Internet a klasifikační metody, 4. přednáška

Kdy dělá klasifikátor nejméně chyb na nových datech?

volitelný předmět pro magisterské studium

Martin Holeňa

O čem to bude?

- Schopnost klasifikátoru generalizovat pro nové vstupy
  - souvislost s šířkou pásu mezi třídami
- Generalizující klasifikátory pro 2 lineárně separabilní třídy
  - opěrné vektory a jejich význam
- Zobecnění pro 2 lineárně nesaparabilní třídy
- Zobecnění pro více tříd + další zobecnění

Generalizační schopnost klasifikátoru

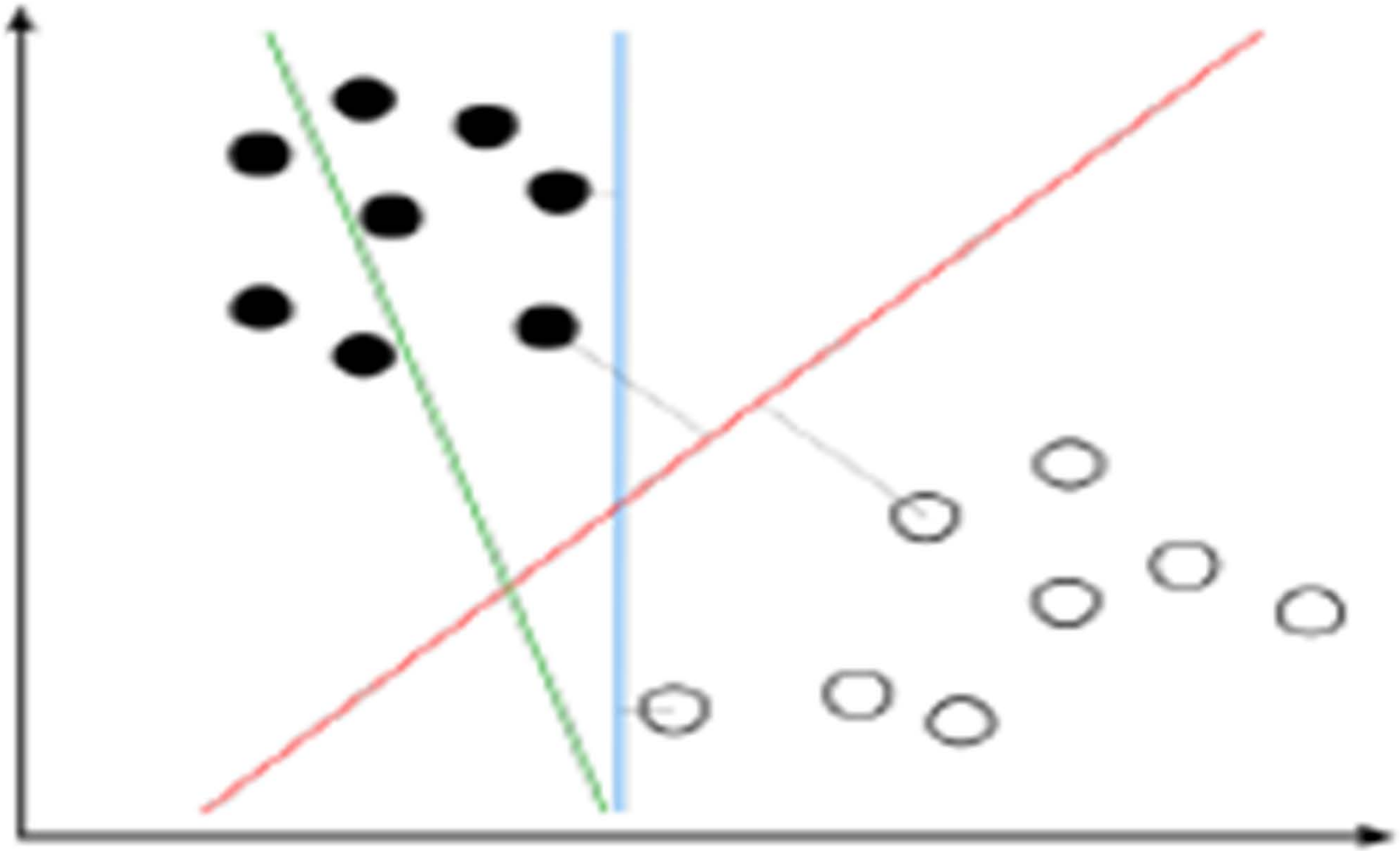
- = Schopnost správně klasifikovat možné nové vstupy
  - při zohlednění pravděpodobnosti jednotlivých nových vstupů
  - $\rightarrow$ formálně nejčastěji E_{vstupy} přesnost $= 1 - E_{vstupy}$ chybovost
- Pro začátek: 2 třídy $C_{\pm} \subset \mathbb{R}^d$, lineárně separabilní, klasifikátor $F: \mathbb{R}^d \rightarrow \{1,-1\}$

$$F(x) = \begin{cases} 1 \text{ při } w^{\intercal}x + b \geq 0 \\ -1 \text{ při } w^{\intercal}x + b < 0 \end{cases}, \; E_{vstupy} \text{ chybovost} = P\left(C_{+} \cap F^{-1}(-1)\right) + P\left(C_{-} \cap F^{-1}(1)\right)$$

Souvislost s pásem mezi třídami

- Předpokládáme, že nadrovina oddělující $C_1 \cap \{x_1, \ldots x_q\}$ a $C_2 \cap \{x_1, \ldots x_q\}$ se

$C_1 \cap \{x_1, \ldots x_q\}$ ani $C_2 \cap \{x_1, \ldots x_q\}$ nedotýká → vzdálena $\rho > 0$ od učicích dat



Souvislost s pásem mezi třídami

- Předpokládáme, že nadrovina oddělující $C_1 \cap \{x_1, \ldots x_q\}$ a $C_2 \cap \{x_1, \ldots x_q\}$ se $C_1 \cap \{x_1, \ldots x_q\}$ ani $C_2 \cap \{x_1, \ldots x_q\}$ nedotýká → vzdálena $\rho > 0$ od učicích dat
- *Domněnka: E_{vstupy} chybovost je nerostoucí funkcí ρ(?)*
  - nezkazí to vliv směru oddělující nadroviny??
- Při omezených w v definici F $((\exists \Lambda > 0)\|w\| \leq \Lambda)$ a $R = \max_{i=1,\ldots,q} \|x_i\|$ lze odvodit: $(\exists \gamma > 0) E_{vstupy} \text{chybovost} < \sqrt{\frac{\gamma}{q}\left(\frac{\Lambda^2 R^2}{\rho^2} \ln^2 q + |\ln P(\text{vstup } x: \|x\| > 0)|\right)}$

Takže maximalizační úloha pro ρ

- Označme $\delta = \rho\|w\|, w$ – směrový vektor nadroviny $H: w^T x + b = 0$
  - $\max \rho = \max \frac{\delta}{\|w\|} \rightarrow \max \delta$ a současně $\min \|w\|$
  - přitom omezení $\rho \leq$ vzdálenost x_i od $H, i = 1, \ldots, q$
- Budeme řešit dvoustupňově, při omezeních pro $i = 1, \ldots, q$

  1. stupeň: fixujeme δ, minimalizujeme $\|w\|$

  2. stupeň: přidáme *maximalizaci δ*

Minimalizační úloha pro $\|w\|$

- Raději minimalizujeme $\|w\|^2$ (má stejné minimum jako $\|w\|$)
  - při častých vyhodnocováních dost ušetříme nepočítáním $\sqrt{\ }$
- Přeformulujeme *omezení* $\rho \leq$ *vzdálenost* x_i *od* $H, i = 1, \ldots, q$:

$$\rho \leq \frac{|w^T x_i + b|}{\|w\|} \rightarrow \frac{\delta}{\|w\|} \leq \frac{|w^T x_i + b|}{\|w\|} \rightarrow \delta \leq |w^T x_i + b| = c_i(w^T x_i + b), c_i \in \{-1,1\}$$

- připomínka: *Lagrangeova funkce* $L = \|w\|^2 + \sum_{i=1}^{q} \alpha_i \underbrace{\left(\delta - c_i(w^T x_i + b)\right)}_{= 0}$

Lagrangeovy koeficienty α_i při $\min\|w\|^2$ splňují KKT podmínky:

Řešení Lagrangeovy funkce

- K $(\hat{w},\hat{b}) = \arg\min\|w\|^2$ splňujícím $\delta \le c_i(\hat{w}^\top x_i + \hat{b})$ existují $\hat{\alpha}_1,\ldots,\hat{\alpha}_q$ že

  1. L má v $(\hat{w},\hat{b},\hat{\alpha}_1,\ldots,\hat{\alpha}_q)$ *minimum vůči* (w,b), *maximum vůči* $(\alpha_1,\ldots,\alpha_q)$

  2. ve $(\hat{w},\hat{b},\hat{\alpha}_1,\ldots,\hat{\alpha}_q)$ platí *KKT podmínky*, $\hat{\alpha}_i\left(\delta - c_i(\hat{w}^\top x_i + \hat{b})\right) = 0$, $i = 1,\ldots,q$

- Z $\frac{\partial L}{\partial w}, \frac{\partial L}{\partial b}(\hat{w},\hat{b},\hat{\alpha}_1,\ldots,\hat{\alpha}_q) = 0$ vychází $\hat{w} = \sum_{i=1}^{q} \frac{\hat{\alpha}_i}{2} c_i\, x_i, \sum_{i=1}^{q} \hat{\alpha}_i c_i = 0$

  - zjednodušíme přeznačením $\hat{\alpha}_i^{\text{new}} = \frac{1}{2}\hat{\alpha}_i^{\text{old}} \rightarrow \hat{w} = \sum_{i=1}^{q} \hat{\alpha}_i c_i\, x_i, \sum_{i=1}^{q} \hat{\alpha}_i c_i = 0$
  - $\rightarrow \|\hat{w}\|^2 = \hat{w}^\top \hat{w} = \sum_{i,j=1}^{q} \hat{\alpha}_i \hat{\alpha}_j c_i c_j x_i^\top x_j$, $L = -\sum_{i,j=1}^{q} \hat{\alpha}_i \hat{\alpha}_j c_i c_j x_i^\top x_j + 2\delta \sum_{i=1}^{q} \hat{\alpha}_i$

Výsledek: kvadratická optimalizace

- L má v $(\hat{w}, \hat{b}, \hat{\alpha}_1, \ldots, \hat{\alpha}_q)$ maximum vůči $(\alpha_1, \ldots, \alpha_q) \rightarrow (\hat{\alpha}_1, \ldots, \hat{\alpha}_q)$ je řešením

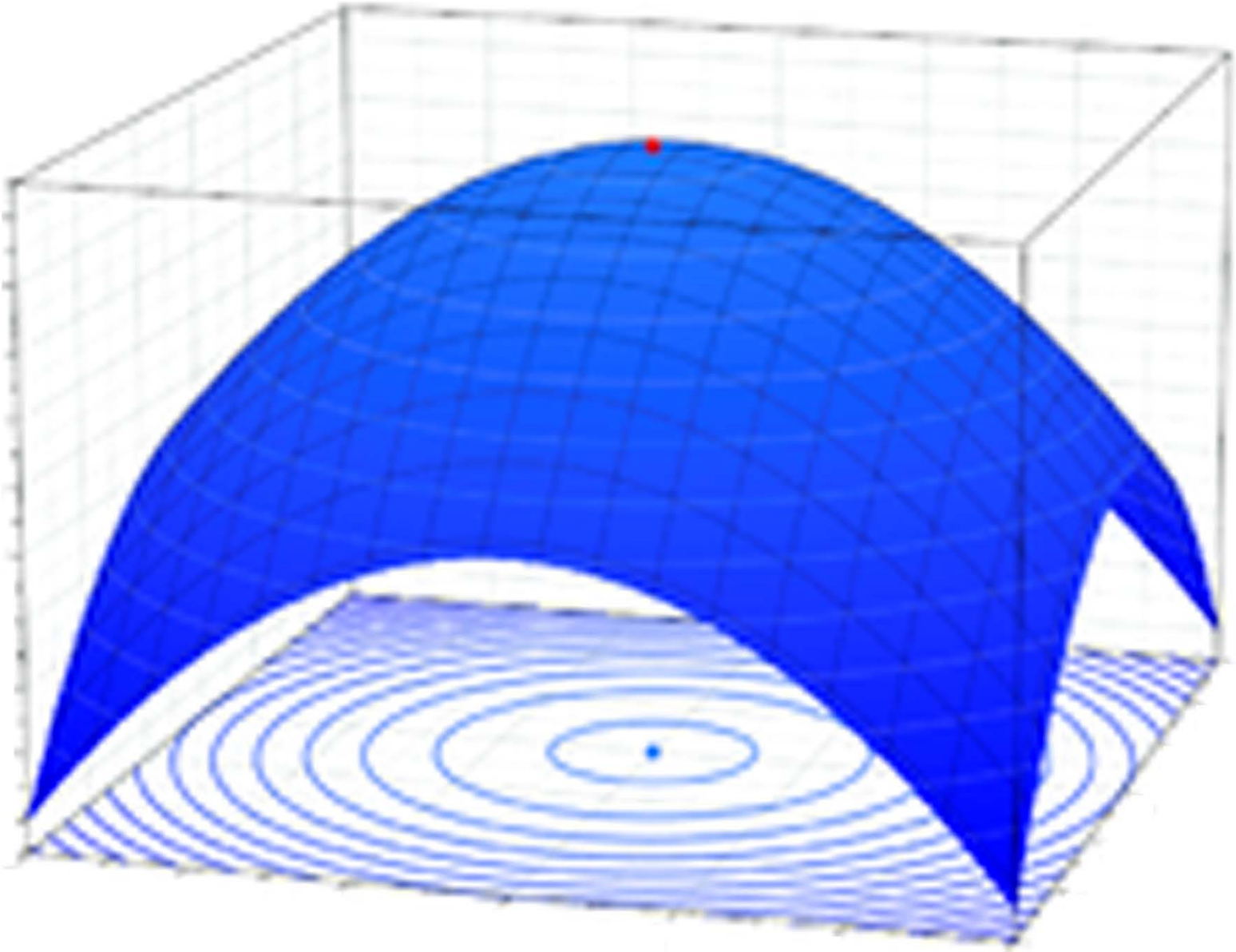
  úlohy $2\delta \sum_{i=1}^{q} \hat{\alpha}_i - \sum_{i,j=1}^{q} \hat{\alpha}_i \hat{\alpha}_j c_i c_j x_i^{\mathsf{T}} x_j = \max_{\alpha_1,\ldots,\alpha_q} 2\delta \sum_{i=1}^{q} \alpha_i - \sum_{i,j=1}^{q} \alpha_i \alpha_j c_i c_j x_i^{\mathsf{T}} x_j$

  při omezeních $\sum_{i=1}^{q} \alpha_i c_i = 0$ a KKT: $\hat{\alpha}_i \left(\delta - c_i \left(\hat{w}^{\mathsf{T}} x_i + \hat{b} \right) \right) = 0$ – *duální úloha*

- \+ *přidáme* 2. stupeň maximalizační úlohy pro ρ: *maximalizaci* δ

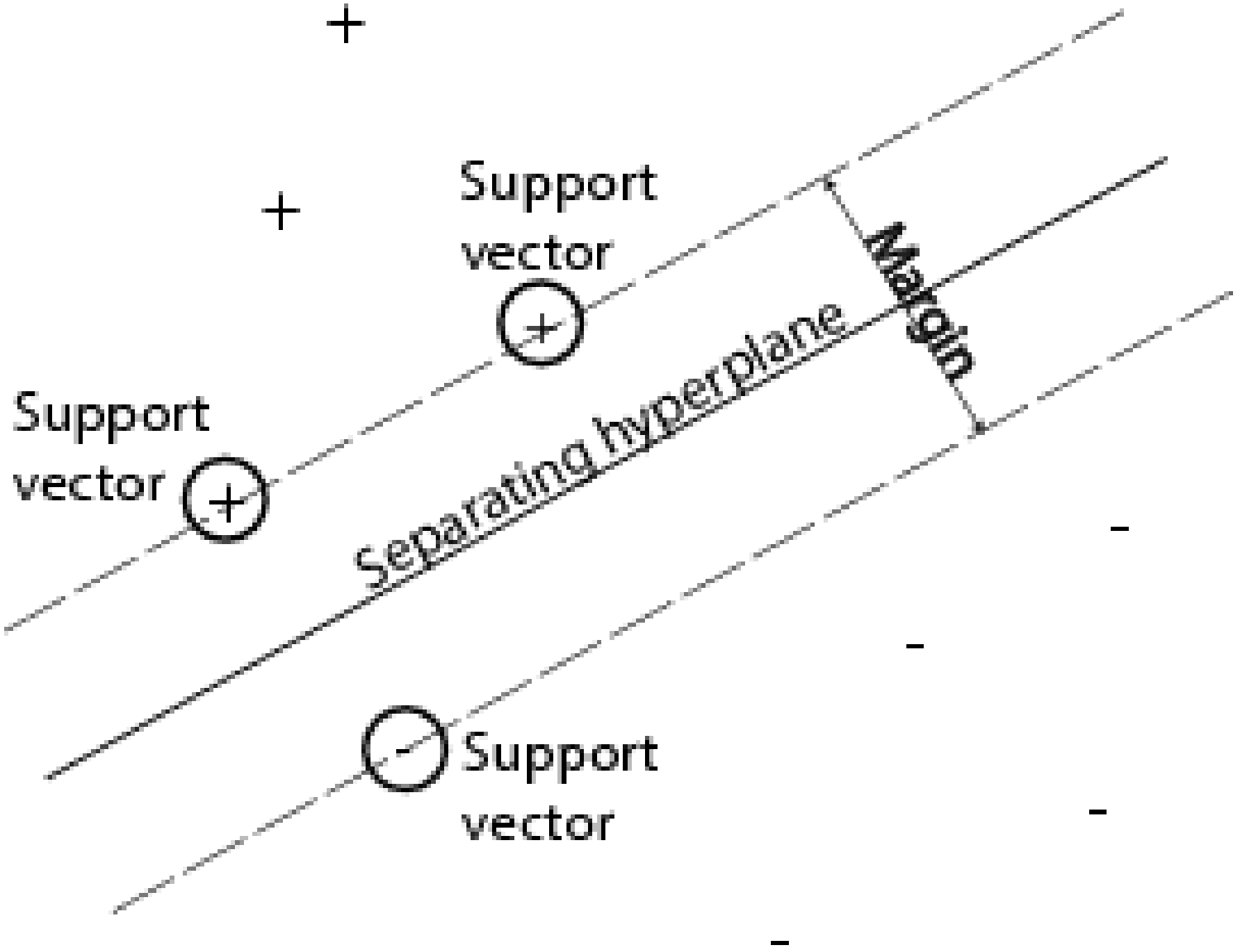
- → *kvadratická optimalizace*: 1 lokální optimum = globální

  - snadno vyřeší každý optimalizační software



Opěrné vektory

- Získané $\hat{\alpha}_1,\ldots,\hat{\alpha}_q$ dosadíme do $\hat{w}=\sum_{i=1}^{q}\hat{\alpha}_i c_i x_i$; uvědomme si, že:
  - $\hat{w}=\sum_{i\in S}\hat{\alpha}_i c_i x_i$, kde $S=\{i:\hat{\alpha}_i>0\}$, takže pro $i\in S, \delta-c_i\left(\hat{w}^{\mathrm{T}}x_i+\hat{b}\right)=0$ díky KKT
  - pomocí získaného $\hat{\delta}=\max\delta$ pro libovolné $i\in S$ dopočítáme $\hat{b}=c_i\hat{\delta}-\hat{w}^{\mathrm{T}}x_i$
- $\rightarrow$ Vektory $x_i, i\in S$ leží v nadrovině $H_+: \hat{w}^{\mathrm{T}}x_i+\hat{b}+\hat{\delta}$ nebo $H_-: \hat{w}^{\mathrm{T}}x_i+\hat{b}-\hat{\delta}$
  - názvy: $H_+|H_-$ – *opěrné nadroviny*, $C_+|C_-\cap\{x_1,\ldots x_q\}$, $x_i, i\in S$ – opěrné vektory
  - $\rightarrow$ označení metody: *Support Vector Machine (SVM)*


Support
vector
Support
vector
Separating hyperplane
Margin
Support
vector

Rekapitulace SVM algoritmu

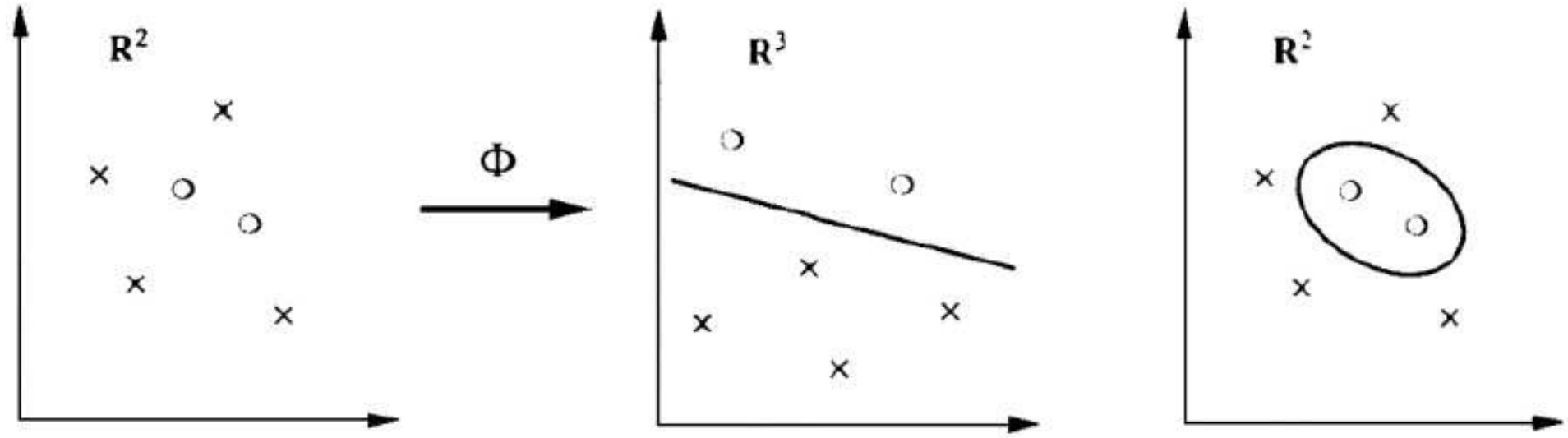
- Vstup: učicí data $(x_1, c_1), \ldots, (x_q, c_q)$

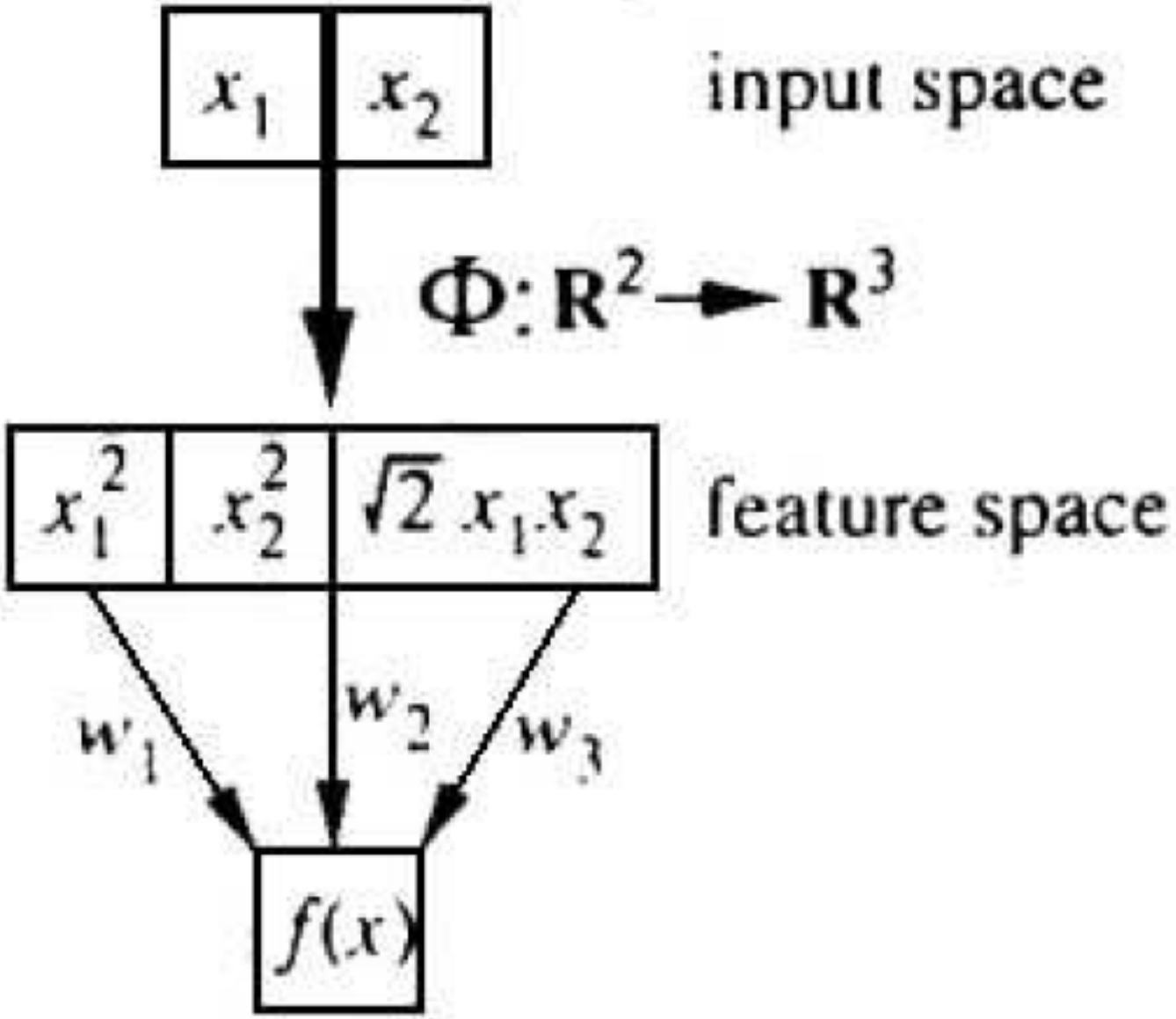
1. $(\hat{\alpha}_1, \ldots, \hat{\alpha}_q, \hat{\delta}) = \arg\max_{\alpha_1,\ldots,\alpha_q} 2\delta \sum_{i=1}^{q} \alpha_i - \sum_{i,j=1}^{q} \alpha_i \alpha_j c_i c_j x_i^\top x_j$ při $\sum_{i=1}^{q} \alpha_i c_i = 0$ a KKT
2. $S = \{i : \hat{\alpha}_i > 0\}$ a vybereme nějaké $i_b \in S$
3. $\hat{w} = \sum_{i \in S} \hat{\alpha}_i c_i x_i$, $\hat{b} = c_{i_b} \hat{\delta} - \hat{w}^\top x_{i_b}$

- Výstup: klasifikátor F, $F(x) = \begin{cases} 1 \text{ při } \sum_{i \in S} \hat{\alpha}_i c_i x^\top x_i + \hat{b} \geq 0 \\ -1 \text{ při } \sum_{i \in S} \hat{\alpha}_i c_i x^\top x_i + \hat{b} < 0 \end{cases}$

Připomínka: jak k lineární separabilitě?

- Transformací do prostoru *vyšší dimenze*


R2
Φ
R3
R2



$$f(x)=\mathrm{sgn}(w_1 x_1^2+w_2 x_2^2+w_3\sqrt{2}\, x_1 x_2+b)$$

Připomínka: jak k lineární separabilitě?

- Transformací do prostoru *vyšší dimenze*
- Metoda pomocí jádra: zobrazení $\kappa: X^2 \to \mathcal{R}$, $\kappa(x,y) = \kappa(y,x)$,

  s každou $\begin{pmatrix} \kappa(x_i,x_i) & \cdots & \kappa(x_i,x_j) \\ \vdots & \ddots & \vdots \\ \kappa(x_j,x_i) & \ldots & \kappa(x_j,x_j) \end{pmatrix}$ pozitivně semidefinitní

  - q-dimenzionální množina $M = \{\varphi_1, \ldots, \varphi_q\}, \varphi_i: X \to R, \varphi_i(x) = \kappa(x_i, x)$
  - vektorový prostor $V = \left\{\sum_{i=1}^{q} a_i \varphi_i \,|a_1, \ldots, a_q \in \mathcal{R}\right\}$ = lineární obal M
  - skalární součin na V: $\left\langle \sum_{i=1}^{q} a_i \varphi_i , \sum_{i=1}^{q} b_i \varphi_i \right\rangle = \sum_{i,j=1}^{q} a_i b_j \kappa(x_i, x_j)$

Pro SVM se hodí již známý trik

- Připomínka: *kernel trick* – počítáme skalární součin

  v prostoru *V dimenze* q pomocí $\kappa(x_i, x_j)$, tj. pomocí $x_i, x_j \in \mathbb{R}^d$

  - q – počet dat → většinou $d \ll q$

- Vhodnost pro SVM: skalárním součinem klasifikátor F

kompletně definujeme: $F(x) = \begin{cases} 1 \text{ při } \sum_{i \in S} \hat{\alpha}_i c_i \kappa(x, x_i) + \hat{b} \geq 0 \\ -1 \text{ při } \sum_{i \in S} \hat{\alpha}_i c_i \kappa(x, x_i) + \hat{b} < 0 \end{cases}$

SVM pro vícetřídní klasifikaci

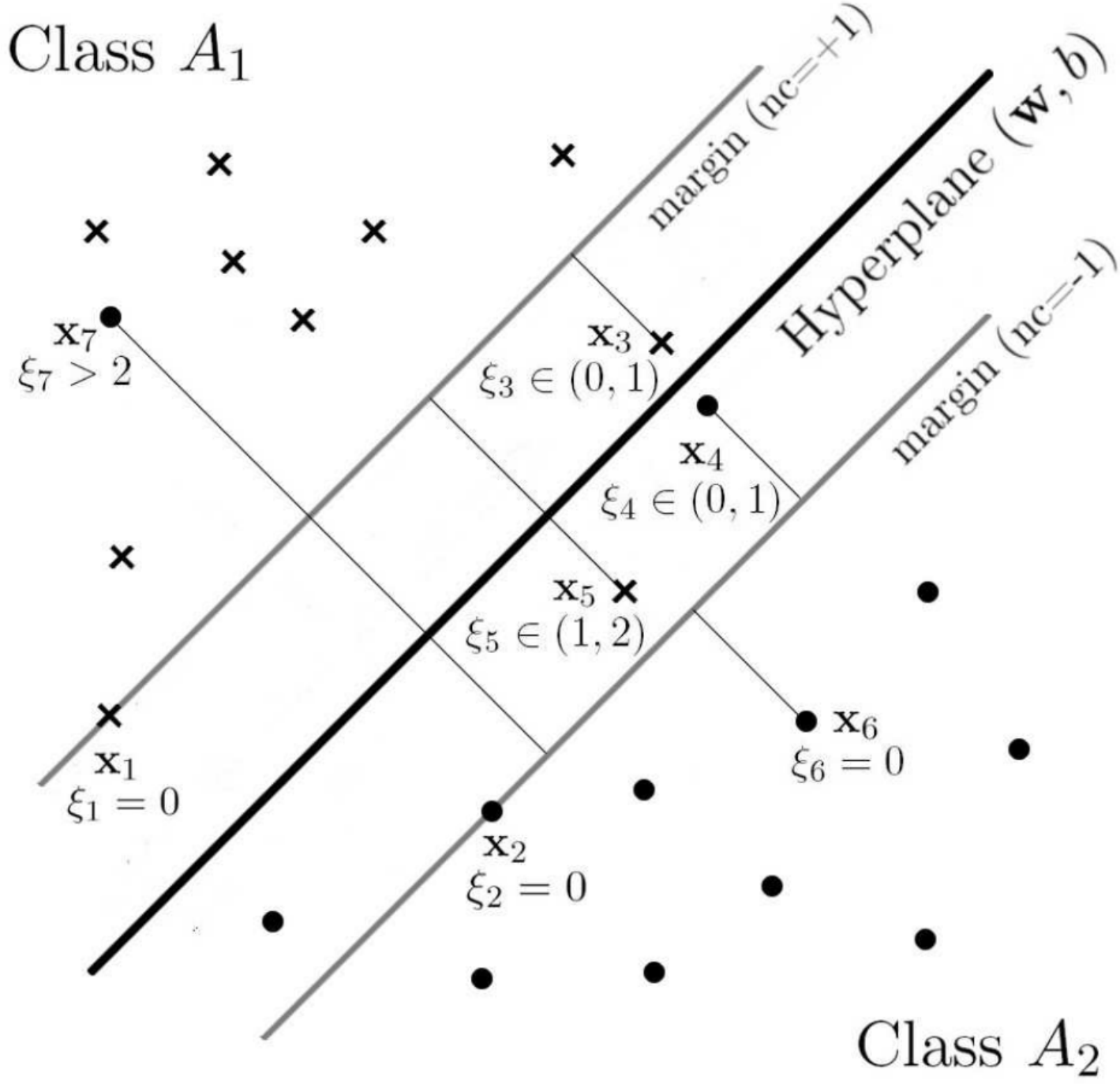
- Výrazně nejpoužívanější přístup: pomocí více 2-třídních

1. *Každá* ze tříd $C_1, \ldots, C_m$ *proti každé* ($\frac{m(m-1)}{2}$ klasifikací)
   - výsledek: C_k s nejvíce z $m-1$ možných vítězství
2. *Každá* třída C_k *proti sjednocení* $\bigcup_{j \neq k} C_j$ zbývajících
   - $c_{i,k} \in \{1, -1\}$ volíme tak, aby $c_{i,k} = 1 \leftrightarrow x_i \in C_k, c_{i,k} = -1 \leftrightarrow x_i \in \bigcup_{j \neq k} C_j$
   - výsledek: C_k s $k = \arg \max_{j=1,\ldots m} \sum_{i \in S} \hat{\alpha}_{i,k} c_{i,k} \kappa(x, x_i) + \hat{b}$

SVM s tolerujícími omezeními

- Místo omezení $\rho \leq$ vzdálenost x_i od $H \rightarrow \delta \leq c_i\left(w^{\mathsf{T}} x_i + b\right)$

  uvažujeme $\rho - \xi_i \leq$ vzdálenost x_i od $H \rightarrow \delta - \xi_i \leq c_i\left(w^{\mathsf{T}} x_i + b\right), \xi_i \geq 0$

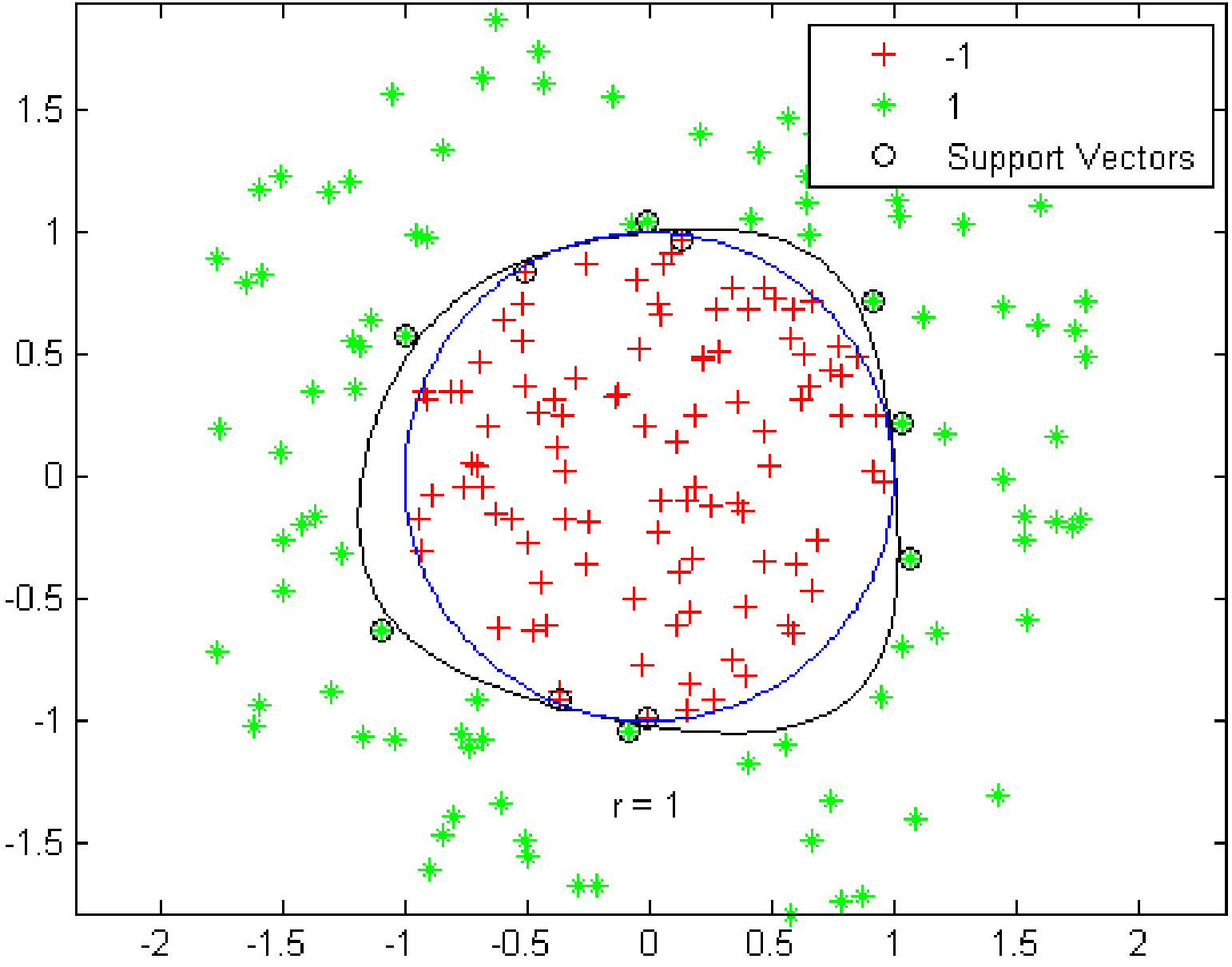

Class A_1
margin (nc=+1)
Hyperplane $(\mathbf{w}, b)$
margin (nc=-1)
$\mathbf{x}_7$
$\xi_7 > 2$
$\mathbf{x}_3$
$\xi_3 \in (0,1)$
$\mathbf{x}_4$
$\xi_4 \in (0,1)$
$\mathbf{x}_5$
$\xi_5 \in (1,2)$
$\mathbf{x}_1$
$\xi_1 = 0$
$\mathbf{x}_6$
$\xi_6 = 0$
$\mathbf{x}_2$
$\xi_2 = 0$
Class A_2

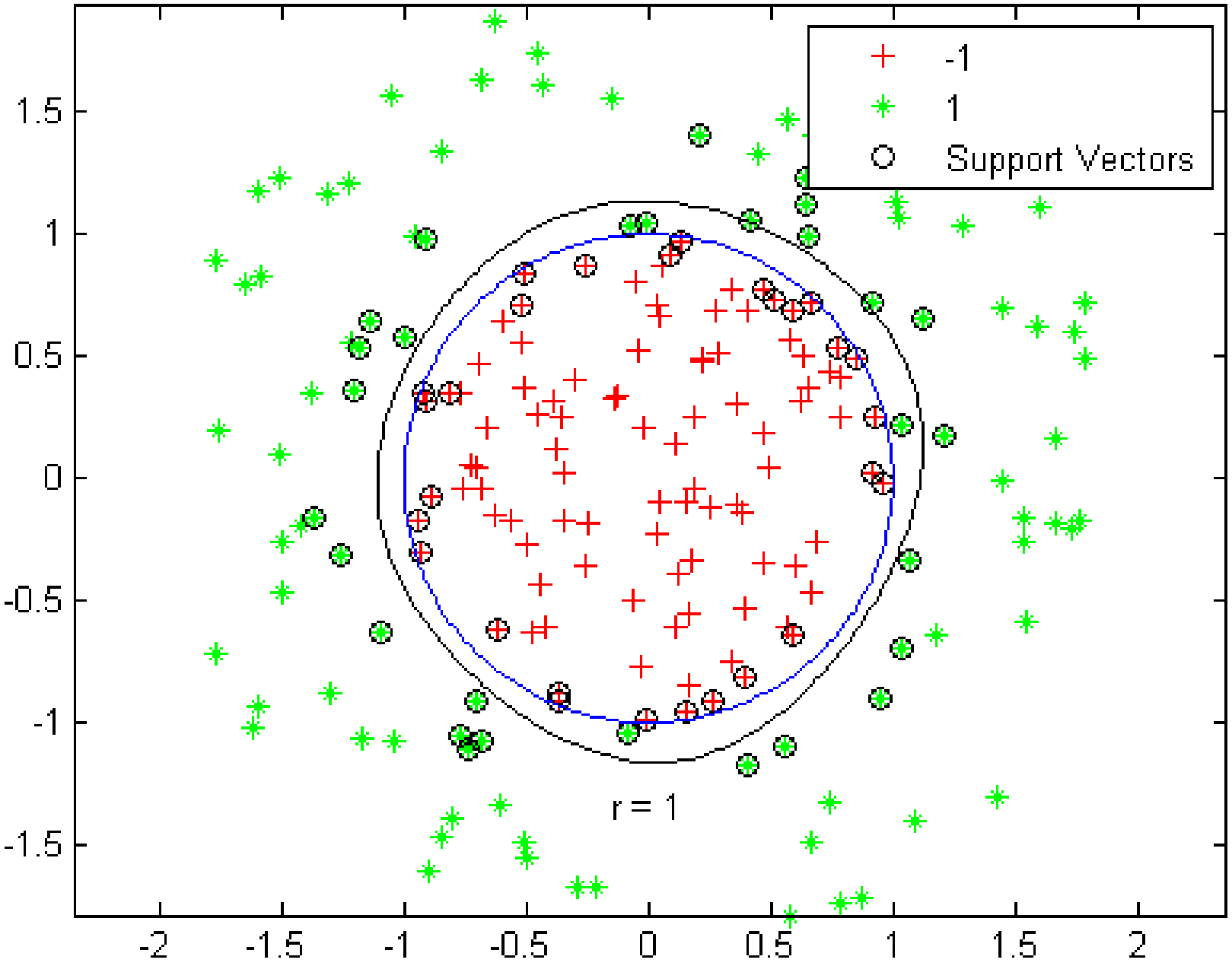
SVM s tolerujícími omezeními

- Místo omezení $\rho \leq$ vzdálenost x_i od $H \rightarrow \delta \leq c_i\left(w^\intercal x_i + b\right)$

  uvažujeme $\rho - \xi_i \leq$ vzdálenost x_i od $H \rightarrow \delta - \xi_i \leq c_i\left(w^\intercal x_i + b\right), \xi_i \geq 0$

  - význam: menší vliv šumu v datech


-1
1
Support Vectors
r = 1
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2


-1
1
Support Vectors
r = 1
1.5
1
0.5
0
-0.5
-1
-1.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

SVM s tolerujícími omezeními

- Místo omezení $\rho \leq$ vzdálenost x_i od $H \rightarrow \delta \leq c_i\left(w^{\mathsf{T}} x_i + b\right)$

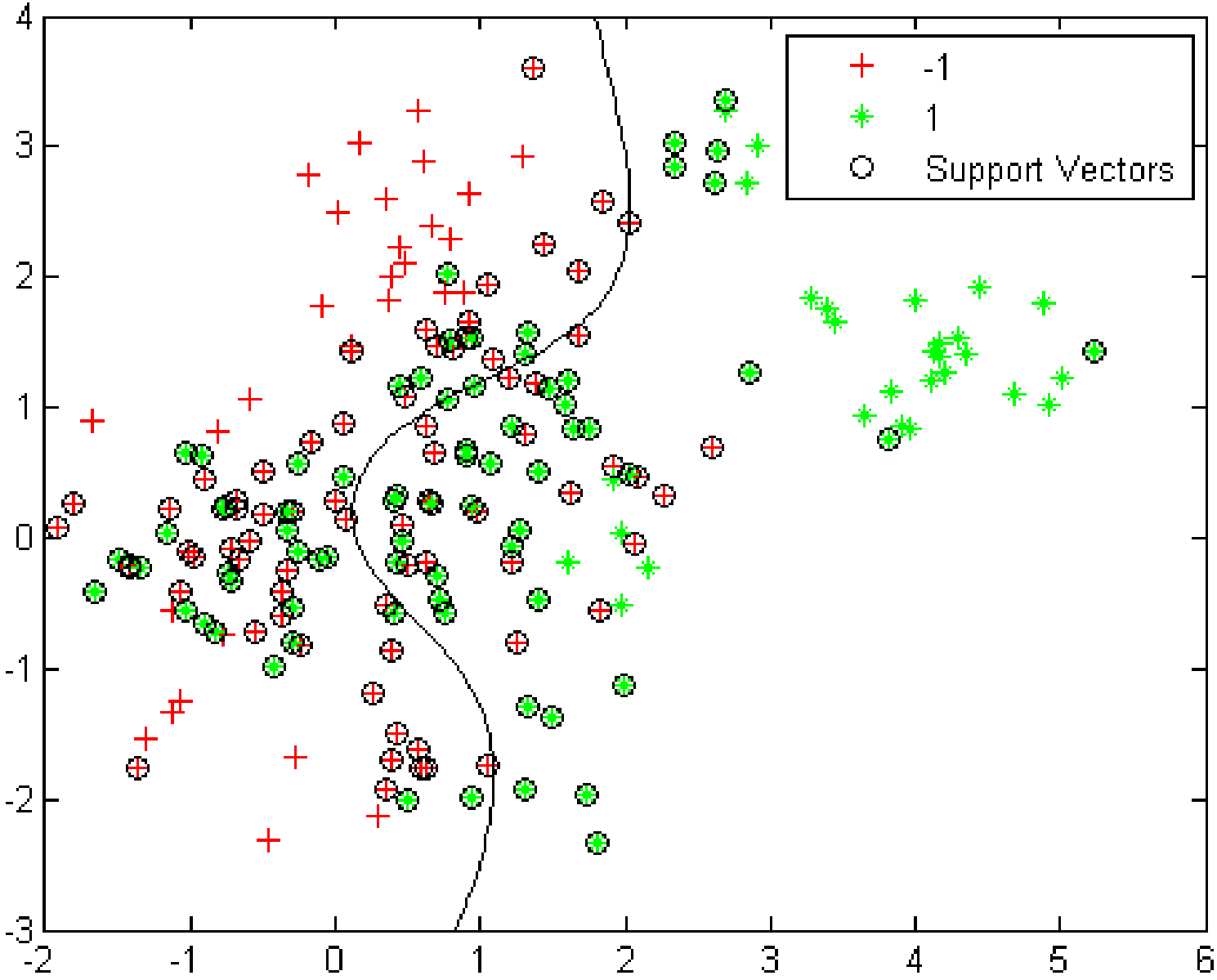
  uvažujeme $\rho - \xi_i \leq$ vzdálenost x_i od $H \rightarrow \delta - \xi_i \leq c_i\left(w^{\mathsf{T}} x_i + b\right), \xi_i \geq 0$

  - význam: menší vliv šumu v datech

- Snaha udržet *tolerující (slack) proměnné* ξ_i malé →

  v 1. stupni maximalizace pásu *minimalizujeme* $\|\mathrm{w}\|^2 + C \sum_{i=1}^{q} \xi_i,\ C > 0$

  - C – poměr mezi tolerancí ξ_i a původní minimalizací $\|\mathrm{w}\|^2$ (~ 0)


-1
1
Support Vectors
4
3
2
1
0
-1
-2
-3
-2
-1
0
1
2
3
4
5
6

SVM při filtraci spamu

- Výhoda: lepší *generalizační schopnost*
  - často využívány specificky pro *obrázkový spam*
  - + příbuzný problém odfiltrování pornografických webových stránek
- Nevýhoda: velká *výpočetní náročnost*
  - kvůli řešení optimalizačního problému
  - protiopatření: paralelní a distribuované implementace

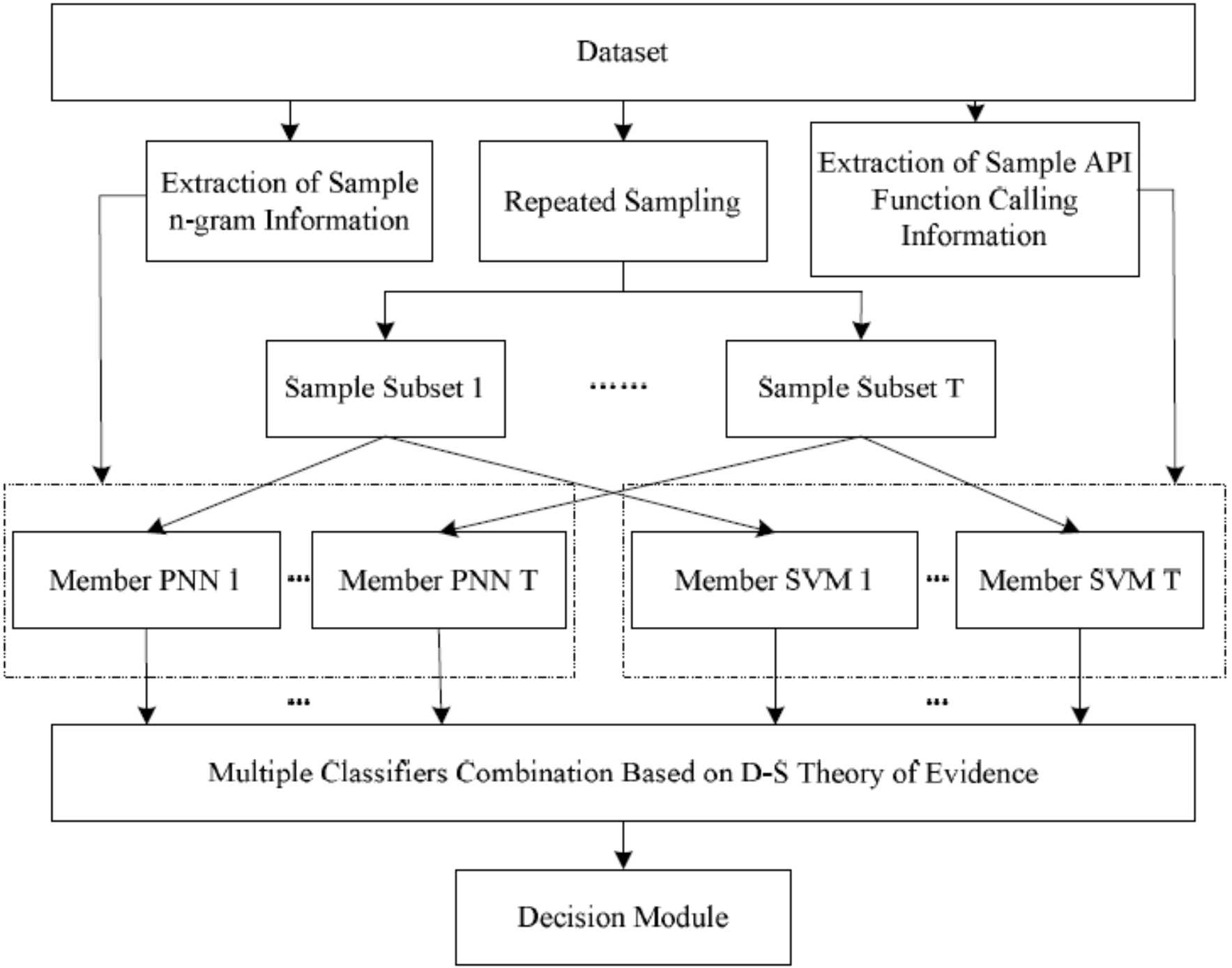
SVM v doporučovacích systémech

- Obsahové filtrování, hlavně obrázků a videí
  - *obrázky*: přiřazování částí obrázku *sémantickým konceptům*
  - *videa*: rozpoznávání *událostí*, slov *mluveného* jazyka
- Nelineární, kromě mluveného jazyka (← vysocedimenzionální příznaky)
- Vícetřídní, každá třída proti sjednocení zbývajících,

  kromě sémantických konceptů: každá třída proti každé

SVM při detekci malware

- Používají *dynamické vlastnosti* software = průběh interakcí s OS – kódován posloupností volání *API*
  - kombinovány s klasifikátorem používajícím statické vlastnosti
- *Nelineární*, používají se s různými kernely
- Buď *2-třídní*: neškodný software × malware, nebo *vícetřídní*: neškodný software × různé druhy malware


Dataset
Extraction of Sample n-gram Information
Repeated Sampling
Extraction of Sample API Function Calling Information
Sample Subset 1
.....
Sample Subset T
Member PNN 1
...
Member PNN T
Member SVM 1
...
Member SVM T
...
...
Multiple Classifiers Combination Based on D-S Theory of Evidence
Decision Module

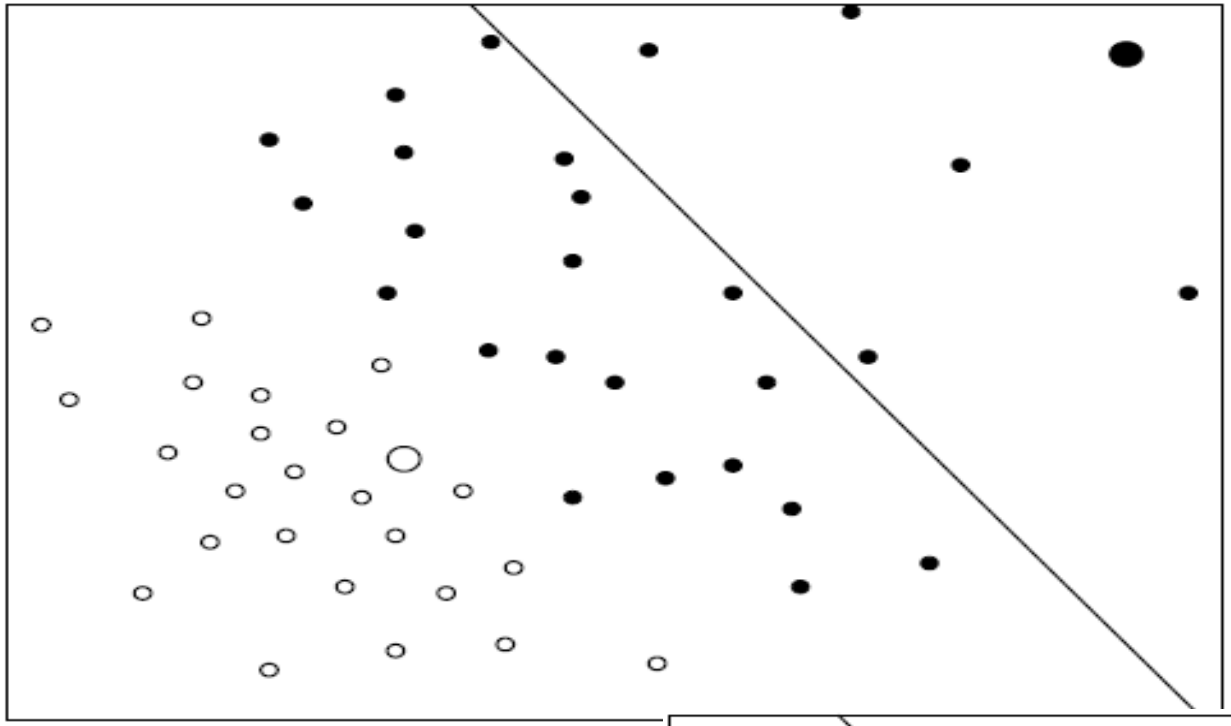
Aktivně se učící SVM

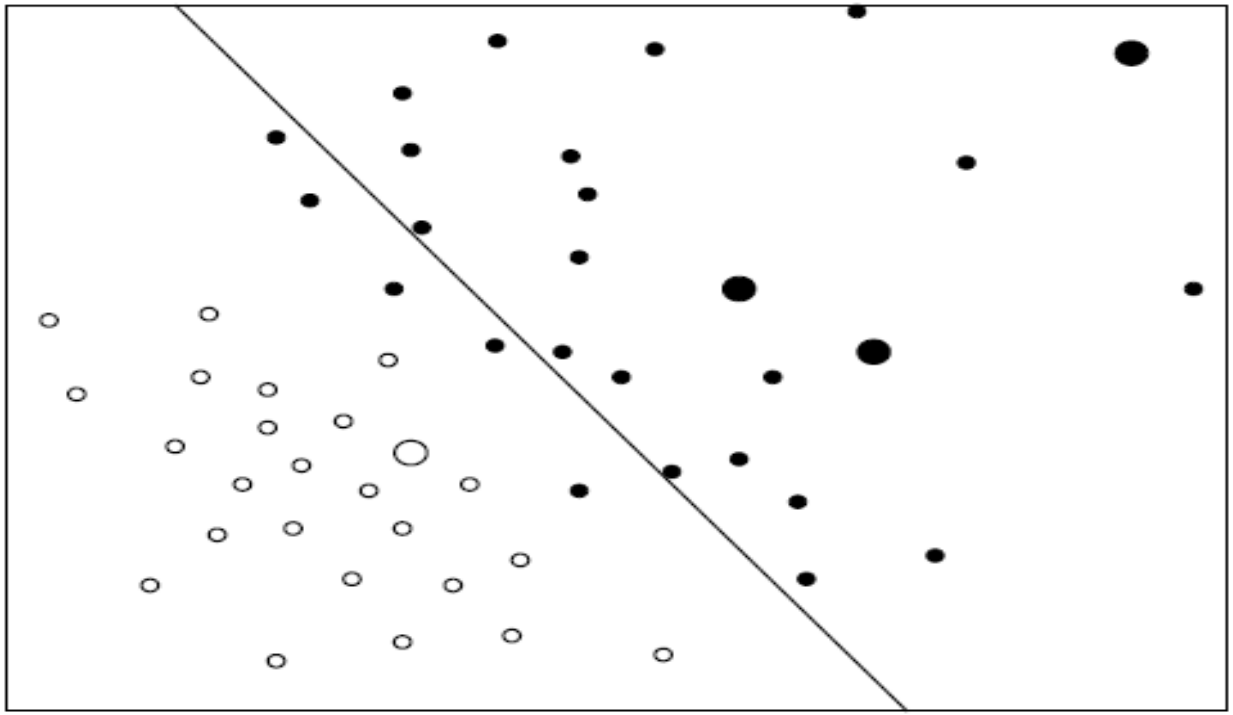
- *Aktivní učení* klasifikátoru: učení, při kterém

1. část učicích dat dodávána na vyžádání
   - obvykle x_i od začátku, od uživatele *vyžádáno* $c_i \in \{-1,1\}$
2. výběr vyžádaných určený *očekávaným zlepšením kvality*

- Použitelné pro libovolný klasifikátor; nejrelevantnější SVM ←

← očekávané zlepšení počítáno přes nové vstupy

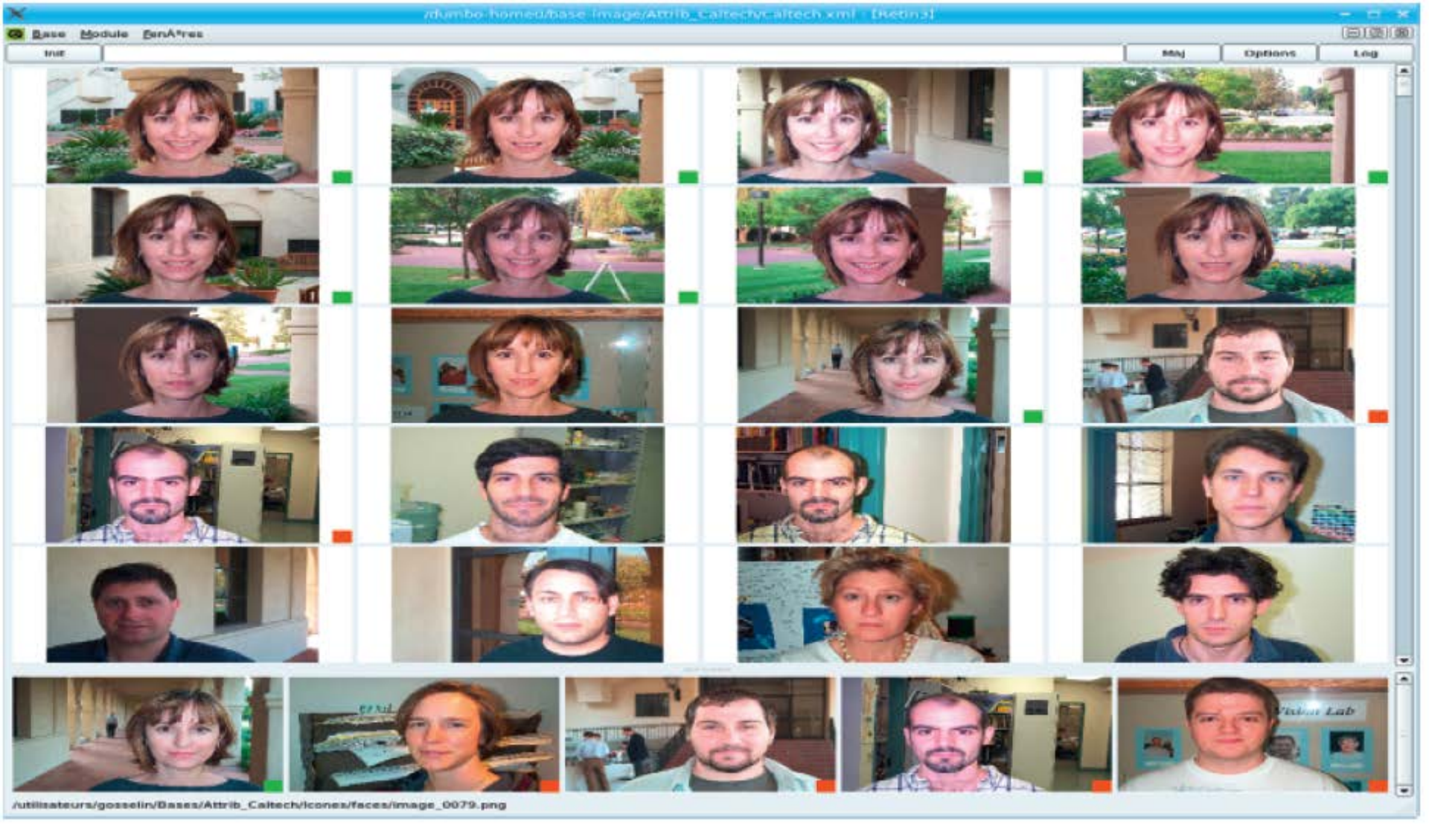




Internet a SVM s aktivním učením

- Doporučovací systémy: *obsahové filtrování* podle *příkladů*
  - velké množství možností → uživatel požádán ohodnotit

    příklady nejvíce pomáhající zpřesnit jeho přání


/dumbo/homed/base/image/Attrib_caltech/caltech.xml - [Retin3]
Base Module BenÂ°res
Init
Maj
Options
Log
/utilisateurs/gosselin/Bases/Attrib_Caltech/icones/faces/image_0079.png

Internet a SVM s aktivním učením

- Doporučovací systémy: *obsahové filtrování* podle *příkladů*
  - velké množství možností → uživatel požádán ohodnotit

    příklady nejvíce pomáhající zpřesnit jeho přání
- Detekce *malware* nebo *síťových útoků*
  - množství nového software | záznamů síťového provozu →

    → aktivní učení: kde nejpotřebnější klasifikace člověkem