ANALÝZA A KLASIFIKACE DAT

prof. Ing. Jiří Holčík, CSc.

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

IV. LINEÁRNÍ KLASIFIKACE

pokračování

ALGORITMUS PODPŮRNÝCH VEKTORŮ
(SUPPORT VECTOR MACHINE – SVM)

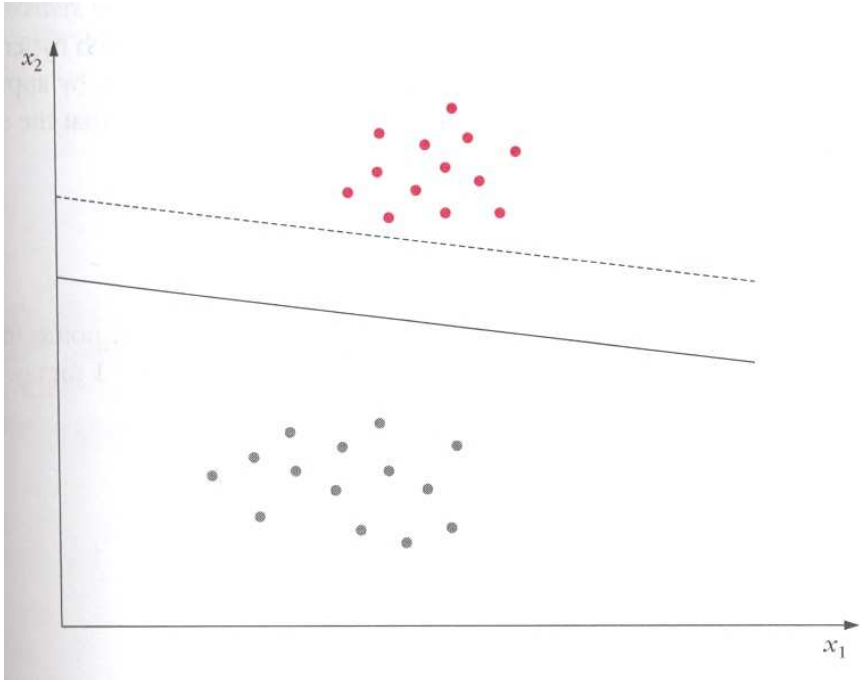
SEPARABILNÍ TŘÍDY

mějme v učební množině obrazy $\mathbf{x}_i$, i=1,2,…,n, ze dvou lineárně separabilních klasifikačních tříd ω_1 a ω_2

cílem je určení parametrů definující hranici

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0 = 0,$$

jejíž pomocí klasifikátor správně zařadí všechny obrazy z učební množiny

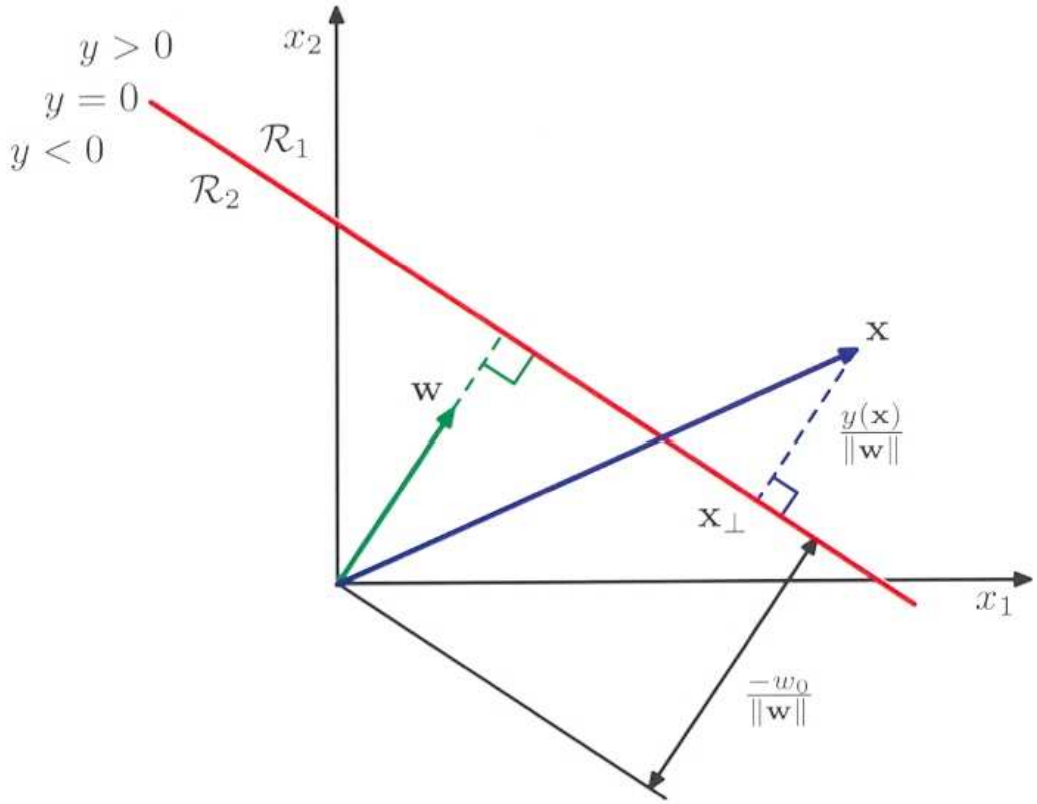


ALGORITMUS PODPŮRNÝCH VEKTORŮ
SEPARABILNÍ TŘÍDY

- připomenutí:

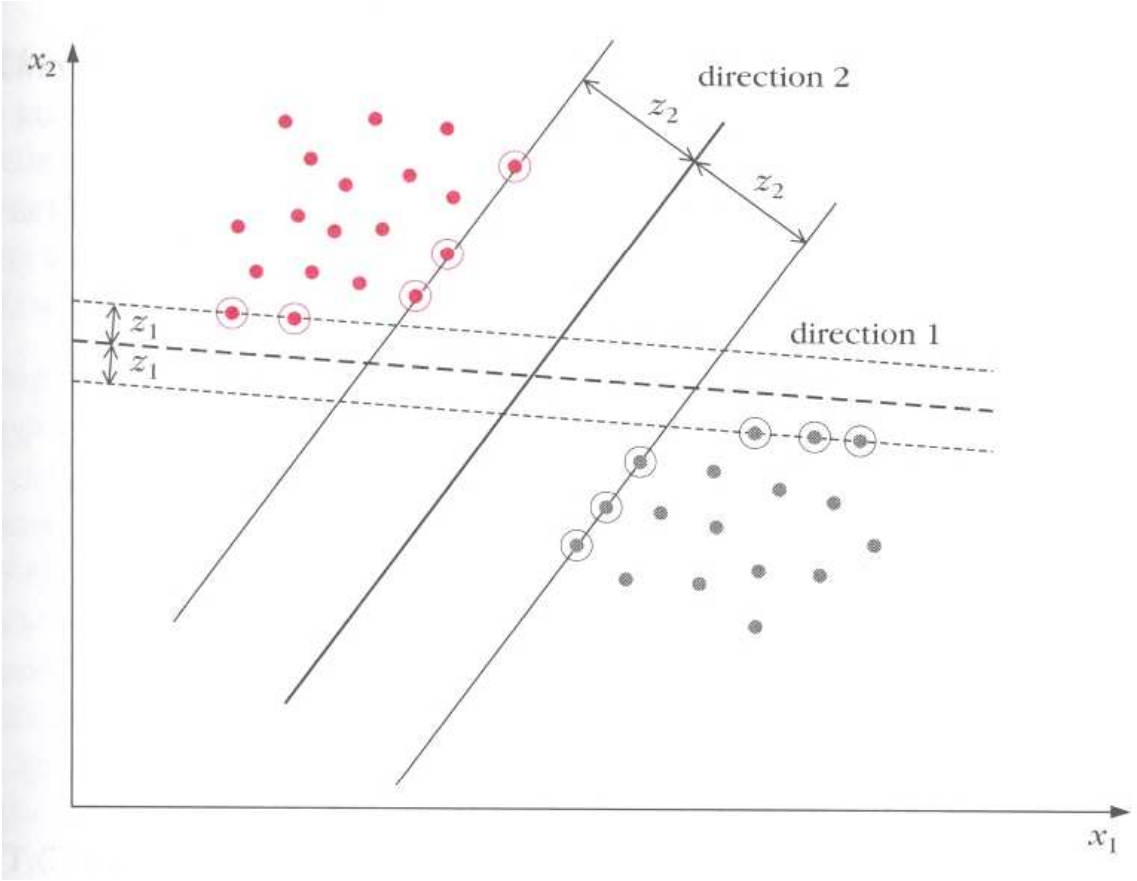
  vzdálenost jakéhokoliv bodu od klasifikační hranice je

$$d = \frac{|y(\mathbf{x})|}{\|\mathbf{w}\|}$$



ALGORITMUS PODPŮRNÝCH VEKTORŮ
SEPARABILNÍ TŘÍDY

- určeme hodnoty váhového vektoru **w** a w_0 tak, aby hodnota y(**x**) v nejbližším bodě třídy ω_1 byla rovna 1 a pro ω_2 rovna -1



ALGORITMUS PODPŮRNÝCH VEKTORŮ
SEPARABILNÍ TŘÍDY

- máme „ochranné“ klasifikační pásmo o šířce

$$\frac{1}{\|\mathbf{w}\|}+\frac{1}{\|\mathbf{w}\|}=\frac{2}{\|\mathbf{w}\|}$$

a chceme

$$\mathbf{w}^T\mathbf{x}+w_0 \geq 1 \quad \text{pro } \forall \mathbf{x} \in \omega_1$$

$$\mathbf{w}^T\mathbf{x}+w_0 \geq -1 \quad \text{pro } \forall \mathbf{x} \in \omega_2$$

nebo také - chceme najít minimální

$$J(\mathbf{w},w_0)=\frac{1}{2}\|\mathbf{w}\|^2$$

za předpokladu, že

$$t_i(\mathbf{w}^T\mathbf{x}+w_0) \geq 1, \quad i=1,2,\ldots,n$$

kde $t_i = +1$ pro ω_1 a $t_i = -1$ pro ω_2

(minimalizace normy maximalizuje klasifikační pásmo)

ALGORITMUS PODPŮRNÝCH VEKTORŮ
SEPARABILNÍ TŘÍDY

- nelineární kvadratická optimalizační úloha se soustavou podmínek formulovaných pomocí lineárních nerovností
- Karushovy-Kuhnovy-Tuckerovy podmínky praví, že pro to musí být splněno

$$\frac{\partial}{\partial \mathbf{w}} L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \mathbf{0}$$

$$\frac{\partial}{\partial w_0} L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = 0$$

$$\lambda_i \geq 0, i = 1, 2, \ldots, n$$

$$\lambda_i [t_i (\mathbf{w}^T \mathbf{x} + w_0) - 1] = 0, i = 1, 2, \ldots, n,$$

kde $\boldsymbol{\lambda}$ je vektor Langrangových součinitelů a $L(\mathbf{w}, w0, \boldsymbol{\lambda})$ je Lagrangova funkce definována vztahem

$$L(\mathbf{w}, w_0, \boldsymbol{\lambda}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^{n} \lambda_i [t_i (\mathbf{w}^T \mathbf{x}_i + w_0) - 1]$$

ALGORITMUS PODPŮRNÝCH VEKTORŮ
SEPARABILNÍ TŘÍDY

- když se všechny vztahy z předcházející strany dají dohromady dostaneme

$$\mathbf{w} = \sum_{i=1}^{n} \lambda_i t_i \mathbf{x}_i$$

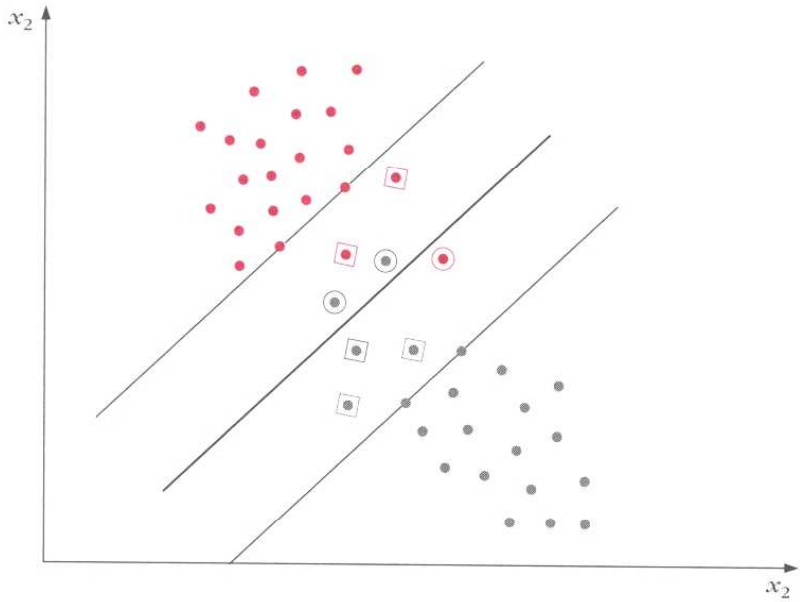
podpůrné vektory

$$\sum_{i=1}^{n} \lambda_i t_i = 0$$

ALGORITMUS PODPŮRNÝCH VEKTORŮ
NESEPARABILNÍ TŘÍDY

- stále ale platí, že klasifikační „ochranné" pásmo je definováno dvěma paralelními „nadrovinami" definovanými

$$\mathbf{w}^T\mathbf{x} + w_0 = \pm 1$$



- obrazy z trénovací množiny patří do následujících tří kategorií:
  - obraz leží **vně** pásma a je **správně** klasifikován [platí podmínka $t_i(\mathbf{w}^T\mathbf{x}+w_0)\geq 1$ $i=1,2,\ldots,n$];
  - obraz leží **uvnitř** pásma a je **správně** klasifikován (čtverečky) [platí pro ně $0\leq t_i(\mathbf{w}^T\mathbf{x}+w_0)<1$];
  - obraz je chybně klasifikován (kolečka) [platí pro něj $t_i(\mathbf{w}^T\mathbf{x}+w_0)<0$]

ALGORITMUS PODPŮRNÝCH VEKTORŮ
NESEPARABILNÍ TŘÍDY

- všechny tři kategorie obrazů mohou být řešeny na základě pro daný typ specifických podmínek

$$t_i(\mathbf{w}^T\mathbf{x}+w_0)\geq 1-\xi_i$$

pomocí nově zavedených proměnných ξ_i (tzv. volné proměnné - slack variables).

První kategorie je pro $\xi_i=0$, druhá $0<\xi_i\leq 1$ a třetí pro $\xi_i>1$.

Cílem návrhu v tomto případě je vytvořit co nejširší „ochranné" pásmo, ale současně minimalizovat počet obrazů s $\xi_i>0$, což vyjadřuje kritérium se ztrátovou funkcí

$$J(\mathbf{w},w_0,\boldsymbol{\xi})=\frac{1}{2}\|\mathbf{w}\|^2+C\sum_{i=1}^{n}I(\xi_i)$$

C je kladná korekční konstanta, která váhuje vliv obou členů v uvedeném vztahu.

kde $\boldsymbol{\xi}$ je vektor parametrů ξ_i a

$$I(\xi_i)=\begin{cases}1 & \xi_i>0\\ 0 & \xi_i=0\end{cases}$$

ALGORITMUS PODPŮRNÝCH VEKTORŮ
NESEPARABILNÍ TŘÍDY

- optimalizace je obtížná, protože ztrátová funkce je nespojitá (díky funkci I(•)). V takových případech se proto používá náhradní ztrátová funkce

$$J(\mathbf{w}, w_0, \boldsymbol{\xi}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i$$

- a cílem návrhu je minimalizovat $J(\mathbf{w}, w_0, \boldsymbol{\xi})$ za podmínek, že

$$t_i(\mathbf{w}^T\mathbf{x}+w_0) \geq 1-\xi_i \text{ a } \xi_i \geq 0,\ i=1,2,\ldots,n.$$

- Problém lze opět řešit pomocí Langrangeovy funkce

$$L(\mathbf{w}, w_0, \boldsymbol{\xi}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{n}\xi_i - \sum_{i=1}^{n}\mu_i\xi_i - \sum_{i=1}^{n}\lambda_i[t_i(\mathbf{w}^T\mathbf{x}_i + w_0) - 1 + \xi_i]$$

ALGORITMUS PODPŮRNÝCH VEKTORŮ
NESEPARABILNÍ TŘÍDY

- příslušné Karushovy-Kuhnovy-Tuckerovy podmínky jsou

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{0} \text{ nebo } \mathbf{w} = \sum_{i=1}^{n} \lambda_i t_i \mathbf{x}_i;$$

$$\frac{\partial L}{\partial w_0} = 0 \text{ nebo } \sum_{i=1}^{n} \lambda_i t_i \mathbf{x}_i;$$

$$\frac{\partial L}{\partial \xi_i} = 0 \text{ nebo } C - \mu_i - \lambda_i = 0, i = 1, 2, \ldots, n;$$

$$\lambda_i \left[t_i (\mathbf{w}^T \mathbf{x} + w_0) - 1 + \xi_i \right] = 0, i = 1, 2, \ldots, n;$$

$$\mu_i \xi_i = 0, i = 1, 2, \ldots, n;$$

$$\mu_i \geq 0, \lambda_i \geq 0, i = 1, 2, \ldots, n;$$

ALGORITMUS PODPŮRNÝCH VEKTORŮ
NESEPARABILNÍ TŘÍDY

- z čehož platí požadavek na maximalizaci $L(\mathbf{w}, w_0, \boldsymbol{\lambda}, \boldsymbol{\xi}, \boldsymbol{\mu})$ za podmínek

$$\mathbf{w} = \sum_{i=1}^{n} \lambda_i t_i \mathbf{x}_i;$$

$$\sum_{i=1}^{n} \lambda_i t_i = 0;$$

$$C - \mu_i - \lambda_i = 0,$$

$$\mu_i \geq 0, \lambda_i \geq 0, i = 1, 2, \ldots, n;$$

ALGORITMUS PODPŮRNÝCH VEKTORŮ
VÍCE KLASIFIKAČNÍCH TŘÍD

- přímé rozšíření řešení případu dichotomického problému podle schématu
  - „jedna versus zbytek" – M dichotomických úloh; každý klasifikátor je trénován podle schématu s hraniční funkcí $y_i(\mathbf{x})>0$ pro obrazy z ω_i a $y_i(\mathbf{x})<0$ pro všechny ostatní;
  - „jedna versus jedna" – M(M-1)/2 binárních klasifikátorů
  - klasifikační schéma používající K binárních klasifikátorů, přičemž jednotlivé shluky obrazů z jednotlivých tříd jsou stanoveny navrhovatelem a jsou kódovány vektory o délce K, jehož hodnoty jsou +1 nebo -1.
    např. pro M=4 a K=6 může být taková matice

$$\begin{bmatrix} -1 & -1 & -1 & +1 & -1 & +1 \\ +1 & -1 & +1 & +1 & -1 & -1 \\ +1 & +1 & -1 & -1 & -1 & +1 \\ -1 & -1 & +1 & -1 & +1 & +1 \end{bmatrix}$$

Příprava nových učebních materiálů pro obor Matematická biologie

je podporována projektem ESF

č. CZ.1.07/2.2.00/07.0318

„VÍCEOBOROVÁ INOVACE STUDIA MATEMATICKÉ BIOLOGIE"

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ