FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

LINEÁRNÍ MODELY, DISKRIMINAČNÍ ANALÝZA A PODPŮRNÉ VEKTORY

Autor textu:
Ing. Petr Honzík, Ph.D.

Květen 2014

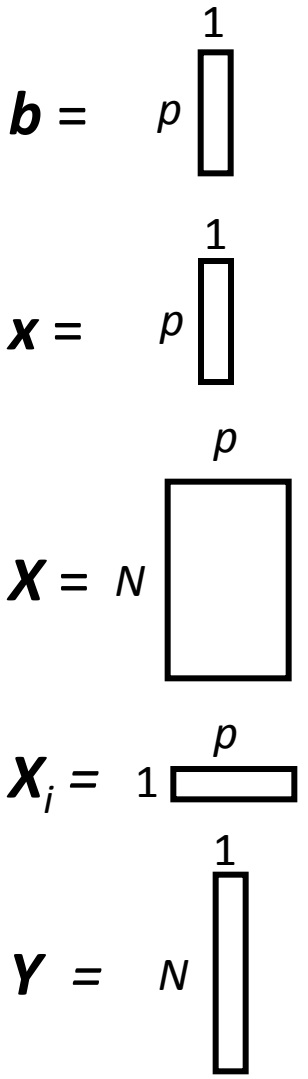
Komplexní inovace studijních programů a zvyšování kvality výuky na FEKT VUT v Brně
OP VK CZ.1.07/2.2.00/28.0193

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Regresní lineární model – symboly

Použité značení

- $\boldsymbol{b}$ – parametry modelu (vektor $p \times 1$)
- p – počet atributů (skalár)
- N – počet příkladů (skalár)
- $\boldsymbol{x}$ – jeden příklad (vektor $p \times 1$)
- $\boldsymbol{x}_i$ – i-tá hodnota příkladu x (skalár)
- $\boldsymbol{X}$ – matice všech příkladů (matice $N \times p$)
- $\boldsymbol{X}_i$ – i-tý příklad (vektor $1 \times p$)
- y – výstupní hodnota (skalár)
- $\boldsymbol{Y}$ – vektor výstupů ($p \times 1$)
- $\boldsymbol{Y}_i$ – i-tý výstup (skalár)



Regresní lineární model – úvod

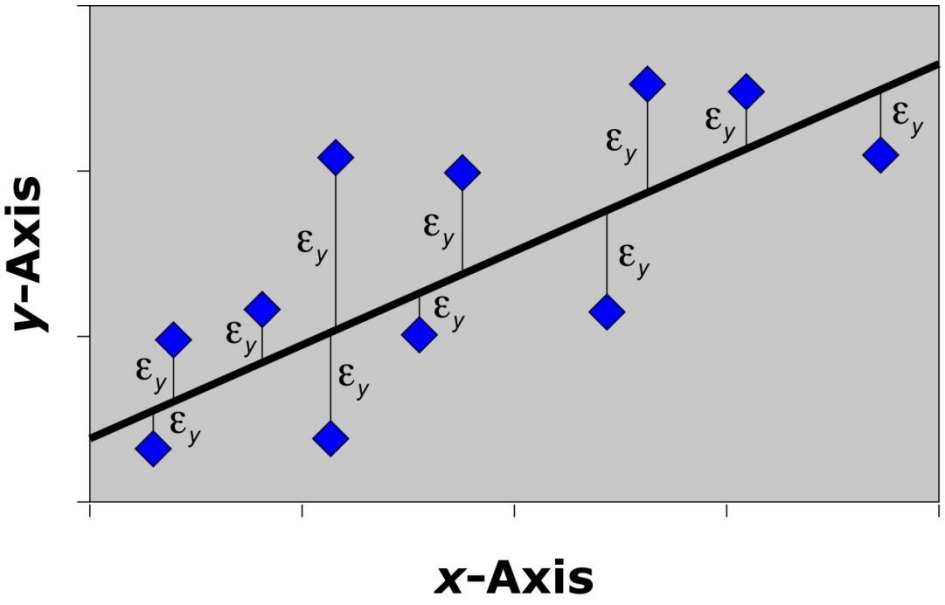
- znalost uložena v koeficientech lineárního modelu
- odvození koeficientů metodou nejmenších čtverců (MNČ)
- existují modely linearizovatelné (např. logitový model)

- Lineární model:

$$y = b_0 + \sum_{j=1}^{p} x_j b_j$$

- Predikce:

$$\hat{y}(x) = x^T \hat{b}$$



palaeomath.palass.org

Regresní lineární model – vlastnosti

- často svou přesností **překonají nelineární** modely, zejména v situaci, když je **málo trénovacích dat**
- silný bias (předpojatost), menší podíl chyby variance
- vstupní veličina $\boldsymbol{x}$ může být různě transformována: x^2, $\sin(\boldsymbol{x})$, $\boldsymbol{x}_1\boldsymbol{x}_2$, $\log(\boldsymbol{x})$,…
- jsou tyto modely lineární ?
  - $y=b_0+b_1x_1$
  - $y=b_0+b_1x_1+ b_0b_1x_1$
  - $y=b_0+b_1x+ b_2e^x$
  - $y=b_0\sin(x_1)$
  - $y=b_0\sin(b_1x_1)$
  - $y=x_1+ x_2^b$
- zajímá nás
  - jak určit parametry $\boldsymbol{b}$
  - interval spolehlivosti těchto parametrů
  - jak regresní model použít ke klasifikaci

Regresní lineární model – *b*

- Odvození hodnot parametrů modelu tak, aby chyba ve výstupní veličině byla co nejmenší.
- Výpočet chyby: metoda nejmenších čtverců

$$RSS(b) = \sum_{i=1}^{N} \left(y_i - X_i\, b\right)^2 = \left(Y - Xb\right)^T \left(Y - Xb\right)$$

- Po derivaci podle ***b*** získáme tvar:

$$\frac{\partial RSS}{\partial b} = -2X^T\left(Y - Xb\right) \qquad \frac{\partial^2 RSS}{\partial b \partial b^T} = -2X^T X$$

Regresní lineární model – *b*

- Postavím 1. derivaci rovnu 0

$$X^T(Y - Xb) = 0 \quad \Rightarrow \quad \hat{b} = (X^T X)^{-1} X^T Y$$

- Pro konečný model tedy platí:

$$\hat{Y} = X^T \hat{b} = X(X^T X)^{-1} X^T Y$$

- Pokud X^TX je matice singulární (není invertibilní), parametry ***b*** nejsou jednoznačně určeny. Lze ubrat atributy (způsobují atributy lineárně závislé – korelace = ±1).

Co způsobuje singularitu matice X^TX? Jak ji lze odstranit?

Regresní lineární model – interval spolehlivosti *b*

- Předpoklad je, že $\widehat{Y} = X\widehat{b} + \varepsilon$, kde $\varepsilon \sim N(0,\sigma^2)$
- Kovarianční matici parametru ***b*** vypočteme jako:

$$Cov(b) = \left(X^T X\right)^{-1}\sigma^2, \; kde \quad \hat{\sigma} = \frac{1}{N-p-1}\left(Y-\hat{Y}\right)^T\left(Y-\hat{Y}\right)$$

- Odhadu parametru ***b*** podléhá normálnímu rozložení

$$\widehat{b} \approx N\left(b, \left(X^T X\right)^{-1}\sigma^2\right)$$

- Hypotézu H_0: $\boldsymbol{b}_i$=0 posoudíme podle *Z*-skóre. Čím je absolutní hodnota *z* větší, s tím větší pravděpodobností lze H_0 zamítnout

$$z_i = \frac{\widehat{b}_i}{\widehat{\sigma}\sqrt{v_i}}$$

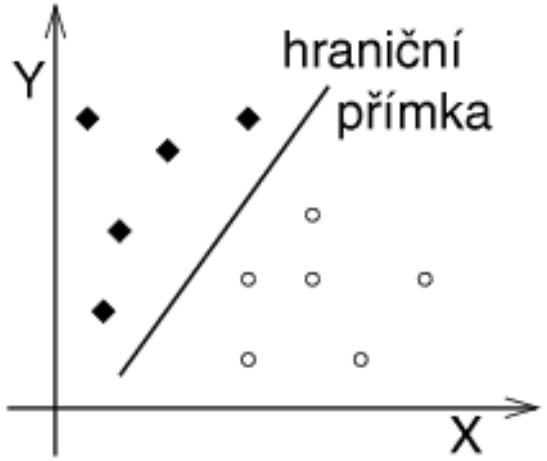
kde v_i je i-tý prvek na hlavní diagonále matice $(X^TX)^{-1}$

Lineární model – klasifikace binární

- o klasifikaci rozhoduje **vzdálenost od roviny**
- třídy nahrazeny čísly 0 a 1 nebo -1 a 1
- je definována **rozhodovací hranice** (= kritická hodnota, prahová hodnota)
  - kódování 0/1 – kritická hodnota = 0,5
  - kódování -1/1 – kritická hodnota = 0

Klasifikace:

$$\hat{G}(x) = \begin{cases} -1: \hat{f}(x) \leq 0 \\ 1: \hat{f}(x) > 0 \end{cases}$$



Lineární model – klasifikace vícerozměrná

- mějme třídy $g_1, \ldots, g_K$
- vytvoření K modelů, přičemž pro každý:
  - vytvořena nová data, respektive nová výstupní veličina ***G***
  - matice ***G*** (***indcator matrix***) má rozměr $N \times K$
  - hodnoty matice tvořeny 1 a 0 (podle příslušnosti do třídy)
  - naučeno K modelů f_i

Klasifikace:

$$\hat{G}(x) = \arg\max_{i=1..K} \hat{f}_i(x)$$

Jak se vytváří tzv. indicator matrix?

Lineární model – klasifikace vícerozměrná

- protože ***G*** je obdélníková matice, i matice parametrů ***B*** je obdélníková
- paralelně se během výpočtu vytvoří *K* lineárních modelů

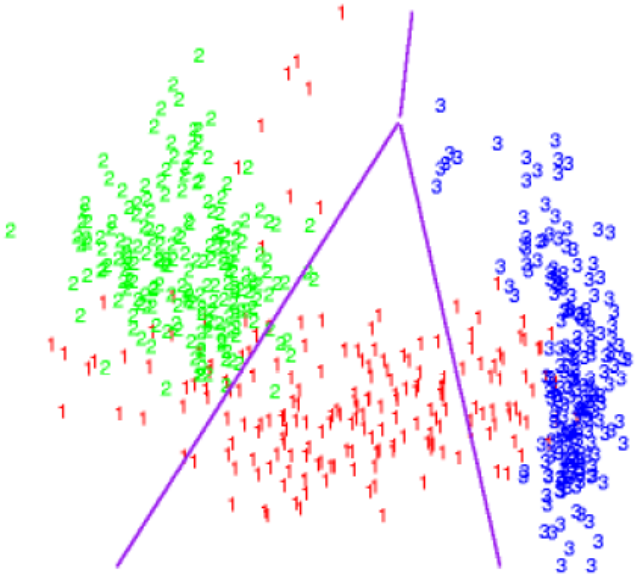
$$\hat{G} = X\hat{B}$$

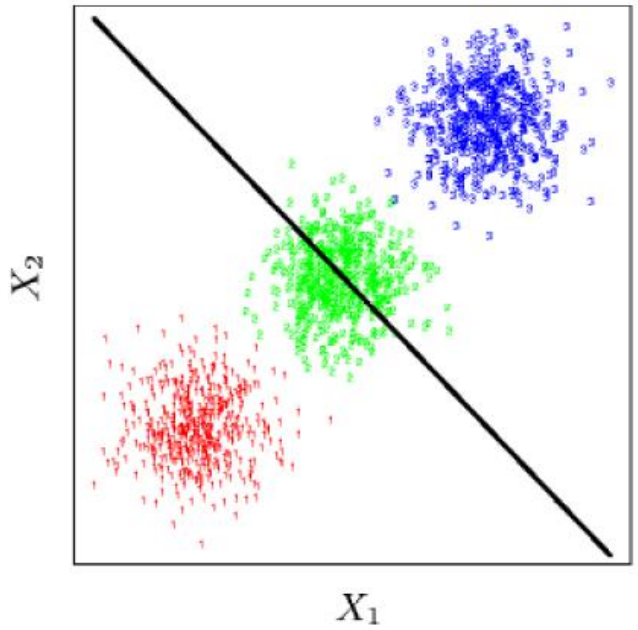
$$\hat{G} = X\left(X^T X\right)^{-1} X^T G \Rightarrow \hat{B} = \left(X^T X\right)^{-1} X^T G$$

$\hat{G} =$ (*N* × *K*) $\qquad \hat{B} =$ (*p* × *K*)

Lineární model – klasifikace vícerozměrná

- problém s **maskováním třídy** (když prostřední shluk = 0, ostatní = 1; funkce této substituce vede středem třídy a bude maskována jednou z krajních tříd)





The elements of statistical learning by T. Hastie and col.

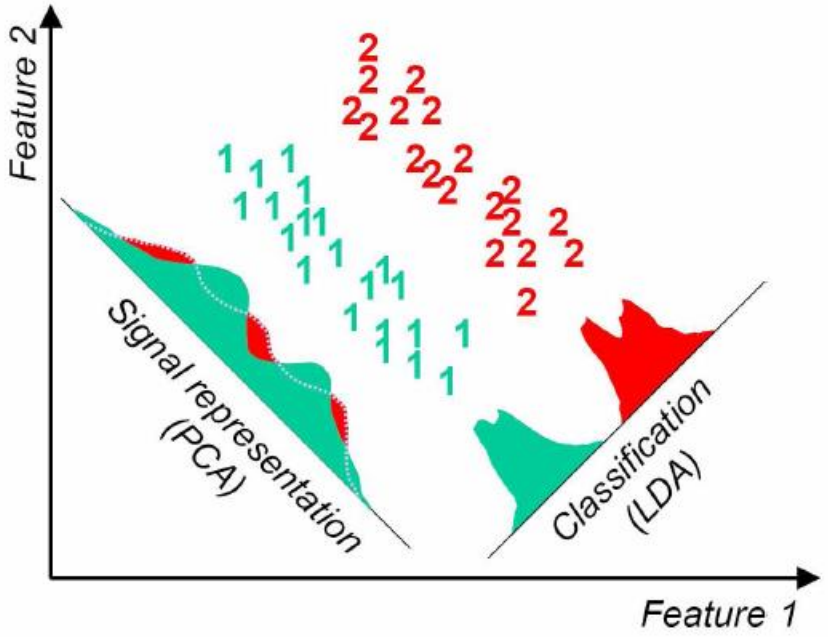
Co je to maskování tříd, čím je způsobeno?

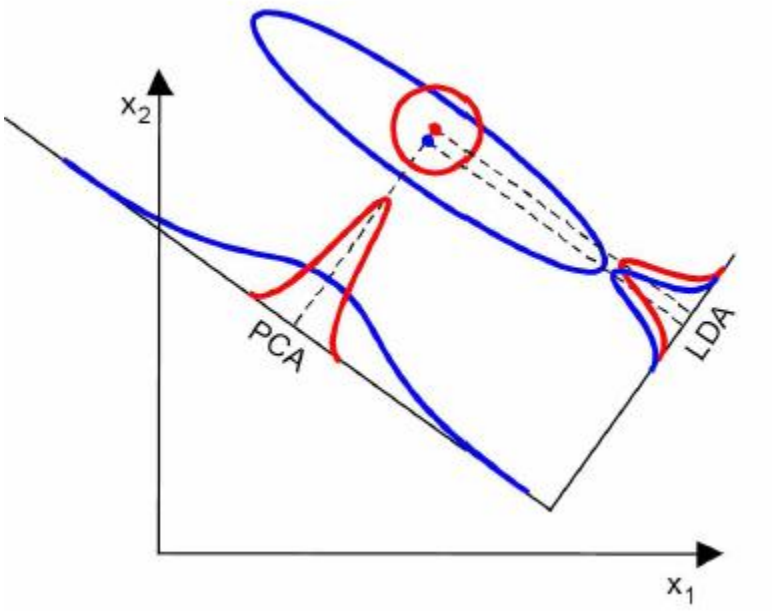
Diskriminační analýza

- o klasifikaci rozhoduje **hustota pravděpodobnosti** dané třídy (nebo vzdálenost od průměru třídy v prostoru transformovaném kovarianční maticí a apriorní pravděpodobností)
- kovarianční matice
  - v LDA (lineární diskriminační analýze) je jedna společná kovarianční matice pro všechny třídy
  - v QDA (kvadratické diskriminační analýze) má každá třída svoji kovarianční matici
- **diskriminační skór** = funkce vyjadřující vzdálenost
- vztahy rozšiřovány o apriorní pravděpodobnost (vypočtená z četnosti v trénovacích datech)

Lineární diskriminační analýza vs. PCA

- PCA mění souřadný systém tak, aby v co nejmenším počtu souřadnic vysvětlila celkový **rozptyl** v datech
- LDA předpokládá stejný rozptyl, maximalizuje rozdíl v **průměrech** jednotlivých tříd





Lineární diskriminační analýza (LDA)

- pro binární problém stejná jako lineární model
- při vícerozměrné klasifikaci zamezuje maskování tříd
- předpokládá vícerozměrné gausovské rozdělení tříd
- společná kovarianční matice pro všechny třídy
- je možné rozšířit prostor vstupních atributů o jejich transformace (mocnina, …)

Klasifikace:

$$\hat{G}(x) = \arg\max_{i=1..k} \partial_k(x) = \arg\max_{i=1..k} \left(x^T \sum^{-1} \mu_i - \frac{1}{2} \mu_i^T \sum^{-1} \mu_i + \ln p_i \right)$$

kde p_i je apriorní pravděpodobnost i-té třídy

LDA – odvození

- odvození vycházející z Bayesova vzorce

$$p(h|D) = p(h)\frac{p(D|h)}{p(D)}$$

$$\left.\begin{matrix}\frac{p(h_1|D)}{p(h_2|D)} > 1 \Rightarrow h_1 \\ \frac{p(h_1|D)}{p(h_2|D)} < 1 \Rightarrow h_2\end{matrix}\right\} \Rightarrow \ln\left(\frac{p(h_1|D)}{p(h_2|D)}\right) \left\{\begin{matrix} > 0 \Rightarrow h_1 \\ < 0 \Rightarrow h_2\end{matrix}\right.$$

- Dále se předpokládá normální rozložení dat $p(D|h)$

$$N(\mu;\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \Rightarrow N(\mu;\Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}$$

LDA – odvození

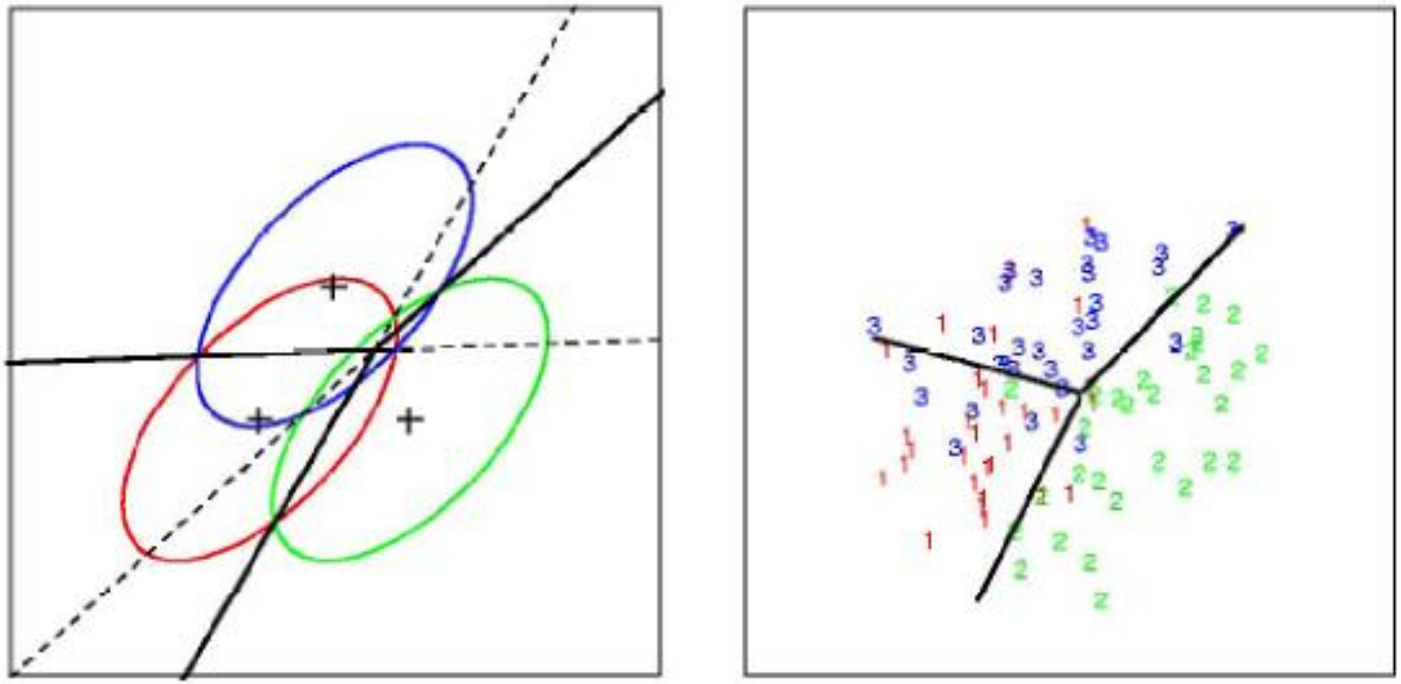
- apriorní pravděpodobnost odvozena z četností $\pi = N_k / N$
- kovarianční matice se vypočte přes všechna data
- pro klasifikaci mezi dvěma třídami lze odvodit:

$$\ln \frac{p(h_1|D)}{p(h_2|D)} = \ln \frac{p(h_1)p(D|h_1)}{p(h_2)p(D|h_2)} = \ln \frac{\pi_1 e^{-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)}}{\pi_2 e^{-\frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2)}} =$$

$$= \ln \frac{\pi_1}{\pi_2} - \frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1) + \frac{1}{2}(x-\mu_2)^T \Sigma^{-1}(x-\mu_2) =$$

$$= \ln \frac{\pi_1}{\pi_2} - \frac{1}{2}(\mu_1+\mu_2)^T \Sigma^{-1}(\mu_1-\mu_2) + x^T \Sigma^{-1}(\mu_1-\mu_2)$$

LDA – rozdělení prostoru



The elements of statistical learning by T. Hastie and col.

Kvadratická diskriminační analýza (QDA)

- pokud odlišné rozložení v jednotlivých třídách – kovarianční matice pro každou třídu

Klasifikace:

$$\hat{G}(x) = \arg\max_{i=1..k} \partial_k(x) =$$

$$\arg\max_{i=1..k} \left(-\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \ln p_i \right)$$

kde p_i je apriorní pravděpodobnost i-té třídy

Příklad

V evidenci je 400 pacientů s vážnou nemocí. U každého je zaznamenána anamnéza a provedeno vyšetření. Pacienti jsou rozděleni do tří skupin. V první skupině jsou ti, kteří zemřeli do 10 dnů po příchodu. Pacienti z druhé skupiny zemřeli po delší době než 10 dnů nebo měli trvalé následky. Třetí skupina je kompletně vyléčená. Následující tabulka ukazuje konkrétní informace získané z těchto údajů. Nový pacient má hodnoty vstupních testů x=(5;7). Do které ze skupin patří? (dle QDA).

třída T_i	T_1	T_2	T_3
N_i	50	50	300
μ_i	(5;5)	(4;8)	(6;9)
Σ_i	$\begin{pmatrix} 3 & 1 \\ 1 & 2 \end{pmatrix}$	$\begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}$	$\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}$

$$\hat{G}(x) = \arg\max_{i=1..k} \left(-\frac{1}{2} \ln|\Sigma_i| - \frac{1}{2} (x-\mu_i)^T \Sigma_i^{-1} (x-\mu_i) + \ln p_i \right)$$

Řešení

$p_i = N_i / N$

$\delta_1(x) = -4{,}08$

$\delta_2(x) = -4{,}58$

$\delta_3(x) = -3{,}17$

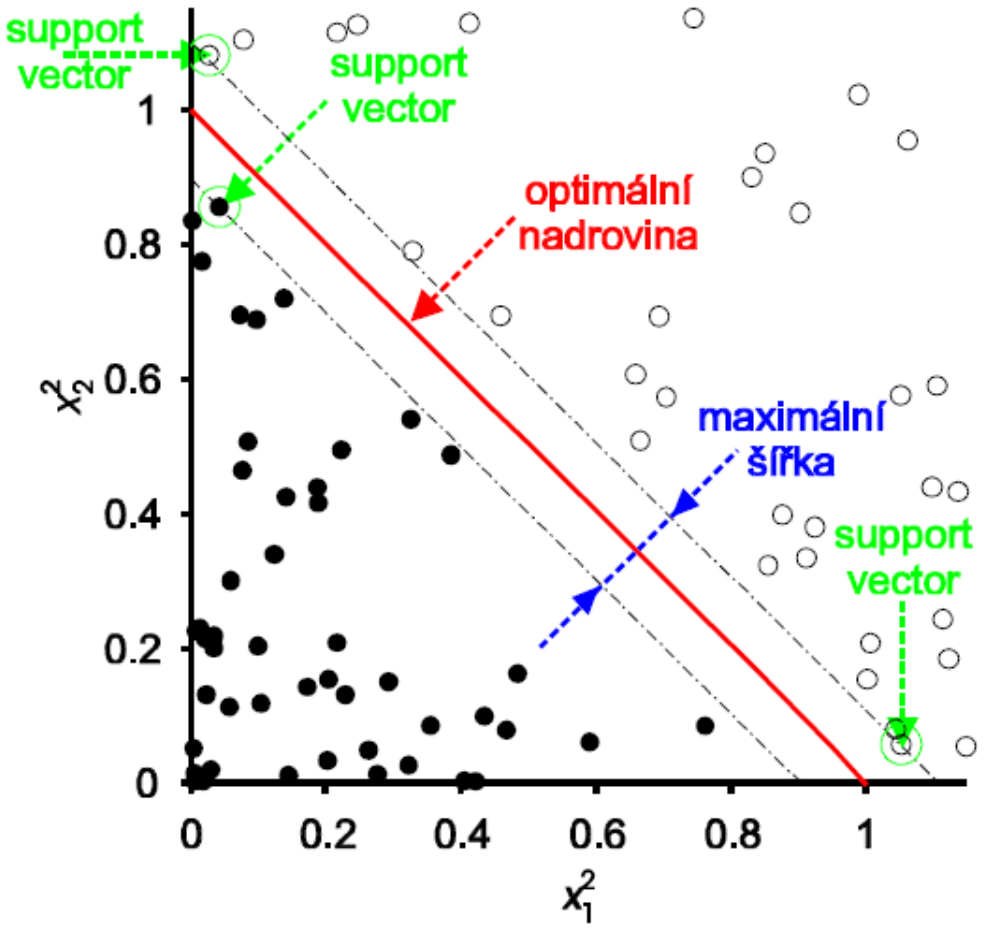
$$\partial_1(x) = -\frac{1}{2}\ln(5) - \frac{1}{2}\begin{pmatrix}0 & 2\end{pmatrix}\frac{1}{5}\begin{pmatrix}2 & -1\\ -1 & 3\end{pmatrix}\begin{pmatrix}0\\ 2\end{pmatrix} + \ln\frac{50}{400}$$

Největší diskriminační skór má poslední (třetí) skupina, pacient má značnou šanci na vyléčení.

Support Vector Machines
Podpůrné vektory

- Jak oddělit následující třídy pomocí lineární rovnice *f(x)=0* ?
  - mapování do nového příznakového prostoru
- Kam umístit hranici?
  - optimální separující nadrovina

SVM – optimální separující nadrovina



Výpočet optimální separující nadroviny

- model má *N* parametrů α, řešením je vektor **nenulových parametrů $\boldsymbol{\alpha}$**
- ty získáme maximalizací účelové funkce:

$$\sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$$

- za podmínek:

$$\alpha_i \geq 0; \qquad \sum_i \alpha_i y_i = 0$$

- k nalezení optima se používá kvadratické programování (obdoba lineárního programování, účelová funkce je však kvadratická v neznámém parametru α)

SVM – nastavený model

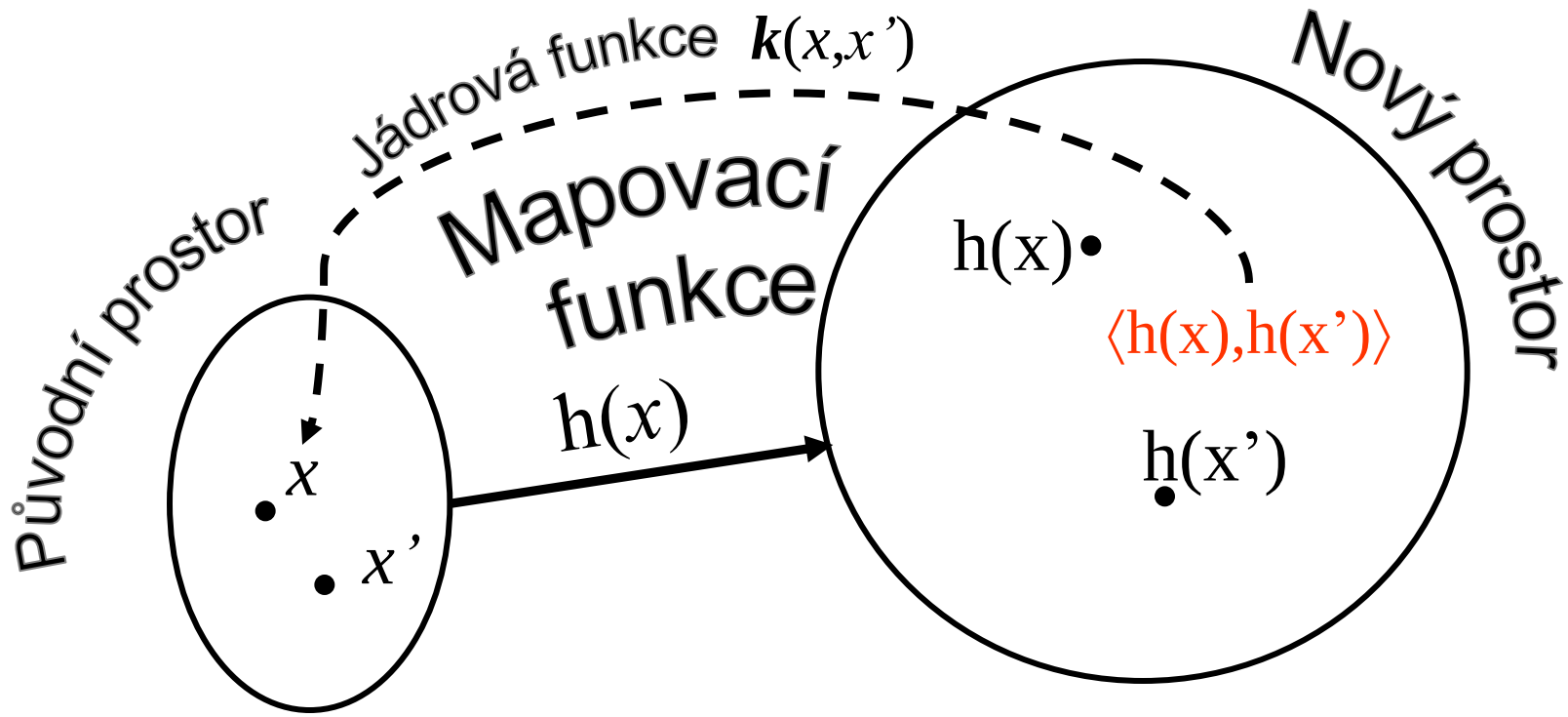
$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i (x \cdot x_i) \right)$$

- N je počet trénovacích prvků
- α_i, váha, nenulová u SV (podpůrné vektory, ty jediné pak má smysl uložit)

SVM – kde je problém

- v účelové funkci a finálním modelu použit skalární součin v původním prostoru atributů ($x.x'$)
- aby bylo možné využít síly SVM, je žádoucí vytvoření celé řady nových atributů (z 5 např. 5.000)
- při transformaci např.(x_1,x_2) na (x_1,x_2,x_1^2,x_2^2,x_1x_2,x_1^3,...) je třeba nejdříve každou instanci přepočítat (rozšířit), až poté lze mezi prvky v novém vícerozměrném prostoru počítat skalární součin – a to stojí spoustu času…
- a takových transformací je třeba vyzkoušet více a hledat tu nejlepší…
- **Lze to zjednodušit?**

SVM – jádrová funkce *k*



$$\langle h(x), h(x') \rangle = k(x,x')$$

SVM – jádrový trik ve vzorci

- Původní vztah pro výpočet optimální separující nadroviny:

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \langle h(x), h(x_i) \rangle\right)$$

- A jeho úprava – skalární součin v novém příznakovém prostoru

$$f(x) = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i k(x, x_i) + b\right)$$

SVM – jádrové funkce

dth Degree polynomial:	$K(x, x') = (1 + \langle x, x' \rangle)^d$
Radial basis	$K(x, x') = exp(\frac{-\|x-x'\|^2}{c})$
Neural network	$K(x, x') = tanh(\kappa_1 \langle x, x' \rangle + \kappa_2)$

d	dimenzí
1	3
2	6
3	10
4	15
5	21
6	28
7	36
8	45
9	55

Př. Polynom stupně 2 na dvourozměrném vstupu:

$K(x, x') = (1 + \langle x, x' \rangle)^2 =$
$(1 + 2x_1x'_1 + 2x_2x'_2 + (x_1x'_1)^2 + (x_2x'_2)^2 + 2x_1x'_1x_2x'_2)$ tj. $M = 6$,
$h_1(x) = 1, h_2(x) = \sqrt{2}x_1, h_3(x) = \sqrt{2}x_2,$
$h_4(x) = x_1^2, h_5(x) = x_2^2, h_6(x) = \sqrt{2}x_1x_2.$

Vypočti skalární součin bodů x=(1;2) a x'=(3;4) A) v původním prostoru a v polynomiálním rozšíření vstupního prostoru při d=2 B) bez C) s kernelovým trikem

A) Skalární součin $\langle x,x' \rangle$ = 1.3+2.4 = 11

B) Skalární součin v 6-rozměrném prostoru (polynomiální jádrová funkce 2. stupně): (1 + 2.1.3 + 2.2.4 + $(1.3)^2$ + $(2.4)^2$ + 2.1.3.2.4) = 1+6+16+9+64+48= **144**

C) Toto lze však vypočíst přímo z jádrové funkce a původního 2-rozměrného prostoru: $K(x,x') = (1+ \langle x,x' \rangle)^2 = (1 + 11)^2 =$ **144**

SVM – příklad

- Kolik bodů (SV = support vectors = podpůrných vektorů) je třeba uložit, aby bylo možno klasifikovat následující prvky?