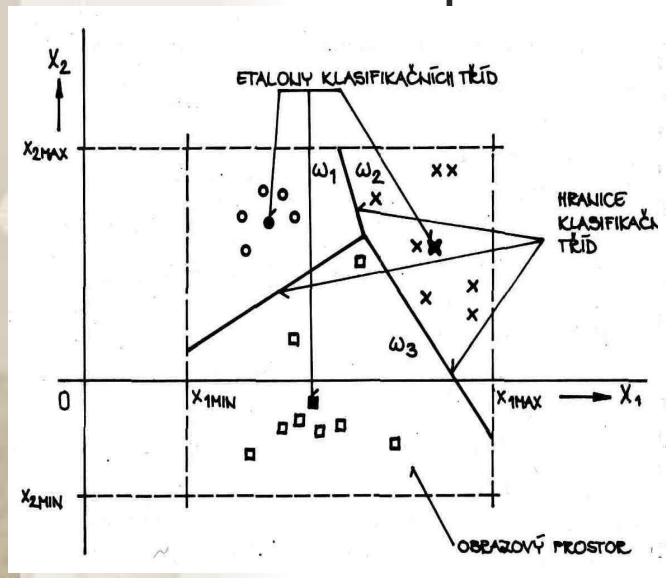
KLASIFIKACE PODLE DISKRIMINAČNÍCH FUNKCÍ

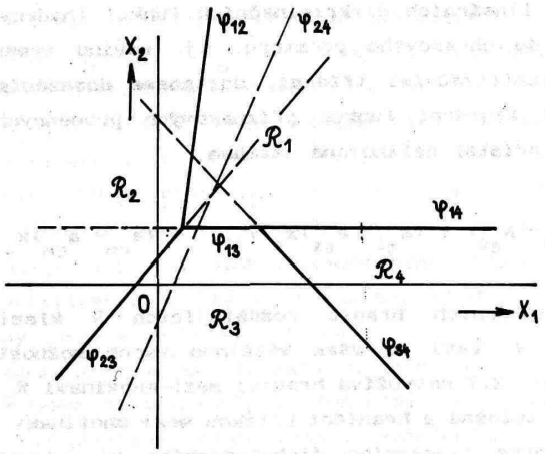
- nejjednodušším tvarem diskriminační funkce je funkce lineární, která má tvar

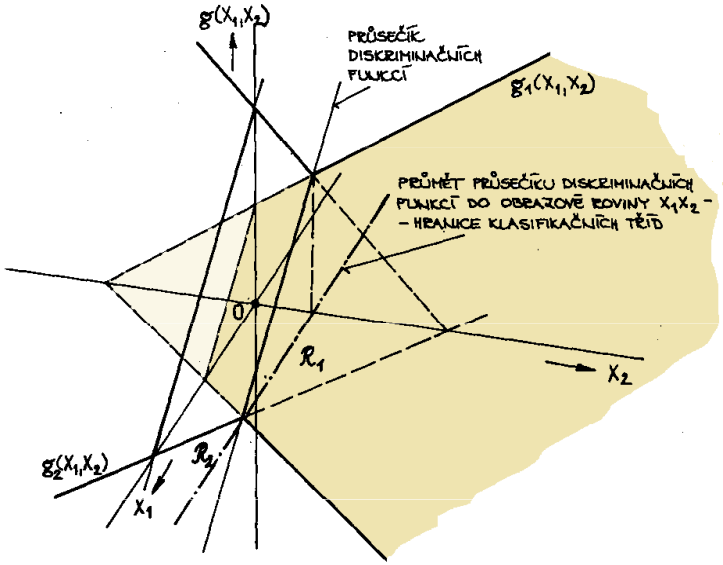
$$g_r(\mathbf{x}) = a_{r0} + a_{r1}x_1 + a_{r2}x_2 + \ldots + a_{rn}x_n$$

kde a_{r0} je práh diskriminační funkce posouvající počátek souřadného systému a a_{ri} jsou váhové koeficienty i-tého příznaku x_i

- lineárně separabilní třídy







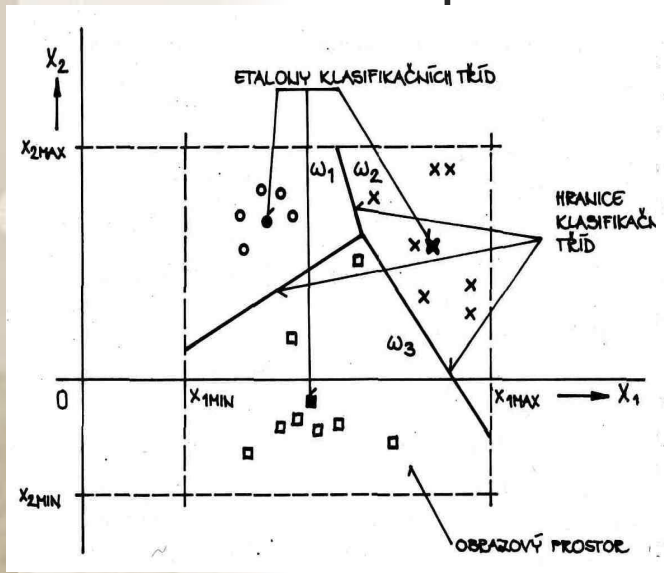
KLASIFIKACE PODLE DISKRIMINAČNÍCH FUNKCÍ

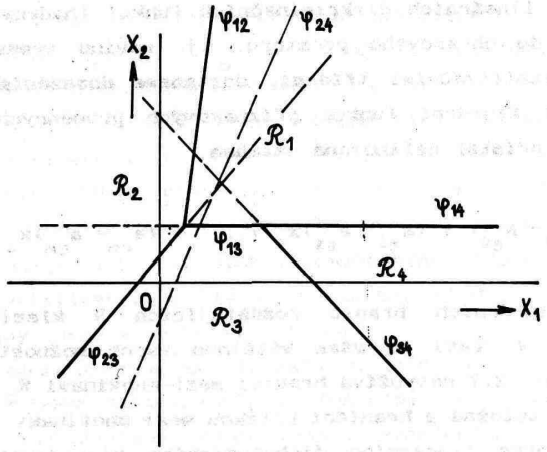
- nejjednodušším tvarem diskriminační funkce je funkce lineární, která má tvar

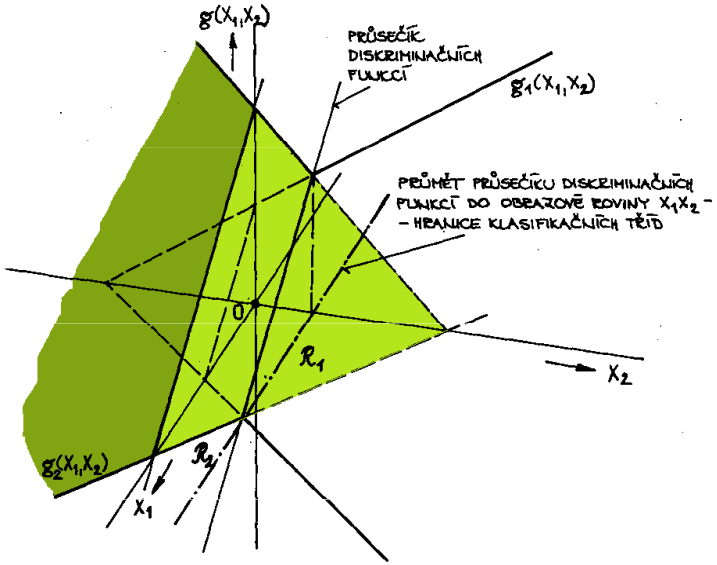
$$g_r(\mathbf{x}) = a_{r0} + a_{r1}x_1 + a_{r2}x_2 + \ldots + a_{rn}x_n$$

kde a_{r0} je práh diskriminační funkce posouvající počátek souřadného systému a a_{ri} jsou váhové koeficienty i-tého příznaku x_i

- lineárně separabilní třídy







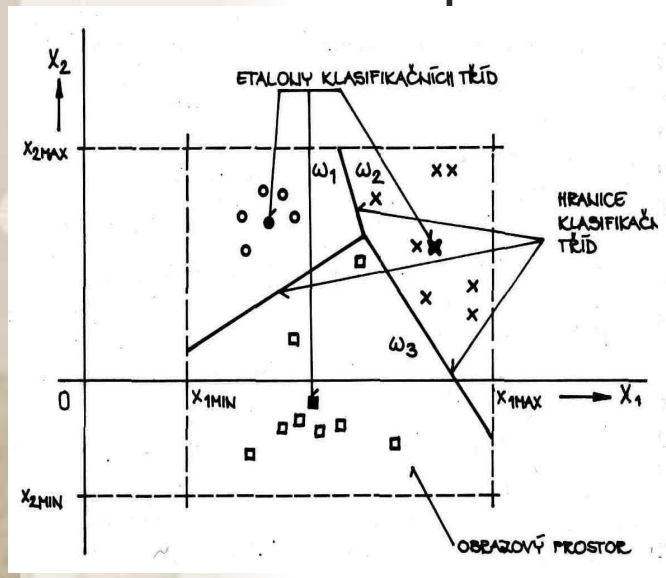
KLASIFIKACE PODLE DISKRIMINAČNÍCH FUNKCÍ

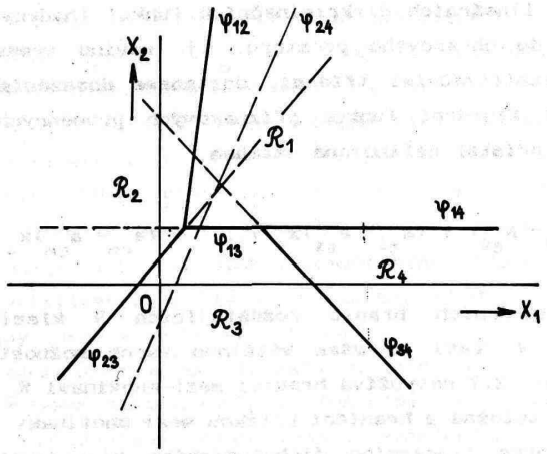
- nejjednodušším tvarem diskriminační funkce je funkce lineární, která má tvar

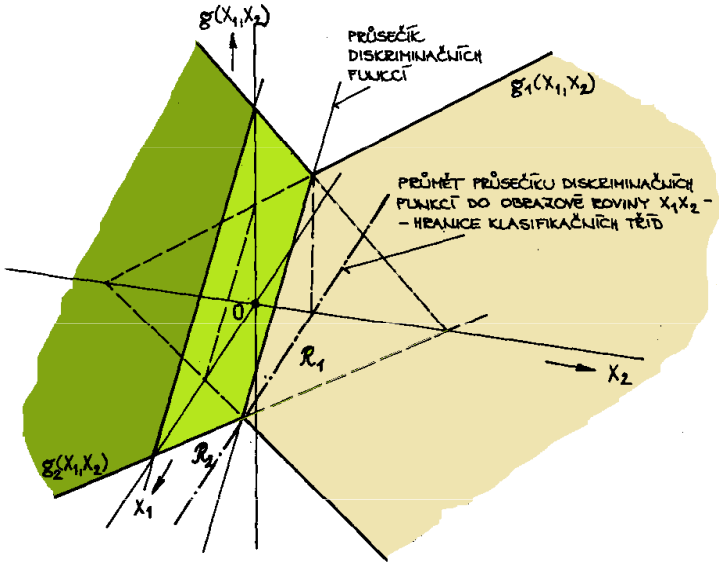
$$g_r(\mathbf{x}) = a_{r0} + a_{r1}x_1 + a_{r2}x_2 + \ldots + a_{rn}x_n$$

kde a_{r0} je práh diskriminační funkce posouvající počátek souřadného systému a a_{ri} jsou váhové koeficienty i-tého příznaku x_i

- lineárně separabilní třídy



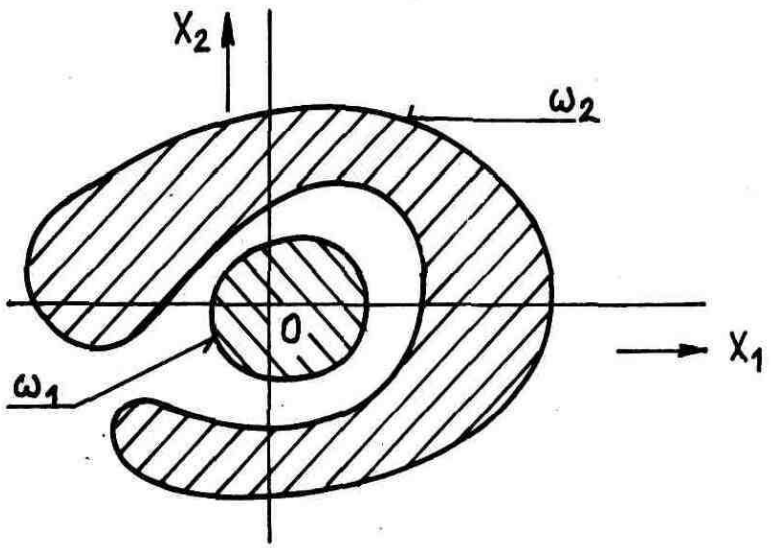




KLASIFIKACE PODLE DISKRIMINAČNÍCH FUNKCÍ

LINEÁRNĚ NESEPARABILNÍ TŘÍDY

- zachováme původní obrazový prostor a zvolíme nelineární diskriminační funkci
  - definovanou obecně
  - složenou po částech z lineárních úseků



- zobrazíme původní n-rozměrný obrazový prostor X^n nelineární transformací $\Phi: X^n \rightarrow X^m$ do nového m-rozměrného prostoru X^m, obecně je $m \neq n$, tak, aby v novém prostoru byly klasifikační třídy lineárně separabilní a v novém prostoru použijeme lineární klasifikátor (Φ převodník)

KLASIFIKACE PODLE MINIMÁLNÍ VZDÁLENOSTI

- reprezentativní obrazy klasifikačních tříd - etalony
- je-li v obrazovém prostoru zadáno R poloh etalonů vektory $\mathbf{x}_{1E}, \mathbf{x}_{2E}, \ldots, \mathbf{x}_{RE}$, zařadí klasifikátor podle minimální vzdálenosti klasifikovaný obraz $\mathbf{x}$ do té třídy, jejíž etalon má od bodu $\mathbf{x}$ minimální vzdálenost. Rozhodovací pravidlo je určeno vztahem

$$d(\mathbf{x}) = \left\|\mathbf{x}_{rE} - \mathbf{x}\right\| = \min_{\forall s} \left\|\mathbf{x}_{sE} - \mathbf{x}\right\|$$

KLASIFIKACE PODLE MINIMÁLNÍ VZDÁLENOSTI

- uvažme případ dvou tříd reprezentovaných etalony $\mathbf{x}_{1E} = (x_{11E}, x_{12E})$ a $\mathbf{x}_{2E} = (x_{21E}, x_{22E})$ ve dvoupříznakovém euklidovském prostoru;
- vzdálenost mezi obrazem $\mathbf{x} = (x_1, x_2)$ a libovolným z obou etalonů je pak definována

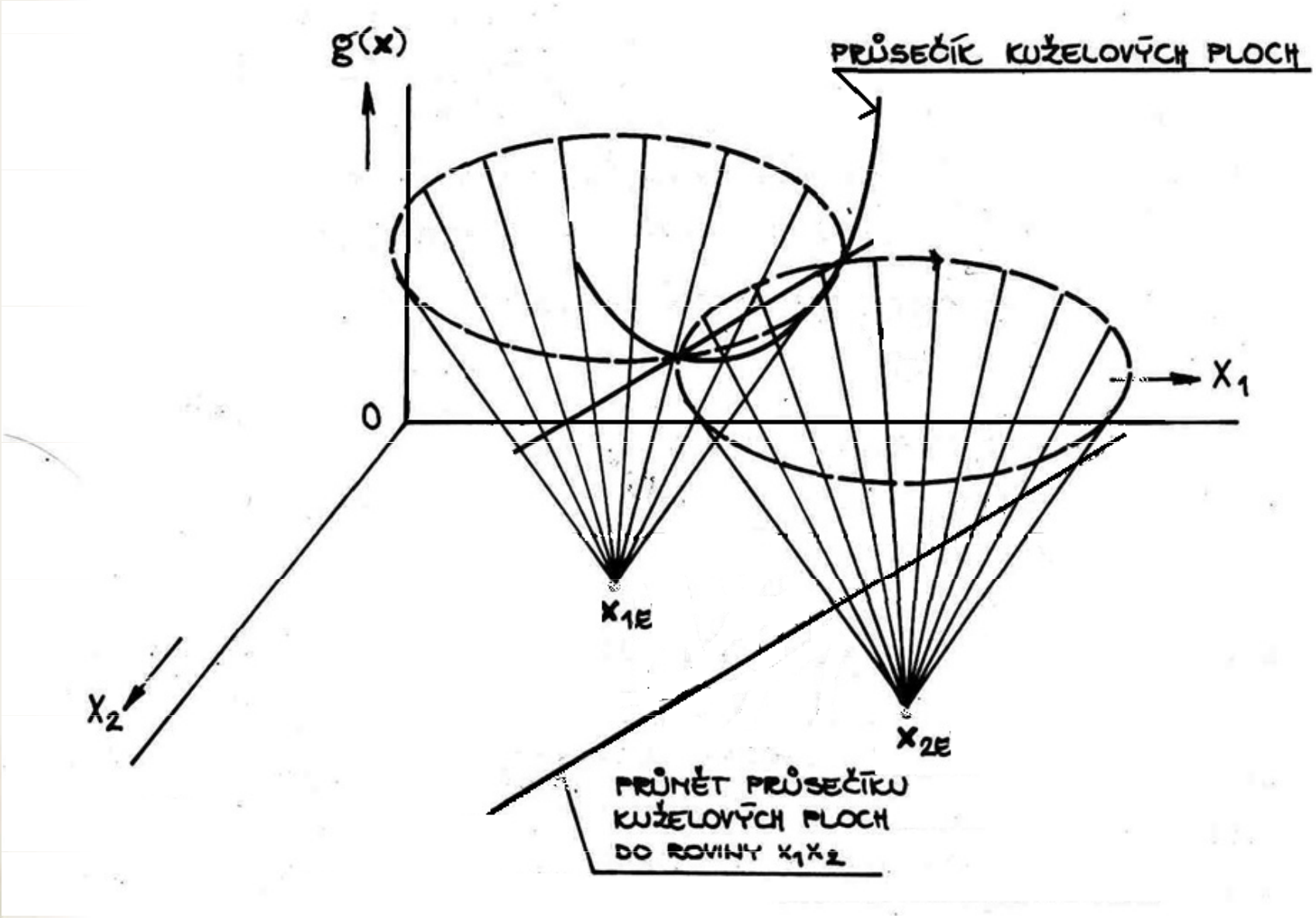
$$v(\mathbf{x}_{sE}, \mathbf{x}) = \left\| \mathbf{x}_{sE} - \mathbf{x} \right\| = \sqrt{(x_{s1E} - x_1)^2 + (x_{s2E} - x_2)^2}$$

- hledáme menší z obou vzdáleností, tj. $\min_{s=1,2} v(\mathbf{x}_{sE}, \mathbf{x})$, ale také $\min_{s=1,2} v^2(\mathbf{x}_{sE}, \mathbf{x})$;

$$\min_{\forall s} v(\mathbf{x}_{sE}, \mathbf{x}) \approx \min_{\forall s} v^2(\mathbf{x}_{sE}, \mathbf{x}) = \min_{\forall s} \left((x_{s1E} - x_1)^2 + (x_{s1E} - x_1)^2 \right) =$$

$$\min_{\forall s} \left(x_1^2 + x_2^2 - 2[x_{s1E} x_1 + x_{s2E} x_2 - (x_{s1E}^2 + x_{s2E}^2)/2] \right)$$

KLASIFIKACE PODLE MINIMÁLNÍ VZDÁLENOSTI



KLASIFIKACE PODLE MINIMÁLNÍ VZDÁLENOSTI

- diskriminační kuželové plochy se protínají v parabole a její průmět do obrazové roviny je přímka definovaná vztahem

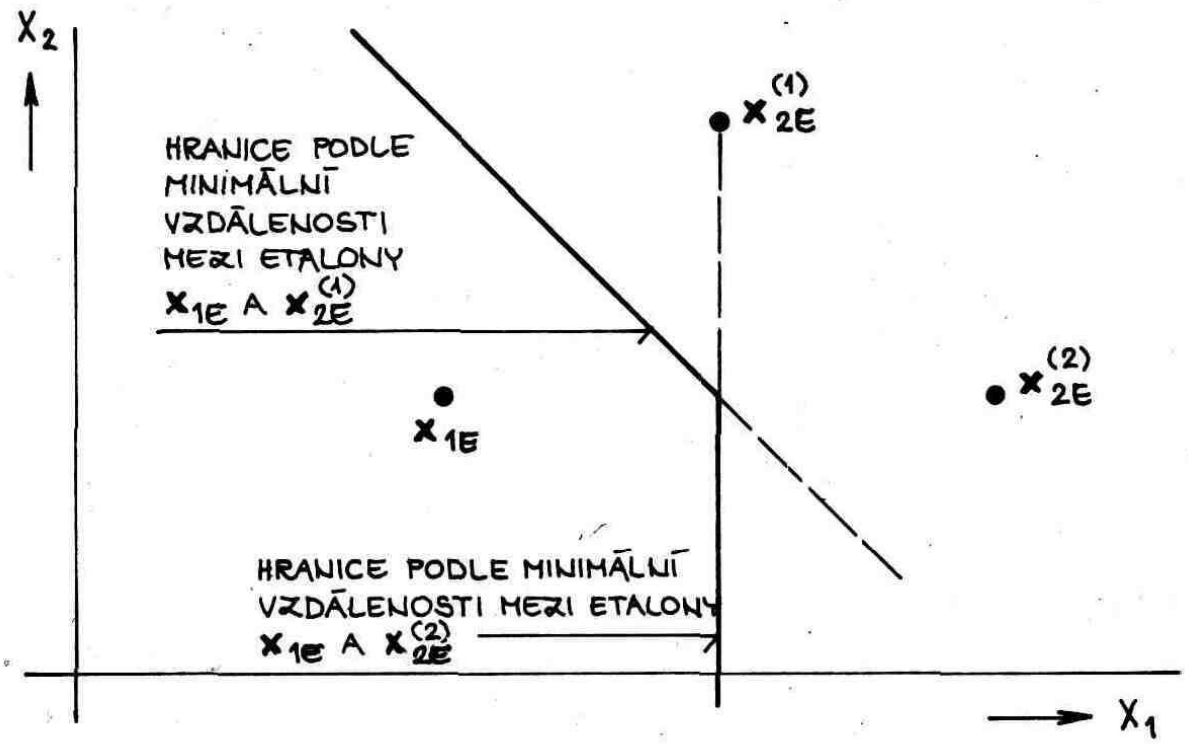
$$x_1(x_{11E} - x_{21E}) + x_2(x_{12E} - x_{22E}) - (x^2_{12E} + x^2_{11E} - x^2_{21E} - x^2_{22E})/2 = 0$$

Tato hraniční přímka mezi klasifikačními třídami je vždy kolmá na spojnici obou etalonů a tuto spojnici půlí

$$\Downarrow$$

klasifikátor pracující na základě kritéria minimální vzdálenosti je ekvivalentní lineárnímu klasifikátoru s diskriminačními funkcemi.

KLASIFIKACE PODLE MINIMÁLNÍ VZDÁLENOSTI



- Klasifikace podle minimální vzdálenosti s třídami reprezentovanými více etalony je ekvivalentní klasifikaci podle diskriminační funkce s po částech lineární hraniční plochou