7 Štatistické (príznakové) metódy

- podrobnejšie popíšeme tri nástroje rozpoznávania, ktoré už čiastočne poznáme, a to diskriminačné funkcie, pravidlo najbližšieho suseda (kritérium minimálnej vzdialenosti) a kritérium minimálnej chyby

Diskriminačné funkcie

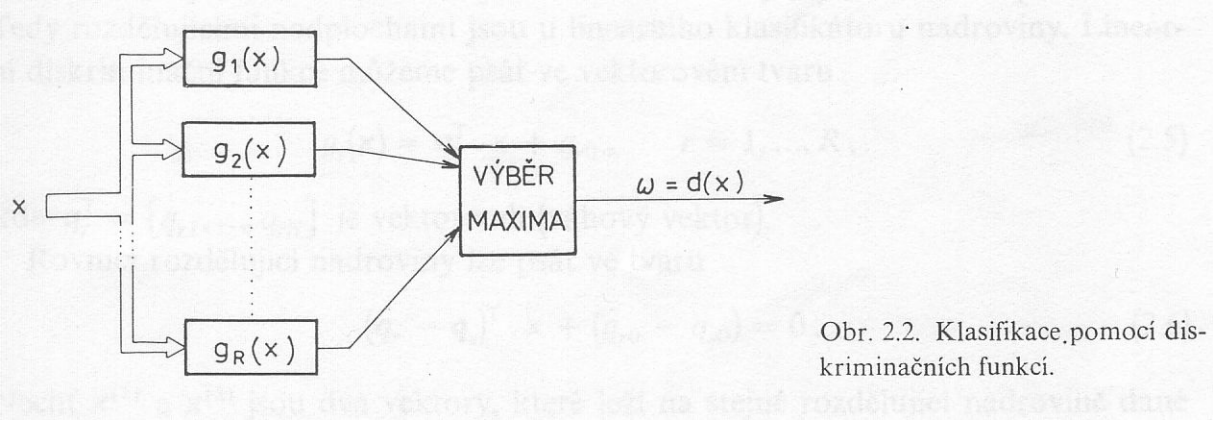
- majme príznakový vektor $\mathbf{x}^{\mathbf{T}} = (x_{1,}, x_2, ..., x_N)$, pričom množina všetkých vektorov tvorí príznakový priestor. Majme *R* vzájomne disjunktných tried, ktoré je možné oddeliť rozdeľujúcimi nadplochami – nadplochy sa dajú definovať pomocou sústavy *R* skalárnych funkcií vektorového argumentu $g_r(\mathbf{x}),\ r = 1, ..., \mathrm{R},$ ktoré sa nazývajú diskriminačné funkcie. Rozhodovacie pravidlo

  $d(\mathbf{x}) = \omega_r \Leftrightarrow g_r(\mathbf{x}) > g_s(\mathbf{x})$, pre každý vektor **x** a každé *s* rozdielne od *r*.

- Pre prvky rozdeľujúcej nadplochy nie je rozhodovacie pravidlo definované. Rovnice rozdeľujúcich nadplôch medzi susednými triedami majú tvar

  $g_r(\mathbf{x}) = g_s(\mathbf{x})$

  a klasifikátor je na obrázku.



Obr. 2.2. Klasifikace pomocí diskriminačních funkcí.

- pre prípad dvoch tried stačí zobrazovať znamienko rozdielu diskriminačných funkcií $\omega = \mathrm{sign}(g_1(\mathbf{x}) - g_2(\mathbf{x}))$. V najjednoduchšom prípade sú diskriminačné funkcie lineárne tvaru

$$g_r(\mathbf{x}) = \sum_{j=1}^{N} q_{rj} x_j + q_{r0}, \qquad r = 1,2,....,R \qquad ,$$

kde koeficient q_{rj} sa nazýva váhou *j*-teho príznaku a q_{r0} sa nazýva prahom. Rozdeľujúca nadplocha medzi dvoma susednými množinami, je definovaná rovnicou

$$g_r(\mathbf{x}) - g_s(\mathbf{x}) = \sum_{j=1}^{N} (q_{rj} - q_{sj}) + (q_{r0} - q_{s0}) = 0\,,$$

ktorá je tiež lineárna a je rovnicou nadroviny v príznakovom priestore. Lineárne diskriminačné funkcie možno napísať vo vektorom tvare

$$g_r(\mathbf{x}) = \mathbf{q}_r^T . \mathbf{x} + q_{r0}, \qquad r = 1,2,....,R\,.$$

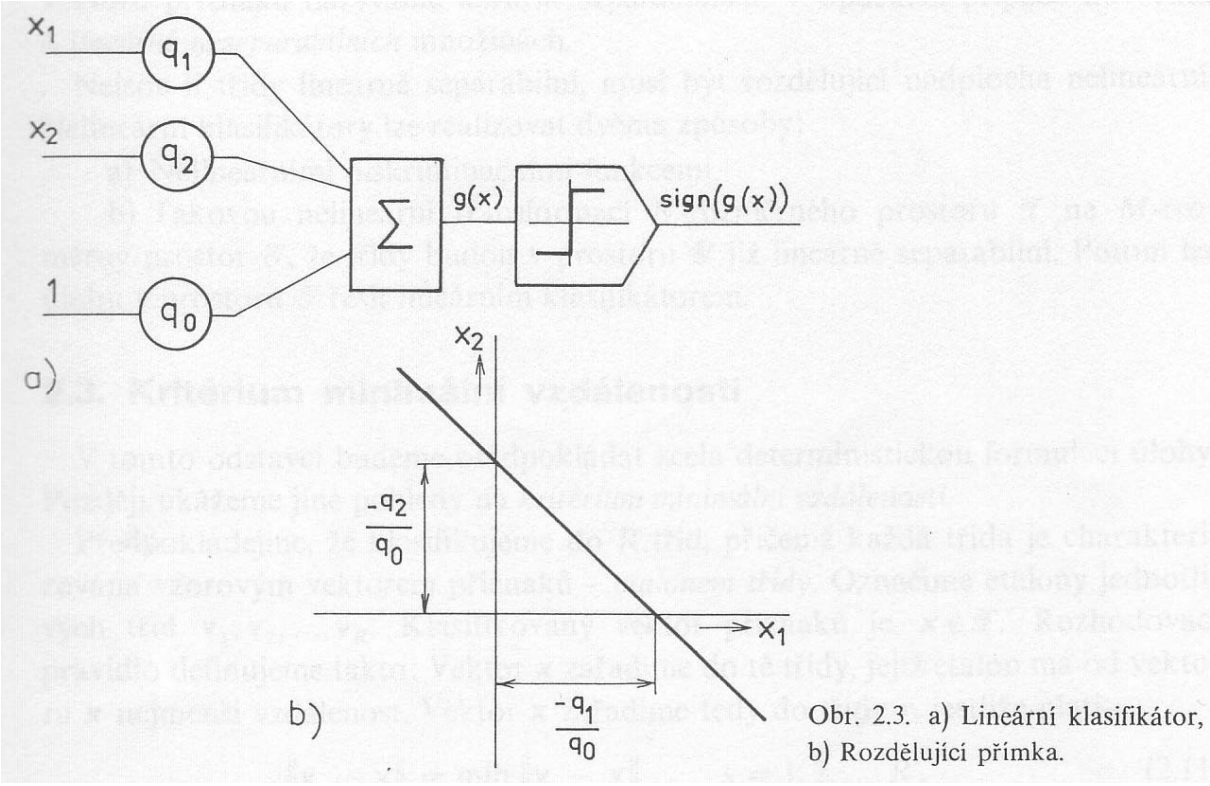
Nech $\mathbf{x}^{(1)}$ a $\mathbf{x}^{(2)}$ sú dva vektory, ktoré ležia na rovnakej rozdeľujúcej nadrovine, potom platí

$$(\mathbf{q}_r - \mathbf{q}_s)^T . (\mathbf{x}^{(1)} - \mathbf{x}^{(2)}) = 0\,.$$

Vektory $(\mathbf{q}_r - \mathbf{q}_s)^T$ a $(\mathbf{x}^{(1)} - \mathbf{x}^{(2)})$ podľa toho sú na seba kolmé, vektor $(\mathbf{x}^{(1)} - \mathbf{x}^{(2)})$ leží v rozdeľujúcej nadrovine a vektor $(\mathbf{q}_r - \mathbf{q}_s)^T$ má smer normály k tejto nadrovine.
Pri dvoch triedach s lineárnym klasifikátorom má diskriminačná funkcia tvar

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x}) = (\mathbf{q}_1 - \mathbf{q}_2)^T . \mathbf{x} - (q_{10} - q_{20}) = \mathbf{q}^T . \mathbf{x} + q_0.$$

Pre $N = 2$ je situácia zobrazená na nasledujúcom obrázku



Obr. 2.3. a) Lineární klasifikátor, b) Rozdělující přímka.

- Pri klasifikácii lineárnym klasifikátorom na *R* tried ukážeme, že triedy príznakového priestoru sú konvexné. Predpokladajme, že $\mathbf{x}_a$ a $\mathbf{x}_b$ sú dva vektory klasifikované do *r*-

tej triedy. Potom aj všetky vektory **x** na úsečke $\overline{\mathbf{x}_a\mathbf{x}_b}$ sú klasifikované do *r*-tej triedy.

Vektor ležiaci na úsečke $\overline{\mathbf{x}_a\mathbf{x}_b}$ vyjadríme parametrickou rovnicou $\mathbf{x} = \mathbf{x}_a + t(\mathbf{x}_b - \mathbf{x}_a)$, kde *t* je parameter z intervalu <0,1>. Vyjadríme diskriminačnú funkciu $g_s(\mathbf{x})$ pre **x** ležiace na $\overline{\mathbf{x}_a\mathbf{x}_b}$ a budeme hľadať pre aké *s* nadobúda maximum. Vzhľadom k linearite funkcie $g_s(\mathbf{x})$ môžeme písať

$$g_s(\mathbf{x}) = g_s(\mathbf{x}_a + t(\mathbf{x}_b - \mathbf{x}_a)) = g_s(\mathbf{x}_a) + t[g_s(\mathbf{x}_b) - g_s(\mathbf{x}_a) = g_s(\mathbf{x}_a)(1-t) + t.g_s(\mathbf{x}_b)\,.$$

Pretože parametre t a $(1 - t)$ sú pre určité **x** konštantné pre všetky diskriminačné funkcie, nadobúda funkcia svoje maximum pre $s = r$, takže vektor **x** bude klasifikovaný do triedy *r*. Tak sme zároveň ukázali, aké podmienky musia spĺňať triedy, ktoré možno *bez chyby* klasifikovať lineárnym klasifikátorom – vektory príznakov, patriacich do tej istej triedy, musia ležať vo vnútri konvexnej množiny. Také množiny nazývame *lineárne separabilné*, v opačnom prípade hovoríme o *lineárne neseparabilných množinách.*

- Ak nie sú triedy lineárne separabilné, musí byť rozdeľujúca nadplocha nelineárna. Nelineárne klasifikátory môžeme realizovať dvoma spôsobmi:
  a) nelineárnymi diskriminačnými funkciami
  b) takou nelineárnou transformáciou *N*-rozmerného priestoru na *M*-rozmerný priestor, že triedy budú v tomto priestore už lineárne separabilné. Takáto transformácia sa nazýva Φ-prevodník.

Pravidlo najbližšieho suseda (kritérium minimálnej vzdialenosti)

- Predpokladajme, že každá trieda je charakterizovaná vzorovým vektorom príznakov – *etalónom triedy,* označeným ako $\mathbf{v}_1, \mathbf{v}_2, ..., \mathbf{v}_R$. Potom rozhodovacie pravidlo vytvoríme takto:

$$d(\mathbf{x}) = \omega_r \Leftrightarrow \|\mathbf{v}_r - \mathbf{x}\| = \min_s \|\mathbf{v}_s - \mathbf{x}\| \qquad s = 1,2,....,R\,,$$

kde vzdialenosť definujeme ako normu $\|\mathbf{v}_r - \mathbf{x}\|$. Ak je norma vektoru **x** definovaná ako

$$\|\mathbf{x}\| = \sqrt{\mathbf{x}^T.\mathbf{x}} = \sqrt{\sum_{j=1}^{N} x_j^2}\,,$$

potom je príznakový priestor *N*-rozmerným Euklidovským priestorom. Minimum vzdialenosti $\|\mathbf{v}_s - \mathbf{x}\|$ nastáva pre rovnaký vektor, pre ktorý nastáva minimum kvadrátu $\|\mathbf{v}_s - \mathbf{x}\|^2$. Môžeme teda minimalizovať výraz

$$\|\mathbf{v}_s - \mathbf{x}\|^2 = (\mathbf{v}_s - \mathbf{x})^T.(\mathbf{v}_s - \mathbf{x}) = \mathbf{v}_s^T.\mathbf{v}_s - 2.\mathbf{v}_s^T.\mathbf{x} + \mathbf{x}^T.\mathbf{x}$$

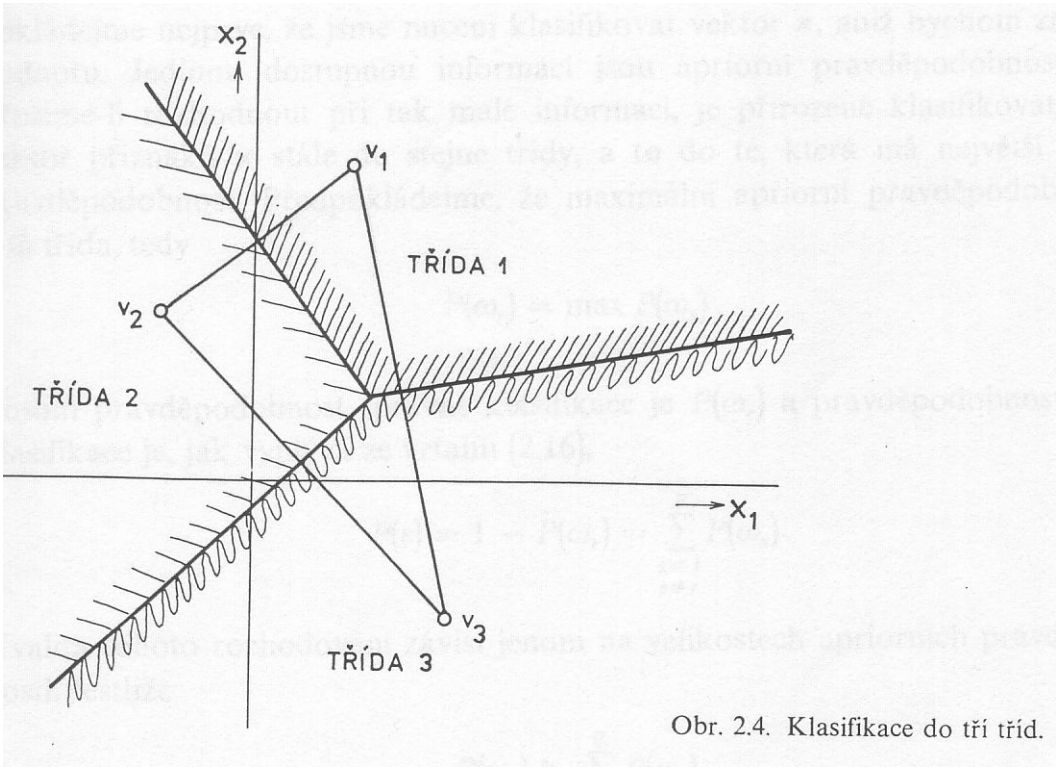
Pre daný vektor príznakov minimalizujeme tento vzťah voľbou $\mathbf{v}_s$. Môžeme teda písať

$$\min_s \|\mathbf{v}_s - \mathbf{x}\|^2 = \mathbf{x}^{\mathrm{T}}.\mathbf{x} - 2.\max_s(\mathbf{v}_s^T.\mathbf{x} - \frac{1}{2}\mathbf{v}_s^T.\mathbf{v}_s).$$

Minimum nastáva pre taký etalón $\mathbf{v}_r$, pre ktorý nadobúda funkcia

$$g_r(\mathbf{x}) = \mathbf{v}_{\mathrm{r}}^{\mathrm{T}}.\mathbf{x} - \frac{1}{2}.\mathbf{v}_r^T.\mathbf{v}_r$$

svoje maximum, pričom táto funkcia je diskriminačnou funkciou *r*-tej triedy. Pretože funkcia je lineárna vzhľadom na **x,** úlohu možno riešiť lineárnym klasifikátorom.



Obr. 2.4. Klasifikace do tří tříd.

Kritérium minimálnej chyby

- existuje veľké množstvo praktických úloh, pri ktorom máme neseparabilné triedy a teda rozhodnutím sa dopúšťame chyby. Vtedy sa snažíme nastaviť klasifikátor tak, aby straty spôsobené chybným rozhodnutím boli minimálne.

- pretože nevieme jednoznačne rozhodnúť o triede, pokladáme ju za náhodnú premennú s možnými hodnotami $\omega_1, \omega_2, ..., \omega_R$ s apriórnymi pravdepodobnosťami $P(\omega_1), P(\omega_2), ..., P(\omega_R)$, pre ktoré platí

$$\sum_{r=1}^{R} P(\omega_r) = 1.$$

- predstavme si, že máme rozhodnúť o zaradení neznámeho vektora **x** bez toho, aby sme poznali jeho hodnotu. Ak zvolíme tú triedu, ktorá má najväčšiu apriórnu pravdepodobnosť, tak potom pravdepodobnosť správnej klasifikácie je $P(\omega_R)$ a pravdepodobnosť nesprávnej klasifikácie je 1 - $P(\omega_R)$. Ak pravdepodobnosť $P(\omega_R)$ je výrazne väčšia ako suma ostatných apriórnych pravdepodobností, potom je toto rozhodnutie väčšinou správne, inak sa dopustíme veľkej chyby.

- väčšinou máme viac informácie ako iba apriórne pravdepodobnosti tried, nech máme aj podmienené pravdepodobnosti $p(\mathbf{x}|\omega_r)$, ktoré vyjadrujú pravdepodobnosť objavenia sa príznakového vektora **x** za podmienky, že je známa trieda ω_R.
- Potom podmienená pravdepodobnosť, že vektor príznakov **x** patrí do triedy ω_R je daná vzťahom

$$P(\omega_r|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_r).P(\omega_r)}{p(\mathbf{x})},$$

kde

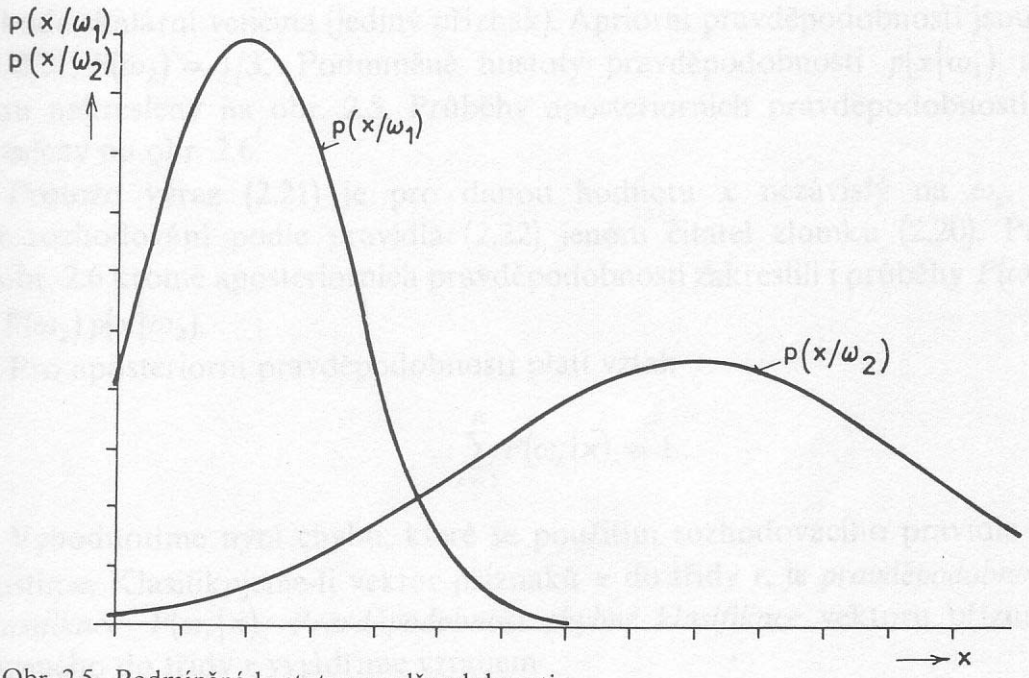
$$p(\mathbf{x}) = \sum_{r=1}^{R} p(\mathbf{x}|\omega_r).P(\omega_r)$$

je hustota pravdepodobnosti rozloženia príznakového vektora v príznakovom priestore bez ohľadu na triedu.

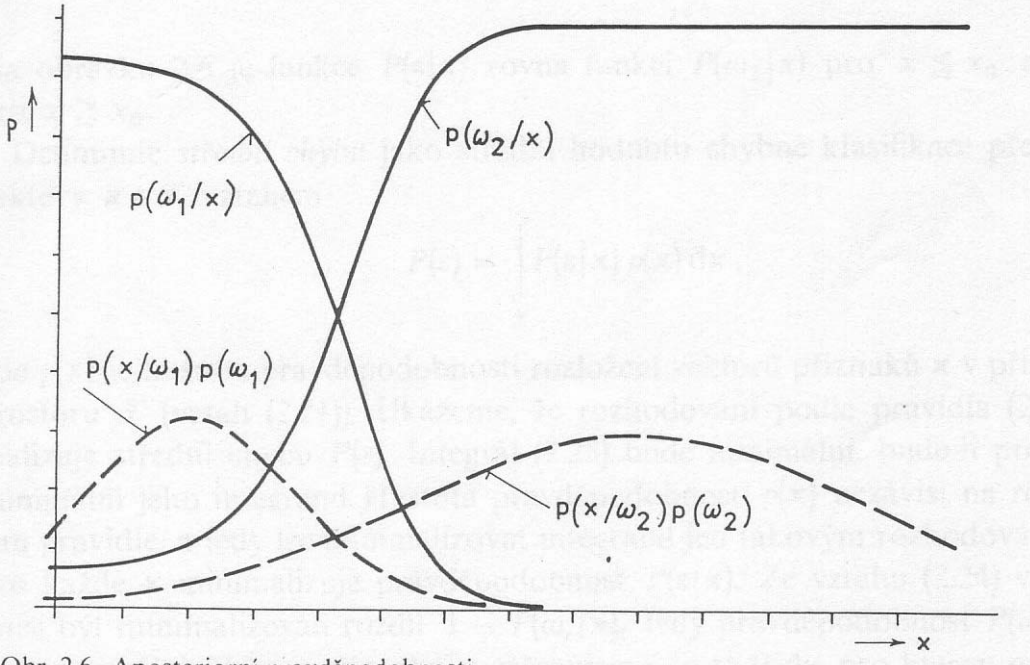
- Pravdepodobnosti $P(\omega_r|\mathbf{x})$ nazývame aposteriórne pravdepodobnosti a ich výpočet nazývame *Bayesovo pravidlo.* Pomocou aposteriórnych pravdepodobností vytvoríme rozhodovacie pravidlo

$$d(\mathbf{x}) = \omega_r \Leftrightarrow P(\omega_r|\mathbf{x}) = \max_s P(\omega_s|\mathbf{x})$$

- Pre ilustráciu ukážme rozhodovacie pravidlo na úlohe klasifikácie do dvoch tried s apriórnymi pravdepodobnosťami tried 2/3 a 1/3 a s podmienenými pravdepodobnosťami ako na obrázku.



Obr. 2.5. Podmíněné hustoty pravděpodobnosti.



Obr. 2.6. Aposteriorní pravděpodobnosti.

- potom pravdepodobnosť chybnej klasifikácie vektoru **x** zaradeného do triedy ω_r je daná vzťahom

$$P(\varepsilon|\mathbf{x}) = 1 - P(\omega_r|\mathbf{x}) = \sum_{\substack{s=1 \\ s\neq r}}^{R} P(\omega_s|\mathbf{x})$$

- Na obrázku je táto pravdepodobnosť rovná funkcii $P(\omega_2|\mathbf{x})$ pre $x \le x_0$ a $P(\omega_1|\mathbf{x})$ pre $x \ge x_0$.

- Definujme strednú chybu klasifikácie pre všetky vektory v príznakovom vektore vzťahom

$$P(\varepsilon) = \int_X P(\varepsilon|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

- Ukážeme, že rozhodovanie podľa zvoleného pravidla $d(\mathbf{x})$ minimalizuje strednú chybu. Integrál bude minimálny, ak bude pre každé **x** minimálny jeho integrand. Hustota pravdepodobnosti $p(\mathbf{x})$ nezávisí na rozhodovacom pravidle, a teda integrand možno minimalizovať takým pravidlom, ktoré minimalizuje $P(\varepsilon|\mathbf{x})$ pre každé **x**. Treba maximalizovať pravdepodobnosť $P(\omega_r|\mathbf{x})$, ktorá vystupuje v definícii $P(\varepsilon|\mathbf{x})$. Vektor **x** zaraďujeme do tej triedy, pre ktorú má $P(\omega_r|\mathbf{x})$ najväčšiu hodnotu. Ak budú všetky aposteriórne pravdepodobnosti rovné $1/R$, potom bude $P(\varepsilon) = (R-1)/R$ a táto hodnota bude najväčšia.

- Kritérium minimálnej chyby možno zovšeobecniť, v tom zmysle, že ohodnotíme aj váhu nesprávneho rozhodnutia, ktorú sme doteraz brali pre všetky rozhodnutia ako rovnakú. Majme množinu tried a majme množinu rozhodnutí $d_1, d_2, ..., d_s$. Označme $\lambda(\omega_r, d_s)$ stratu, ktorú utrpíme, ak vektor **x** patrí do triedy ω_r a zvolíme rozhodnutie d_s. Definujme podmienenú strednú hodnotu straty výrazom

$$\Re(d_s|\mathbf{x}) = \sum_{r=1}^{R} \lambda(\omega_r, d_s) P(\omega_r|\mathbf{x}),$$

pričom táto veličina sa niekedy nazýva podmieneným rizikom. Pretože množina rozhodnutí d sa vo všeobecnosti zhoduje s množinou indikátorov tried, upravíme definíciu rozhodovacieho pravidla.

- Potom podmienené riziko môžeme vyjadriť ako $\Re(d(\mathbf{x})|\mathbf{x})$.
- Stredné riziko $\Re$ určíme ako strednú hodnotu podmieneného rizika $\Re(d(\mathbf{x})|\mathbf{x})$ v celom príznakovom priestore

$$\Re = \int_X \Re(d(\mathbf{x})|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

- Podobnou úvahou ako v prípade kritéria minimálnej chyby dospejeme k tomu, že minimum stredného rizika dosiahneme tak, že pre každý vektor príznakov

minimalizujeme podmienené riziko $\Re(d(\mathbf{x})|\mathbf{x})$. Výsledné minimálne stredné riziko sa nazýva Bayesovské riziko a je najlepším výsledkom, aký je možné dosiahnuť.

- Rôznou voľbou stratovej funkcie dostávame rôzny tvar rozhodovacieho pravidla ***d(x)***, ktoré minimalizuje stredné riziko $\Re$. Uvažujme *R* rozhodnutí $d_1, d_2, ..., d_R$, kde rozhodnutie d_r znamená, že vektor **x** zaraďujeme do triedy ω_r. Potom definujme *jednotkovú stratovú funkciu* takto:

$$\lambda(\omega_r, d_s) = \lambda(\omega_r, \omega_s) = 0 \qquad \text{pre } r = s$$
$$\lambda(\omega_r, \omega_s) = 1 \qquad \text{pre } r \neq s$$

to znamená, že správna klasifikácia spôsobuje nulovú stratu a nesprávna jednotkovú stratu. V takom prípade je podmienené riziko

$$\Re(\omega_s|\mathbf{x}) = \sum_{r=1}^{R} \lambda(\omega_s, \omega_r) P(\omega_r|\mathbf{x}) = \sum_{\substack{r=1 \\ r \neq s}}^{R} P(\omega_r|\mathbf{x}) = 1 - P(\omega_s|\mathbf{x})$$

- potom minimalizácia podmieneného rizika podľa tohto vzorca vedie na Bayesovské rozhodovacie pravidlo, z čoho vyplýva že toto pravidlo (alebo kritérium minimálnej chyby je špeciálnym prípadom kritéria minimalizácie stredného rizika, ktoré dostaneme voľbou jednotkovej stratovej funkcie.

- Rozhodovacie pravidlo podľa kritéria minimálnej chyby môžeme vyjadriť aj pomocou diskriminačných funkcií, ich výber však nie je jednoznačný. Môžeme zvoliť ľubovoľnú z nasledovných funkcií:

$$g_r(\mathbf{x}) = P(\omega_r|\mathbf{x}) = \frac{\mathrm{p}(\mathbf{x}|\omega_r) P(\omega_r)}{\sum_{s=1}^{R} \mathrm{p}(\mathbf{x}|\omega_s) P(\omega_s)}$$
$$g_r(\mathbf{x}) = p(\mathbf{x}|\omega_r) P(\omega_r)$$
$$g_r(\mathbf{x}) = \log\ \mathrm{p}(\mathbf{x}|\omega_r) + \log P(\omega_r)$$

- Ak obrázok uvedený vyššie budeme chápať ako klasifikáciu podľa diskriminačných funkcií, plné čiary môžeme chápať ako diskriminačnú funkciu podľa prvej definície, čiarkované podľa druhej.

- Uvažujme, že výskyt vektorov **x** v triedach sa riadi normálnym rozdelením. Hustota pravdepodobnosti je potom daná vzťahom:

$$p(\mathbf{x}|\omega_r) = \frac{1}{(2\pi)^{\mathrm{N}/2} |\eta|^{1/2}} \exp\left[-\frac{1}{2} (\mathbf{x} - \mu_r)^T . \eta_r^{-1} . (\mathbf{x} - \mu_r) \right]$$

kde μ_r je vektor stredných hodnôt vektora príznakov **x** patriaceho do *r*-tej triedy a $|\eta|$ je diskriminant kovariančnej matice η_r.

- Uvažujme najjednoduchší prípad, keď kovariančná matica je pre všetky triedy rovnaká a naviac platí

$$\eta_r = \sigma^2 \mathbf{E}$$

kde **E** je jednotková matica rozmeru *N* x *N* a σ^2 je rozptyl.

- Tento prípad zodpovedá štatisticky nezávislým príznakom, ktoré majú všetky rovnakú disperziu. Ak použijeme diskriminačnú funkciu podľa tretej definície, dostaneme

$$g_r(\mathbf{x}) = \frac{-(\mathbf{x}-\mu_r)^T.(\mathbf{x}-\mu_r)}{2\sigma^2} - \frac{N}{2}\log 2\pi - \frac{1}{2}\log(\sigma^{2N}) + \log P(\omega_r)$$

- druhý a tretí člen sú konštanty, takže môžeme všetky diskriminačné funkcie o tieto konštanty zväčšiť. Potom nová diskriminačná funkcia má tvar

$$g_r'(\mathbf{x}) = \frac{-\|\mathbf{x}-\mu_r\|^2}{2\sigma^2} + \log P(\omega_r)$$

Ak sú apriórne pravdepodobnosti $P(\omega_r)$ rovnaké pre všetky triedy, možno použiť ako diskriminačnú funkciu výraz $-\|\mathbf{x}-\mu_r\|^2$, ktorá je maximálna, ak vzdialenosť $\|\mathbf{x}-\mu_r\|$ je minimálna, čo v tomto zvláštnom prípade vedie na rozhodovanie podľa pravidla najbližšieho suseda (alebo kritéria minimálnej vzdialenosti). Úlohu riešime lineárnym klasifikátorom.

- Ak apriórne pravdepodobnosti $P(\omega_r)$ nie sú pre všetky triedy rovnaké, musíme ich zobrať do úvahy, potom úpravou vzťahu pre diskriminačnú funkciu dostaneme

$$g_r'(\mathbf{x}) = \frac{1}{2\sigma^2}(\mathbf{x}^T.\mathbf{x} - 2\mu_r^T.\mathbf{x} + \mu_r^T.\mu_r) + \log P(\omega_r)$$

- Z toho vytvoríme novú diskriminačnú funkciu

$$g_r''(\mathbf{x}) = \frac{1}{\sigma^2}\mu_r^T.\mathbf{x} - \frac{1}{\sigma^2}\mu_r^T.\mu_r + \log P(\omega_r) = \mathbf{q}_r^T.\mathbf{x} + q_{r0}$$

z čoho vyplýva, že úloha je opäť riešiteľná lineárnym klasifikátorom.

- Dá sa ukázať, že ak kovariančná matica má všeobecný tvar, ale je rovnaká pre všetky triedy, tak úloha opäť vedie na klasifikáciu s lineárnym klasifikátorom.
  V najvšeobecnejšom prípade, keď sú kovariančné matice celkom ľubovoľné, sú rozdeľujúce nadplochy kvadratické.