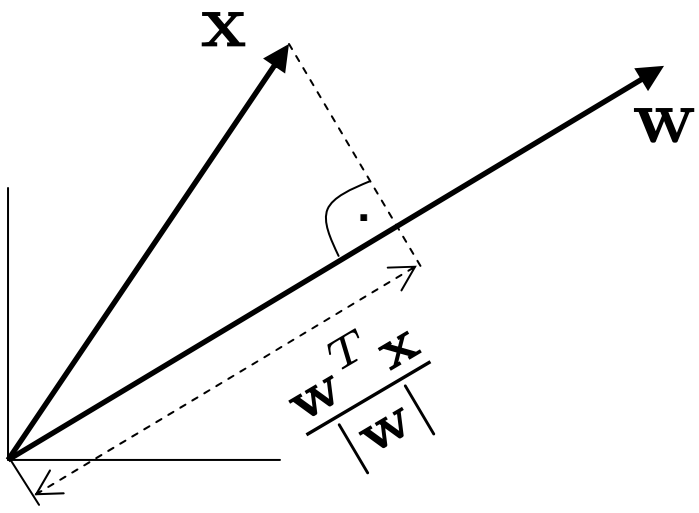
Klasifikace a rozpoznávání

Lineární klasifikátory

Opakování - Skalární součin

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$$

$$\mathbf{w}^T\mathbf{x} = \begin{bmatrix} w_1 & w_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = w_1x_1 + w_2x_2$$



Lineární klasifikátor

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$

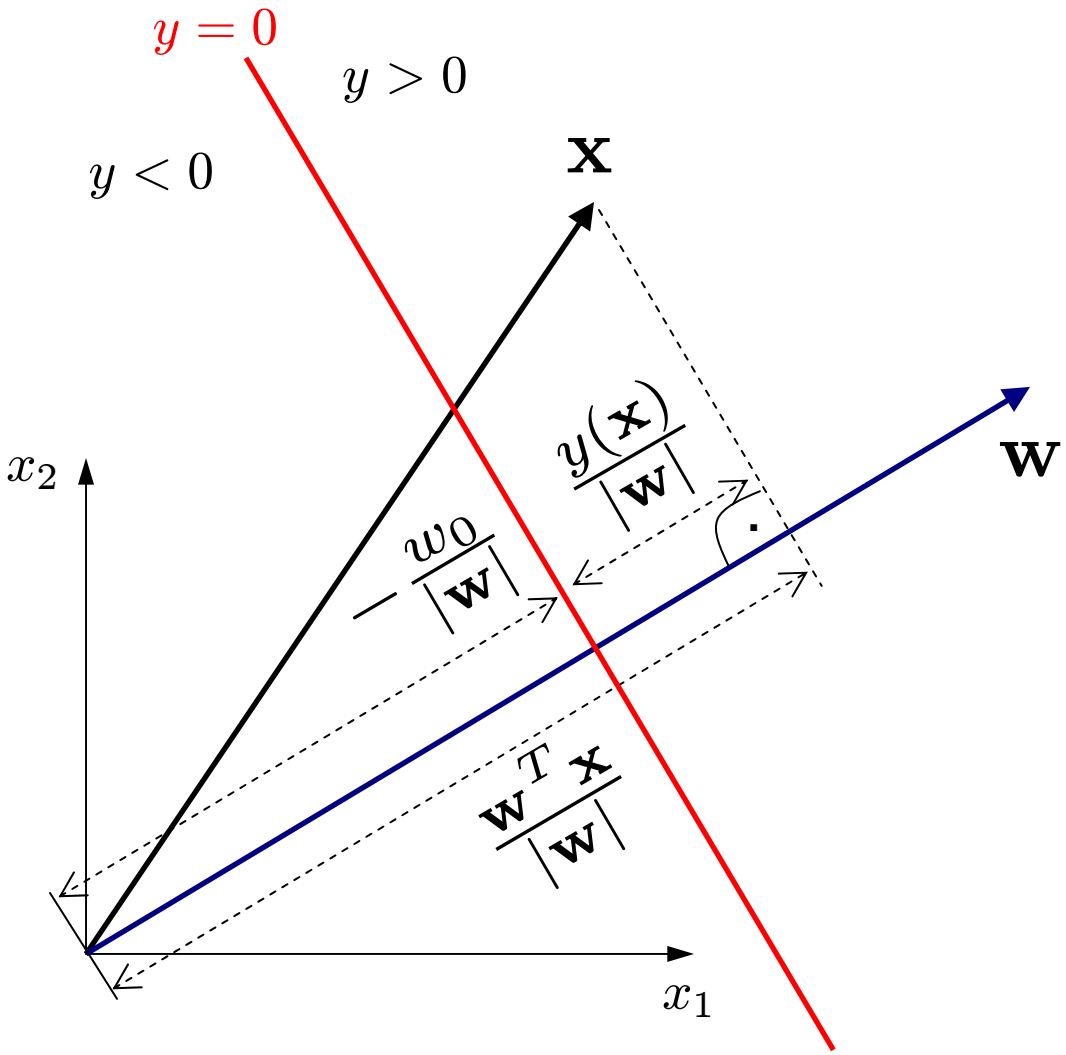
Vyber třídu C_1 pokud y(**x**) > 0 a jinak vyber třídu C_2

Zobecněný lineární klasifikátor

$$y(\mathbf{x}) = f(\mathbf{w}^T\mathbf{x} + w_0)$$

kde f se nazývá aktivační funkce

Lineární klasifikátor



$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$

$$\frac{\mathbf{w}^T\mathbf{x}}{|\mathbf{w}|} = -\frac{w_0}{|\mathbf{w}|} + \frac{y(\mathbf{x})}{|\mathbf{w}|}$$

Perceptron

- Jednoduchý lineární klasifikátor s aktivační funkcí:

$$f(a) = \begin{cases} +1, & a \geqslant 0 \\ -1, & a < 0 \end{cases}$$

- Samotná aktivační funkce v tomto případě nic nezmění – rozhodování na základě y(**x**) > 0 by vedlo ke stejnému výsledku – ale pro učící se algoritmus bude výhodné definovat si požadovaný výstup jako:

$$t \in \{-1, +1\}$$

- Pro další zjednodušení předpokládejme, ze w_0 je “nultý” koeficient vektoru **w** a odpovídající vstup x_0 je vždy 1. Můžeme tedy psát pouze:

$$y(\mathbf{x}) = f(\mathbf{w}^T \mathbf{x})$$

Perceptron – učící algoritmus

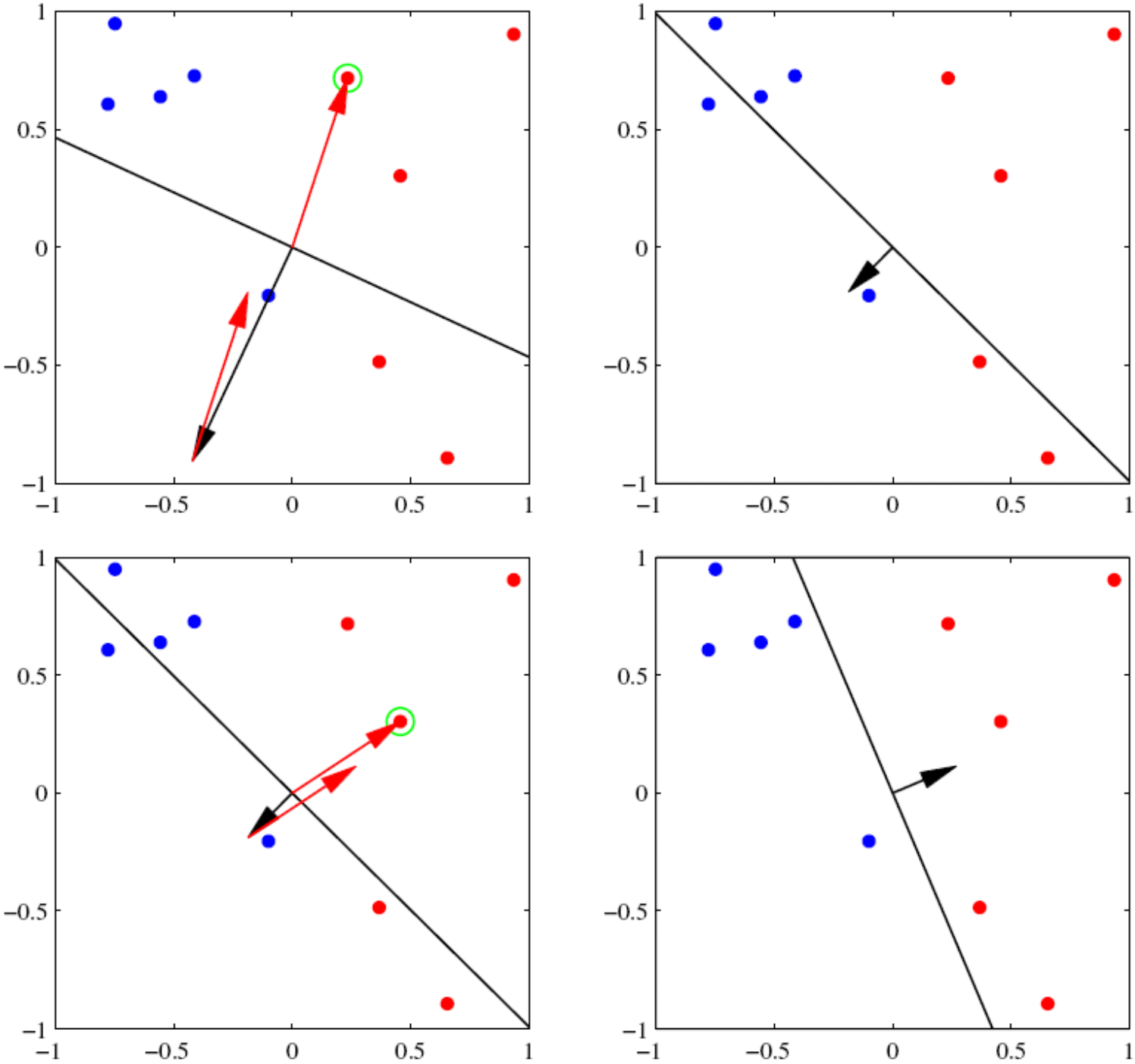
- Cyklicky procházej jednotlivé trénovací vzory a vždy když narazíš na špatně klasifikovaný vzor kde

$$y(\mathbf{x}_n) \neq t_n$$

změň vektor **w** takto:

$$\mathbf{w}^{\tau+1} = \mathbf{w}^{\tau} + \mathbf{x}_n t_n$$

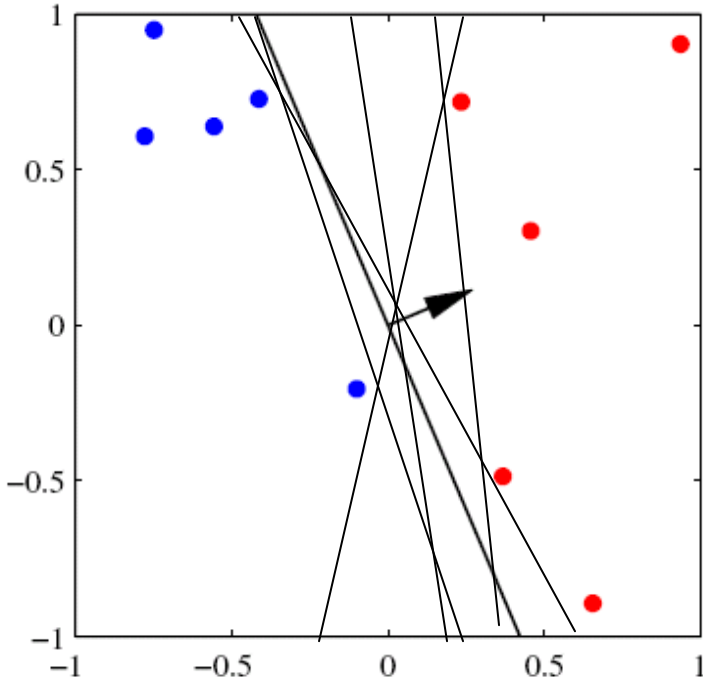
- Lze dokázat, že pokud jsou data lineárně separovatelná, tak, algoritmus vždy nalezne řešení – konverguje. V opačném případě, ale nikdy nekonverguje

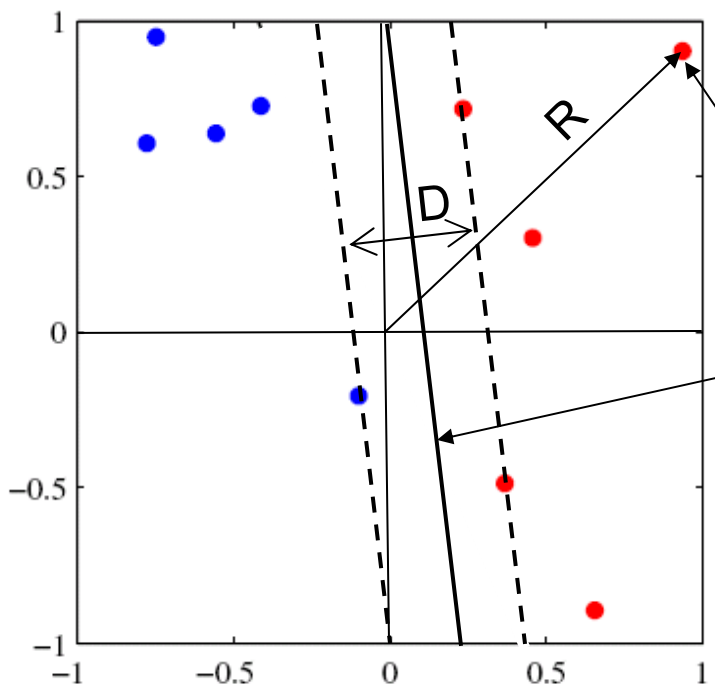

1
0.5
0
−0.5
−1
−1
−0.5
0
0.5
1

Perceptron

- Ale které řešení je to správné?
- Řešení, které poskytne učící algoritmus perceptronu záleží na inicializaci – počátečním **w**

- Algoritmus konverguje v méně než $(R/D)^2$ krocích





Opakování - MAP klasifikátor

- Mějme 2 třídy C_1 a C_2
  - Pro daný příznak x vyber třídu C s větší posteriorní pravděpodobností P(C|x)
  - Vyber C_1 pouze pokud:

$$P(\mathcal{C}_1|x) > P(\mathcal{C}_2|x)$$

$$\frac{P(x|\mathcal{C}_1)P(\mathcal{C}_1)}{P(x)} > \frac{P(x|\mathcal{C}_2)P(\mathcal{C}_2)}{P(x)}$$

$$\ln P(x|\mathcal{C}_1)P(\mathcal{C}_1) > \ln P(x|\mathcal{C}_2)P(\mathcal{C}_2)$$

$$\ln \frac{P(x|\mathcal{C}_1)P(\mathcal{C}_1)}{P(x|\mathcal{C}_2)P(\mathcal{C}_2)} > 0$$

Pravděpodobnostní generativní model

- Modelujme rozložení tříd gaussovskym rozložením:

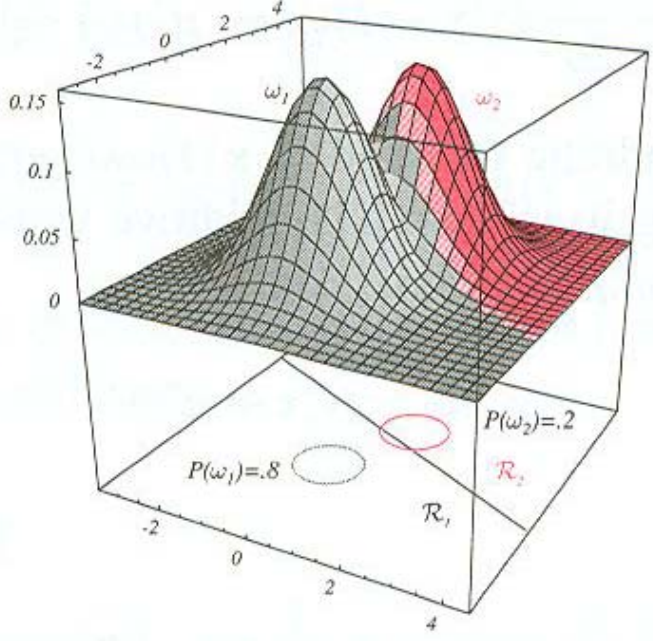
$$p(\mathbf{x}|\mathcal{C}_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\mathbf{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right\}$$

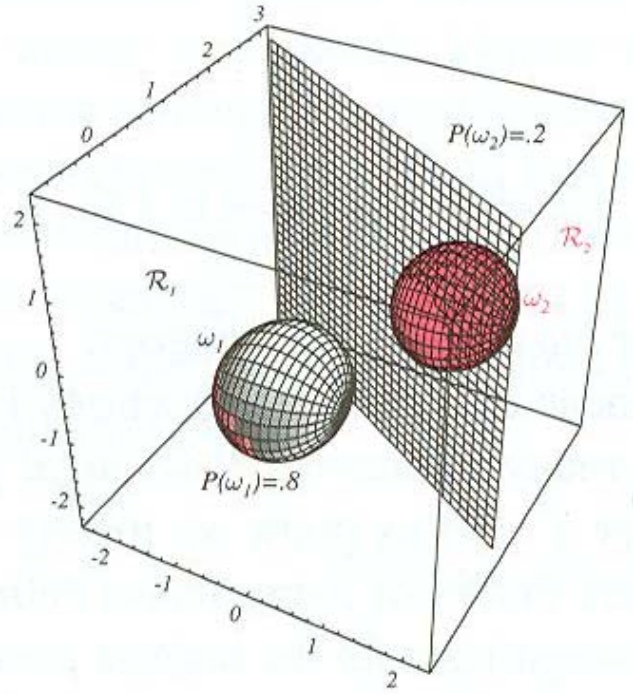
- Pokud náš model omezíme tak, ze každá třída má svou střední $\boldsymbol{\mu}_i$ hodnotu, ale kovarianční matice $\boldsymbol{\Sigma}$ je společná pro obě třídy, tak můžeme psát:

$$y(\mathbf{x}) = \ln \frac{P(x|\mathcal{C}_1)P(\mathcal{C}_1)}{P(x|\mathcal{C}_2)P(\mathcal{C}_2)} = \mathbf{w}^T\mathbf{x} + w_0$$

kde

$$\begin{aligned} \mathbf{w} &= \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ w_0 &= -\frac{1}{2}\boldsymbol{\mu}_1^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2^{\mathrm{T}}\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(\mathcal{C}_1)}{p(\mathcal{C}_2)} \end{aligned}$$


0.15
0.1
0.05
0
-2
0
2
4
ω_1
ω_2
$P(\omega_1)=.8$
$P(\omega_2)=.2$
$\mathcal{R}_1$
$\mathcal{R}_2$


3
2
1
0
-1
-2
-2
-1
0
1
2
$P(\omega_2)=.2$
$\mathcal{R}_2$
$\mathcal{R}_1$
ω_2
ω_1
$P(\omega_1)=.8$

Maximum likelihood odhad parametrů

- Hledáme parametry modelu

$$\{\mu_1, \mu_2, \mathbf{\Sigma}, P(\mathcal{C}_1), P(\mathcal{C}_2)\} = \arg \max_{\{\mu_1, \mu_2, \mathbf{\Sigma}, P(\mathcal{C}_1), P(\mathcal{C}_2)\}} \prod_i p(\mathbf{x}_i | \mu_{t_i}, \mathbf{\Sigma}) P(\mathcal{C}_{t_i})$$

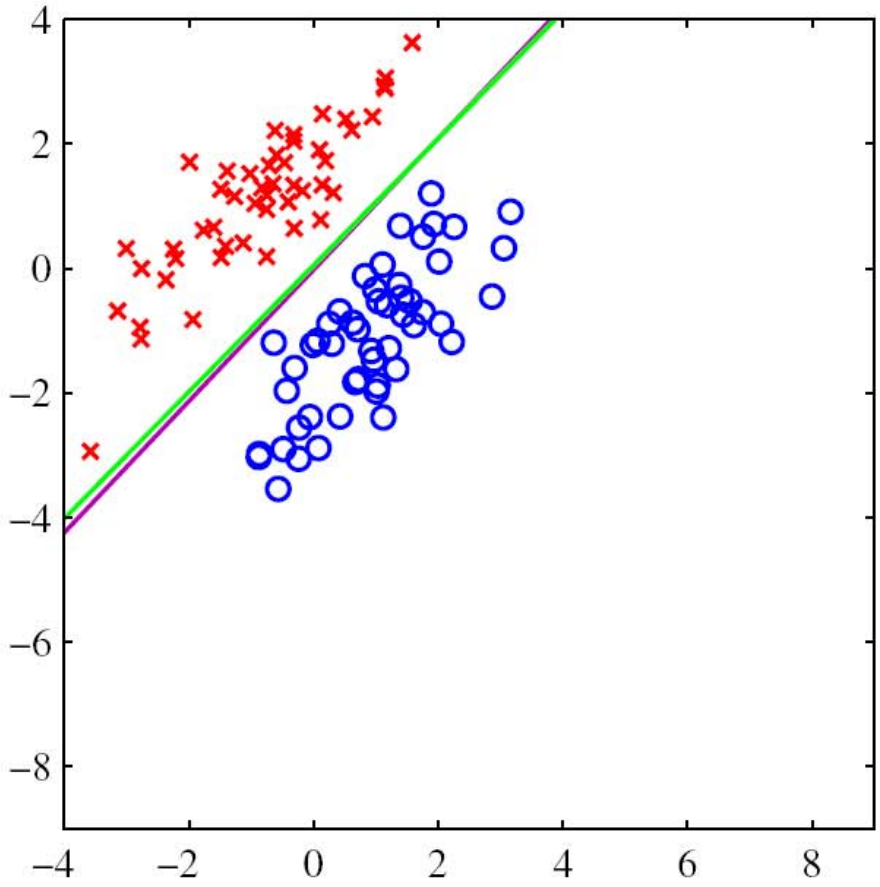
kde t_i je třída do které patří vzor $\mathbf{x}_i$ a $\boldsymbol{\mu}_{t_i}$ je střední hodnota této třídy

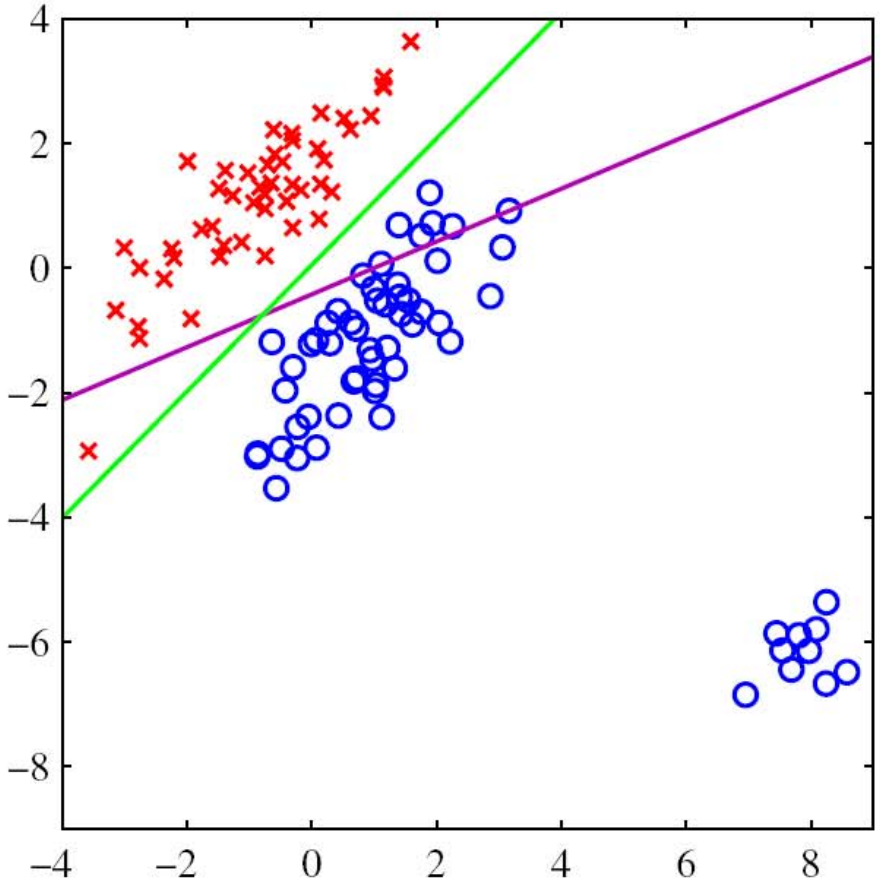
- Řešením jsou :
  - střední hodnoty spočítané z dat jednotlivých tříd
  - Kovarianční matice, která je váhovaným průměrem kovariančních matic spočtených z dat jednotlivých tříd

$$\hat{\boldsymbol{\mu}}^{(j)} = \frac{1}{N_j} \sum_{i=1}^{N_j} \mathbf{x}_i^{(j)}$$

$$\hat{\mathbf{\Sigma}}_{wc} = \frac{1}{N} \sum_{j=1}^{J} N_j \hat{\mathbf{\Sigma}}^{(j)}$$

$$\hat{\mathbf{\Sigma}}^{(j)} = \frac{1}{N_j} \sum_{i=1}^{N_j} (\mathbf{x}_i^{(j)} - \hat{\boldsymbol{\mu}}^{(j)})(\mathbf{x}_i^{(j)} - \hat{\boldsymbol{\mu}}^{(j)})^T$$

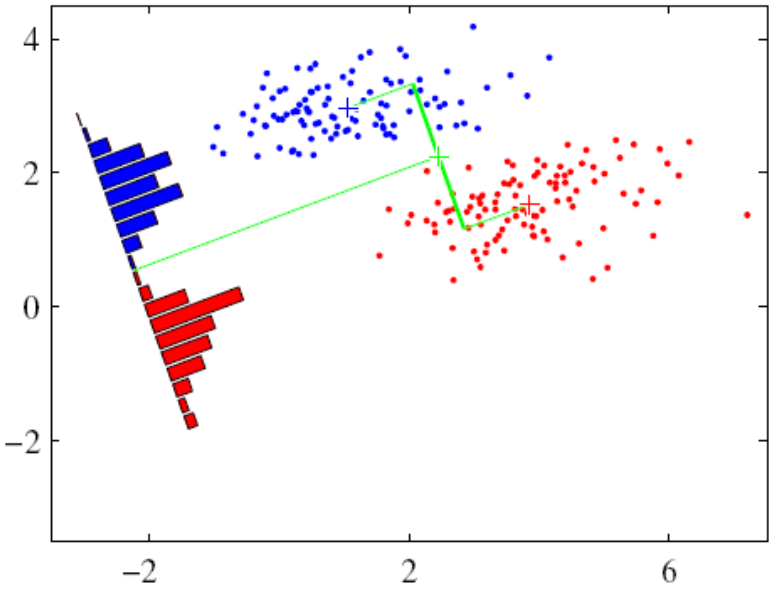




- V případě kdy ovšem naše data nerespektují předpoklad gaussovských rozložení a sdílené kovarianční matice. Klasifikátor může selhat – fialová rozhodovací linie
- Lepší výsledky dostaneme s diskriminativně natrénovaným klasifikátorem, který bude vysvětlen později – zelená rozhodovací linie

Opakování LDA

- Snažíme se data promítnout do takového směru, kde
  - Maximalizujeme vzdálenost mezi středními hodnotami tříd
  - Minimalizujeme průměrnou varianci tříd
- Maximalizujeme tedy



$$m_k = \mathbf{w}^{\mathrm{T}}\mathbf{m}_k$$

$$J(\mathbf{w}) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}$$

$$s_k^2 = \sum_{n \in \mathcal{C}_k} (y_n - m_k)^2$$

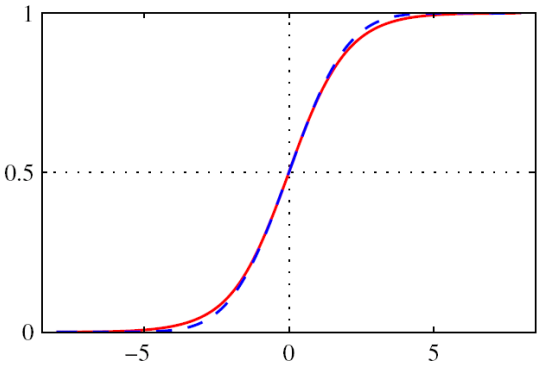
$$\mathbf{w} \propto \mathbf{S}_{\mathrm{W}}^{-1}(\mathbf{m}_2 - \mathbf{m}_1).$$

- Pro dvě třídy je **w** totožné s tím které jsme obdrželi pro náš generativní klasifikátor.
- Generativní klasifikátor ovšem zvolí i práh w_0

Generativní model a zobecněny lineární klasifikátor

Nyní použijme zobecněný lineární klasifikátor

$$y(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0)$$



kde stále platí, že

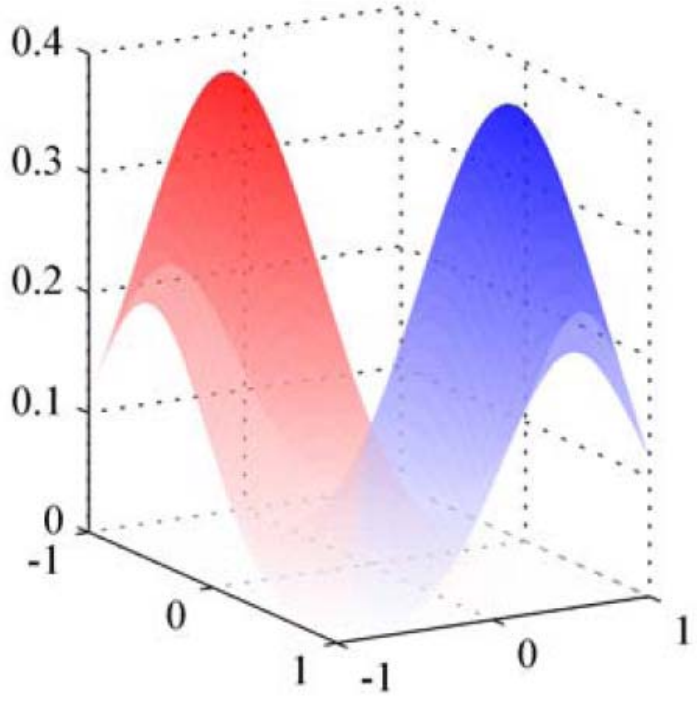
$$\mathbf{w}^T\mathbf{x} + w_0 = \ln \frac{P(x|\mathcal{C}_1)P(\mathcal{C}_1)}{P(x|\mathcal{C}_2)P(\mathcal{C}_2)}$$

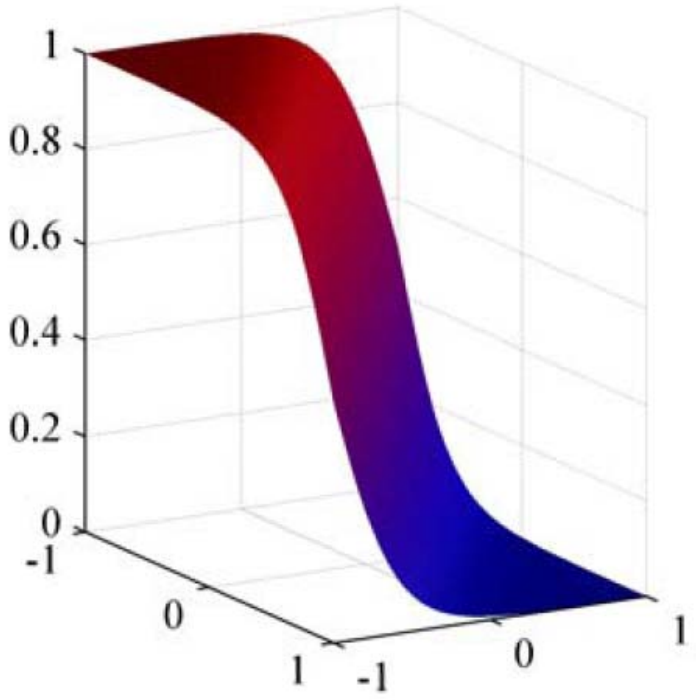
a kde aktivační funkce je logistická sigmoida

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

Potom lze hodnotu tohoto zobecněného linearního klasifikátoru přímo interpretovat jako posteriorní pravděpodobnost třídy C_1

$$p(\mathcal{C}_1|\mathbf{x}) = \sigma(\mathbf{w}^{\mathrm{T}}\mathbf{x} + w_0)$$


0.4
0.3
0.2
0.1
0
-1
0
1
-1
0
1


1
0.8
0.6
0.4
0.2
0
-1
0
1
-1
0
1

Jiné generativní lineární klasifikátory

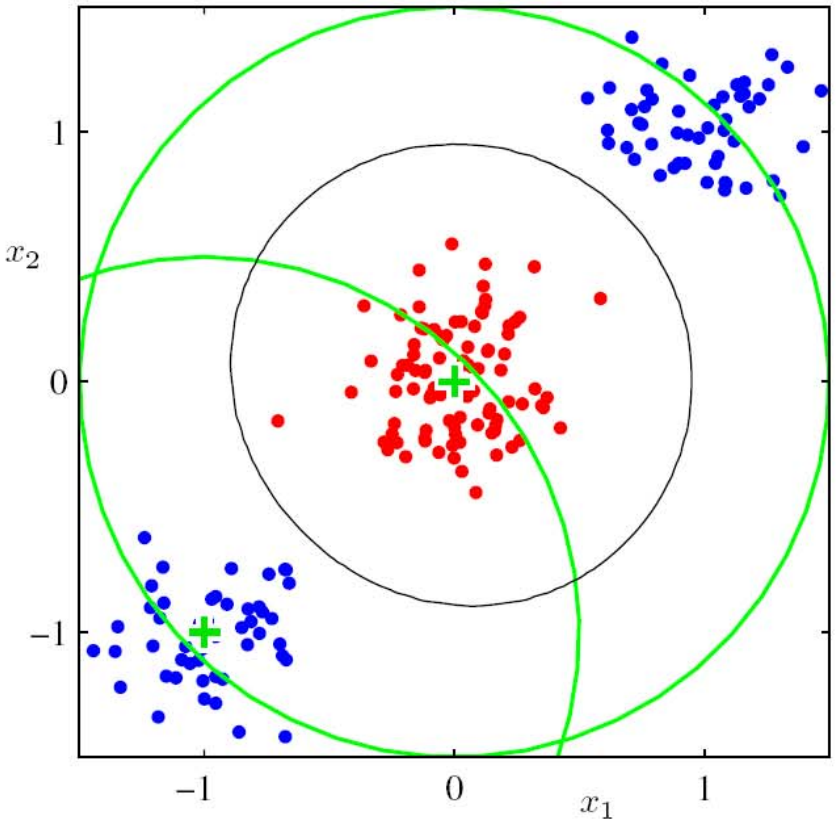
- Lineární klasifikátor dostaneme nejen pro gaussovské rozložení, ale pro celou třídu rozložení s exponencialní rodiny, které lze zapsat v následující formě:

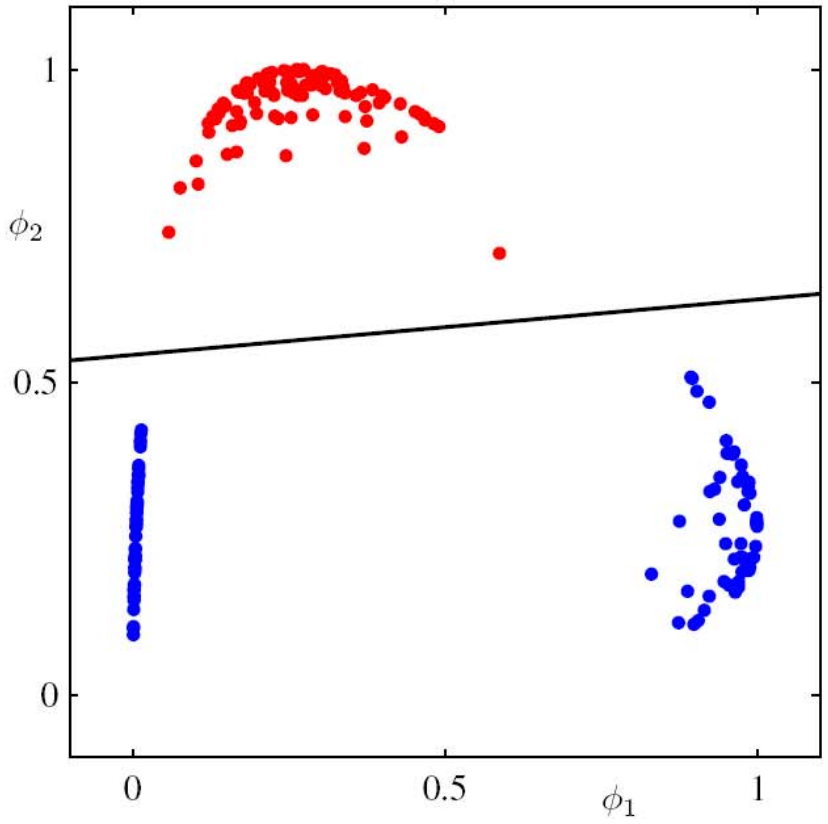
$$p(\mathbf{x}|\boldsymbol{\lambda}_k, s) = \frac{1}{s} h\left(\frac{1}{s}\mathbf{x}\right) g(\boldsymbol{\lambda}_k) \exp\left\{\frac{1}{s}\boldsymbol{\lambda}_k^{\mathrm{T}}\mathbf{x}\right\}$$

kde vektor $\boldsymbol{\lambda}_k$ má každá třída svůj vlastní, zatím co parametr s je sdíleny všemi třídami

Nelineární mapování vstupního vektoru

- Nelze-li původní data lineárně oddělit, možná pomůže jejich nelineární transformace do potenciálně vysocerozměrného prostoru – hlavní myšlenka „kernel methods“ které budou vysvětleny příště
- V našem příkladu pomohlo i mapování dvourozměrných dat do dvou gaussovských funkcí





Lineární logistická regrese

- Uvažujme opět pravděpodobnostní model

$$p(\mathcal{C}_1|\mathbf{x}) = y(\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x})$$

kde opět pro zjednodušení x_0 je vžy 1 a nemusíme tedy explicitně zavádět w_0.

- Nyní ale budeme parametry **w** odhadovat přímo tak abychom maximalizovali pravděpodobnost, že všechna trénovací data budou rozpoznána správně

$$p(\mathbf{t}|\mathbf{X}) = \prod_{n\in\mathcal{C}_1} y(\mathbf{x}_n) \prod_{n\in\mathcal{C}_2} (1 - y(\mathbf{x}_n))$$

Kde **t** je vektor korektních identit tříd t_i pro jednotlivé vstupní vektory $\mathbf{x}_n$. Pro zjednodušení zápisu předpokládejme, že $t_n = 1$, pokud $\mathbf{x}_n$ paří do třídy C_1 a $t_n = 0$ pokud $\mathbf{x}_n$ paří do třídy C_2.Potom muzeme psát

$$p(\mathbf{t}|\mathbf{X}) = \prod_n y_n^{t_n}(1 - y_n)^{1-t_n}$$

Lineární logistická regrese – odhad parametrů

- Lépe se nám bude pracovat s logaritmem naší objektivní funkce, což je chybová funkce známo jako *vzájemná entropie*

$$E(\mathbf{w}) = -\ln p(\mathbf{t}|\mathbf{w}) = -\sum_{n=1}^{N} \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

- Hledáme minimum této funkce, takže derivujeme abychom dostali gradient

$$\nabla E(\mathbf{w}) = \sum_{n=1}^{N} (y_n - t_n)\mathbf{x}_n$$

- Pokud najdeme parametry **w** pro které je gradient nulový, našli jsme optimum chybové funkce. To však není snadné nalézt analyticky. Budeme řešit nymericky, např pomocí *gradient descent*

$$\mathbf{w}^{(\tau+1)} = \mathbf{w}^{(\tau)} - \eta \nabla E_{\mathrm{P}}(\mathbf{w})$$

Lineární logistická regrese – odhad parametrů

- Rychlejší konvergenci dosáhneme pomocí Newton-Raphson optimalizace:
  - Kolem stávajícího řešení $\mathbf{w}^{(\text{old})}$ aproximujeme chybovou funkci ∇E pomoci Taylorova rozvoje druhého řádu, čímž obdržíme kvadratickou formu (vícerozměrné zobecnění kvadratické funkce).
  - Jako nové řešení zvolíme to, kde má tato kvadratická forma minimum.

$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - \mathbf{H}^{-1}\nabla E(\mathbf{w})$$

kde $\mathbf{H} = \mathbf{X}^{\mathbf{T}}\mathbf{R}\,\mathbf{X}$ je matice druhých derivací (Hessian matrix).

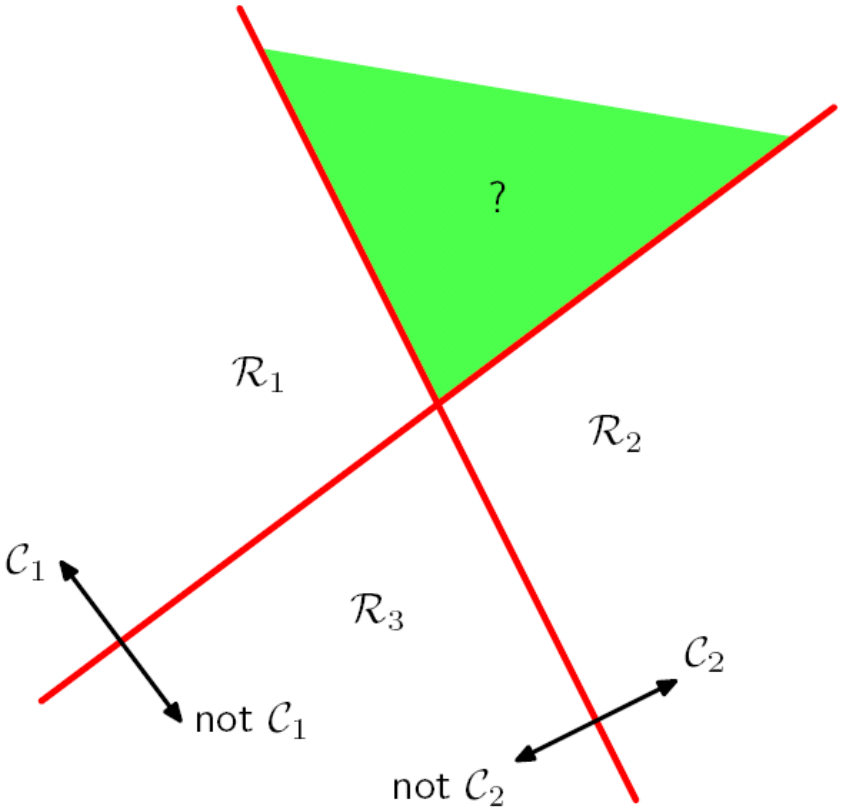
$$\mathbf{w}^{(\text{new})} = \mathbf{w}^{(\text{old})} - (\mathbf{X}^T\mathbf{R}\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} - \mathbf{t})$$

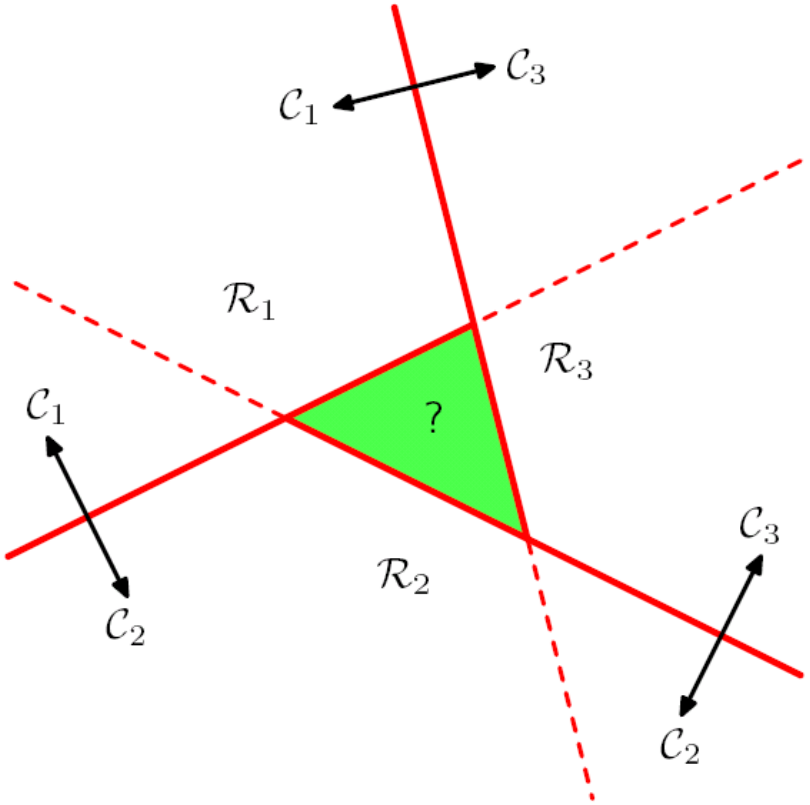
$\mathbf{R}$ je diagonální matice s hodnotami na diagonále:

$$R_{nn} = y_n(1 - y_n)$$

Problém s více třídami

- Klasifikace
  - jeden proti všem
  - Každý s každým





Lineární klasifikátor – více tříd

- Nejlépe je mít jednu lineární funkci pro každou třídu k

$$y_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}$$

- Vyber třídu s největším $y_k(\mathbf{x})$
- Rozhodovací linie je opět lineární dána

$$y_k(\mathbf{x}) = y_j(\mathbf{x})$$

- Kde *k* a *j* jsou dvě nejpravděpodobnější třídy pro dané **x**
- Pro dvě třídy řešení degraduje k tomu co už jsme viděli

$$(\mathbf{w}_k - \mathbf{w}_j)^{\mathrm{T}}\mathbf{x} + (w_{k0} - w_{j0}) = 0$$

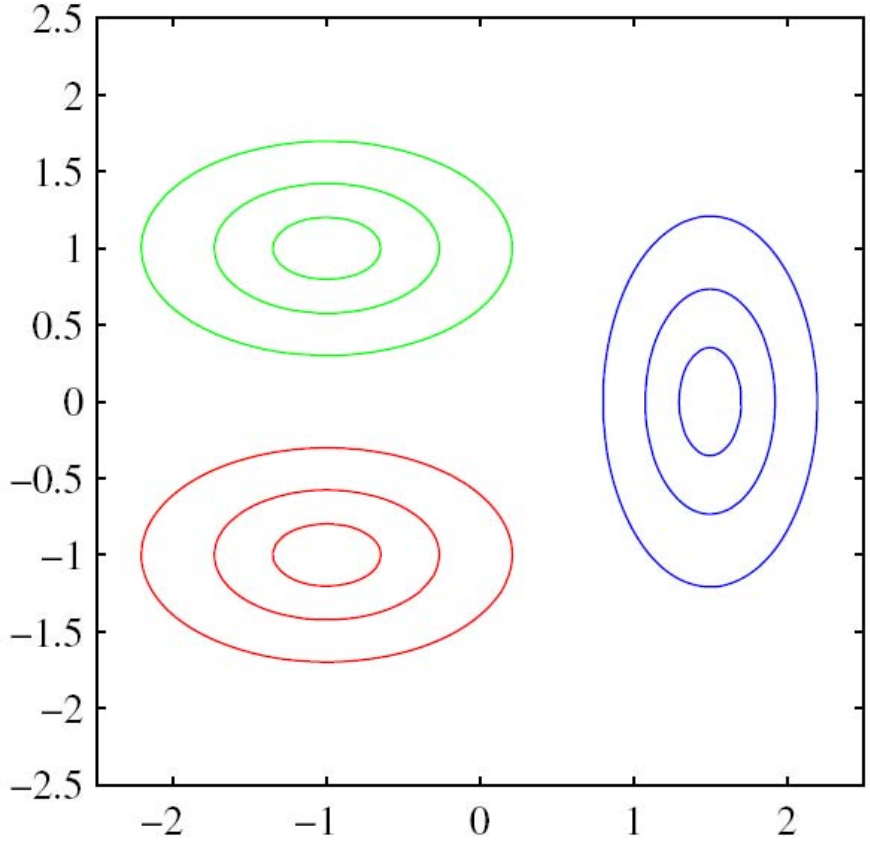
Více tříd – generativní model

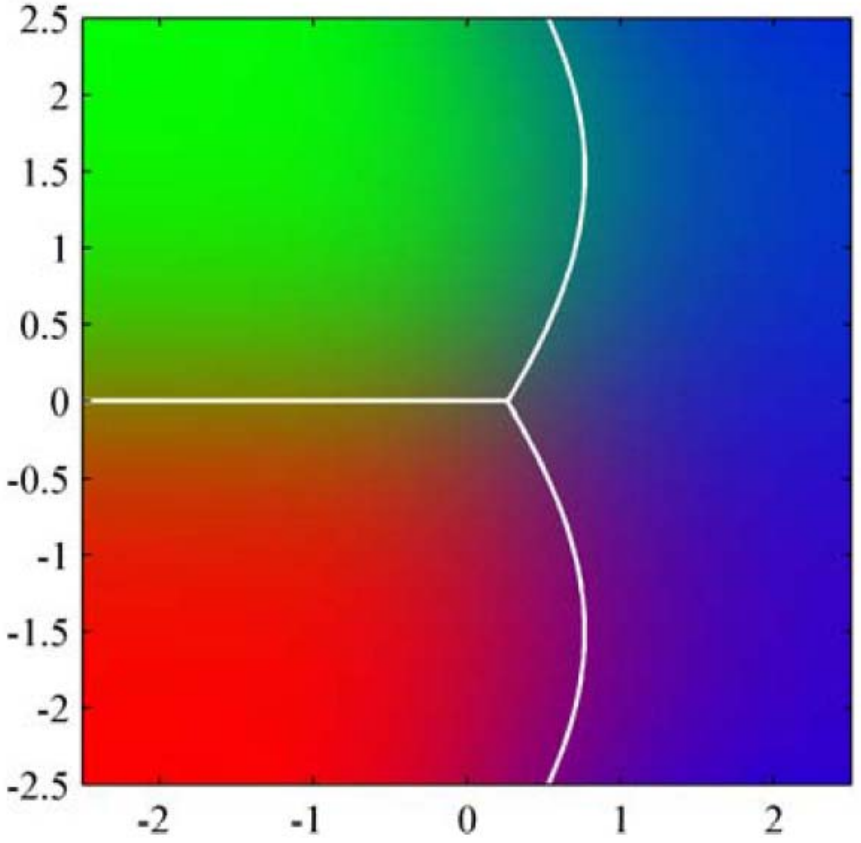
$$
\begin{aligned}
p(\mathcal{C}_k|\mathbf{x}) &= \frac{p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)}{\sum_j p(\mathbf{x}|\mathcal{C}_j)p(\mathcal{C}_j)} \\
&= \frac{\exp(a_k)}{\sum_j \exp(a_j)}
\end{aligned}
$$

$$
a_k = \ln p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)
$$

$$
\begin{aligned}
\mathbf{w}_k &= \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k \\
w_{k0} &= -\frac{1}{2}\boldsymbol{\mu}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k + \ln p(\mathcal{C}_k)
\end{aligned}
$$

$$
a_k(\mathbf{x}) = \mathbf{w}_k^{\mathrm{T}}\mathbf{x} + w_{k0}
$$


2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−2
−1
0
1
2


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-2.5
-2
-1
0
1
2