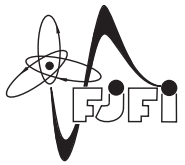
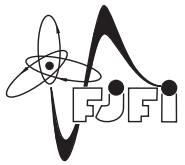
Učící se algoritmy

Michal Odstrčil

Mariánská 2011



Obsah



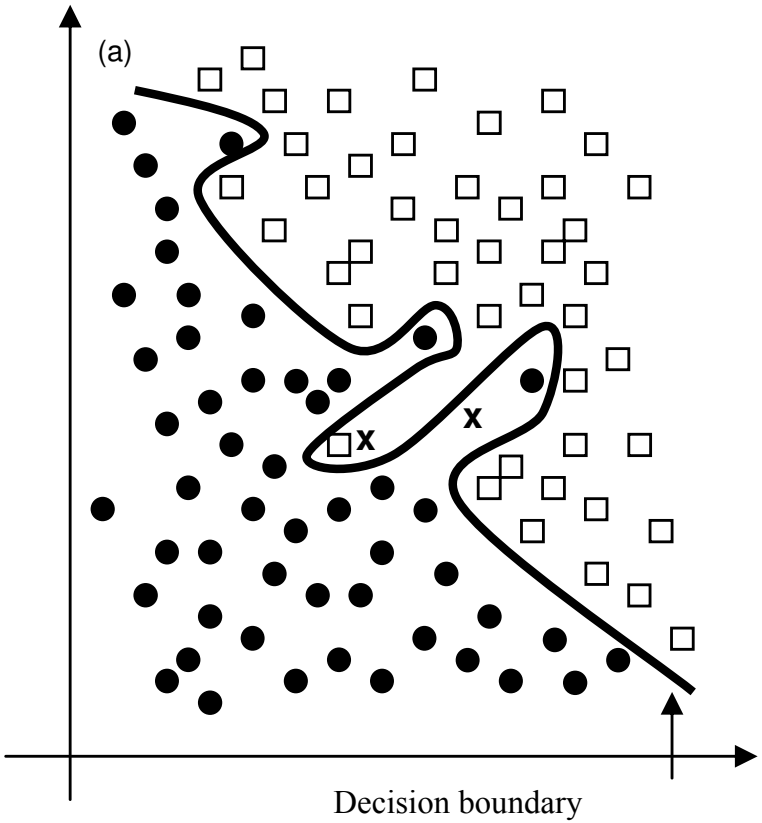
Základní pojmy

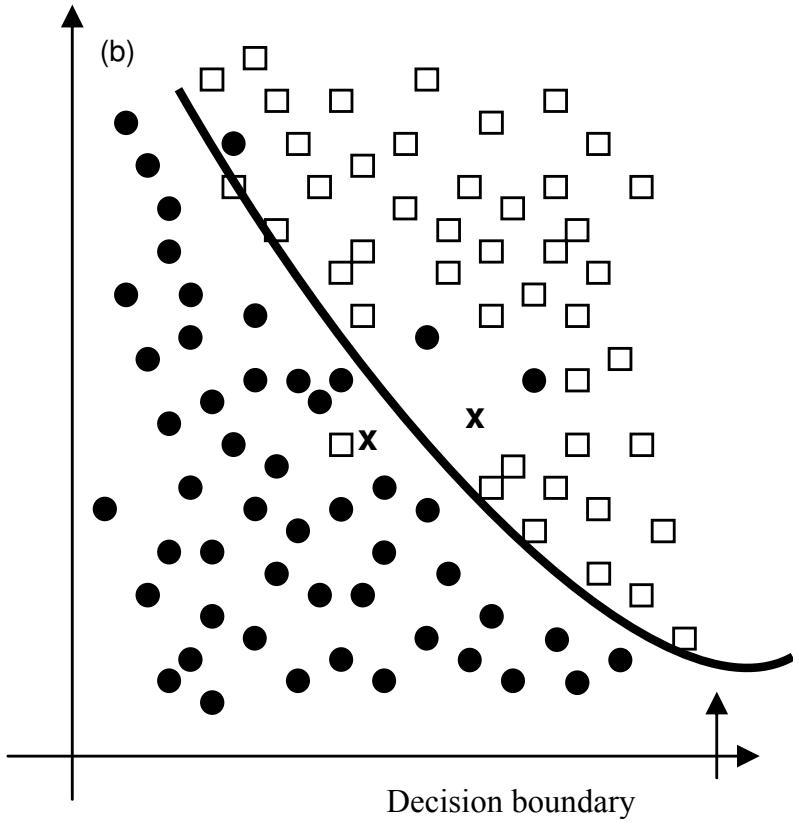
- Špatně podmíněný problém
- Klasifikace vs. regrese
- Overfitting
  - Crossvalidation
  - Soft margins
  - Bayesovský přístup

$$p(\vec{w}|C) = \frac{p(\vec{w}|C)p(\vec{w})}{p(\vec{C})}$$

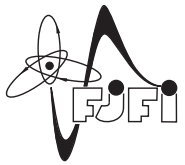
- Supervised/unsupervised learning

Overfitting - Crossvalidation - Soft margins





Outline



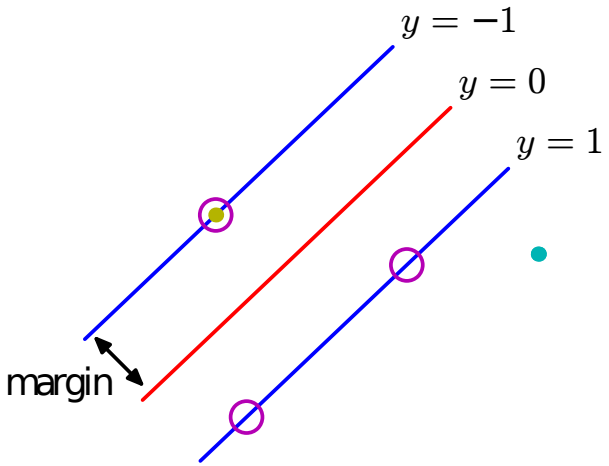
SVM - Support Vector Machine

Princip

- Nejlepší nadplocha oddělující dvě skupiny dat
- Maximalizace okraje
- Lineárně separabilní v transformovaném prostoru
- Sparse kernel method

Parametry:

- C - penalizace „vzdálených pozorování“ (outliers)
- sigma - očekávaná blízkost množin
- Volba jádra



Zdroj [1]

SVM - Support Vector Machine

Výhody

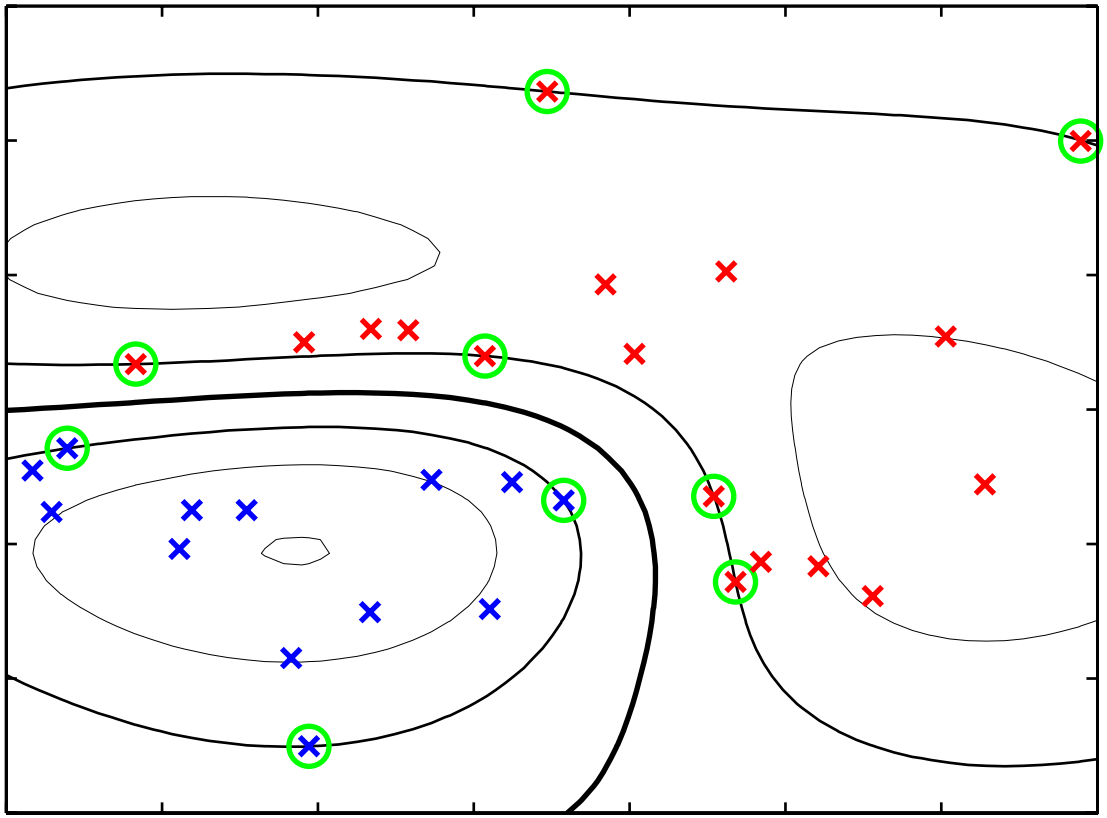
- Jednoduchý princip
- Rychlá klasifikace
- Množství opensource algoritmů na internetu

Nevýhody

- Pomalejší učící část
- Složitost roste s N^3 (až N^2)
- Nelze určit pravděpodobnost správné klasifikace
- Pouze dvě skupiny dat
- Složitost klasifikace obecně $\approx N$
  $y(\vec{x}) = \sum_{n=1}^{N} a_n t_n k(\vec{x}, \vec{x}_n) + b$

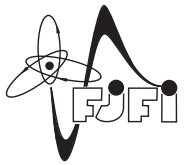
SVM - Support Vector Machine

Nelineárně separovaná data - oddělené „plochou“



Zdroj [1]

Outline



SVR - Support Vector Regression

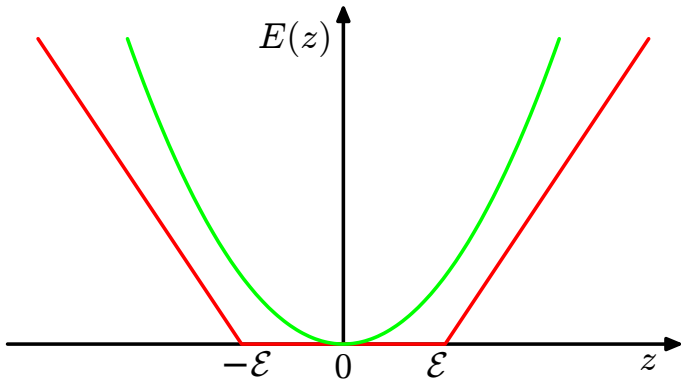
Regrese dat založená na SVM

Princip:

- Nalezení nepřímějšího epsilon-pásu aproximujícího data

Parametry:

- epsilon - odpovídá rozptylu dat
- sigma - velikost očekávaných objektů
- C - penalizace „vzdálených pozorování“ (outliers)



Zdroj [1]

SVR - Support Vector Regression

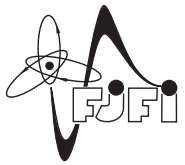
Výhody

- Neparametrická metoda regrese
- Poměrně odolné vůči „vzdáleným pozorováním“
- Body uvnitř ϵ-pásu neovlivňují regresi
- Sparse kernel - poměrně rychlý výpočet předpovědí

Nevýhody

- Složitost úlohy roste s N^3
- Preferuje pouze přímost

Outline



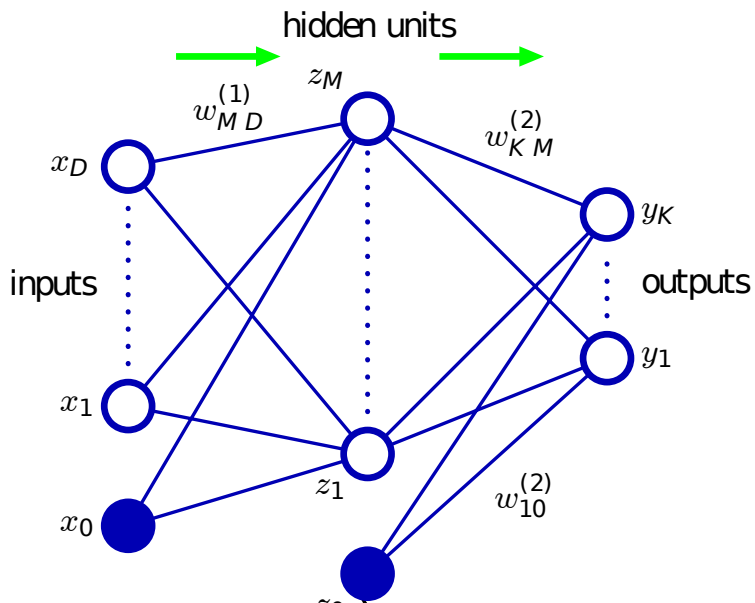
NN - Neural Networks

Princip (obrázek)

- „Vícevrstvý perceptron“
- Hledá se pevný počet parametrů

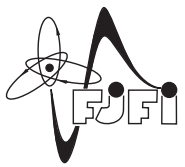
Parametry:

- Každá vrstva různé množství parametrů
- Volba transformační funkce pro každou vrstvu



$$y_k(\vec{x}, \vec{w}) = \sigma\left(\sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{D} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right)$$

Zdroj [1]



NN - Neural Networks

Výhody

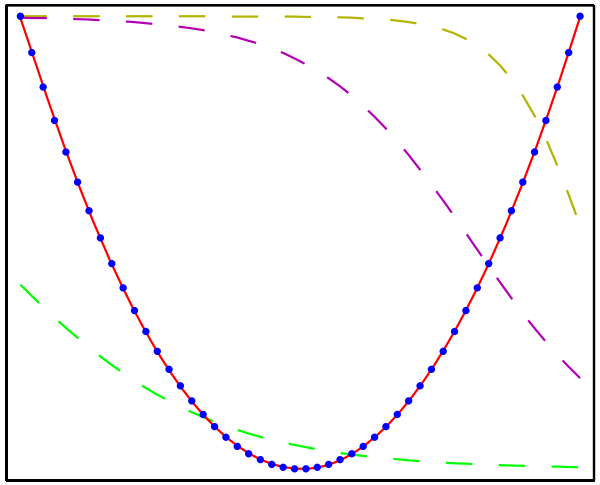
- Široké možnosti využití
- Rychlejší vyhodnocení dat proti SVM
- Složitost klasifikace nezávisí na N
- Možnost algoritmus „zadrátovat“

Nevýhody

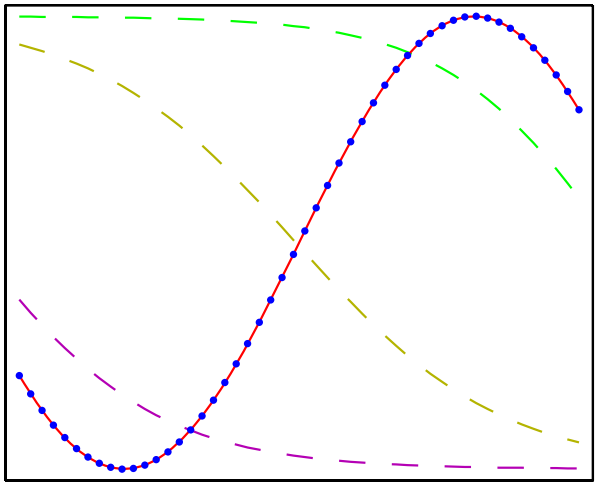
- NN není konvexní funkcí svých parametrů
  Existují lokální minima
- Řešení se hledá iteračně
- Pevný počet parametrů

NN - Neural Networks

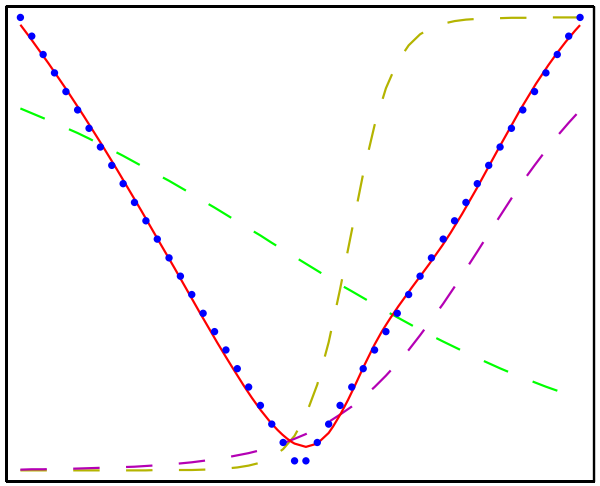
|
)



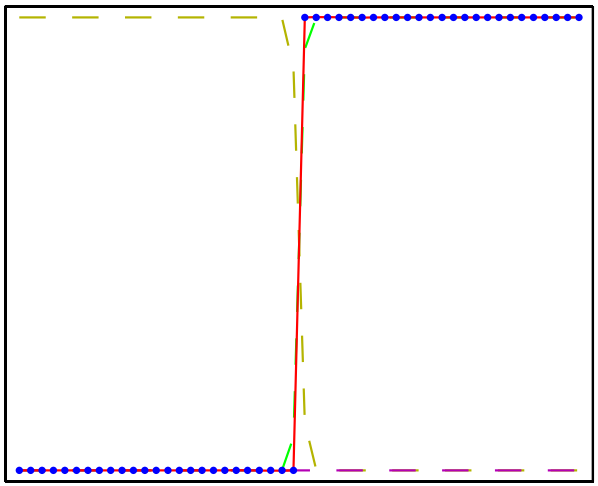
(a)



(b)

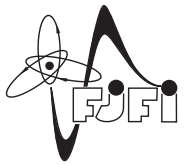


(c)



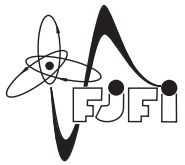
(d)

Outline



SVR - Scaling laws

- Scaling law: $y = \prod x_i^{\alpha_i}$
- Zlogaritmované: $\log(y) = \sum \alpha_i \log(x_i)$
- Plocha v transformovaném prostoru
- Hledání pomocí SVR a lineárního jádra

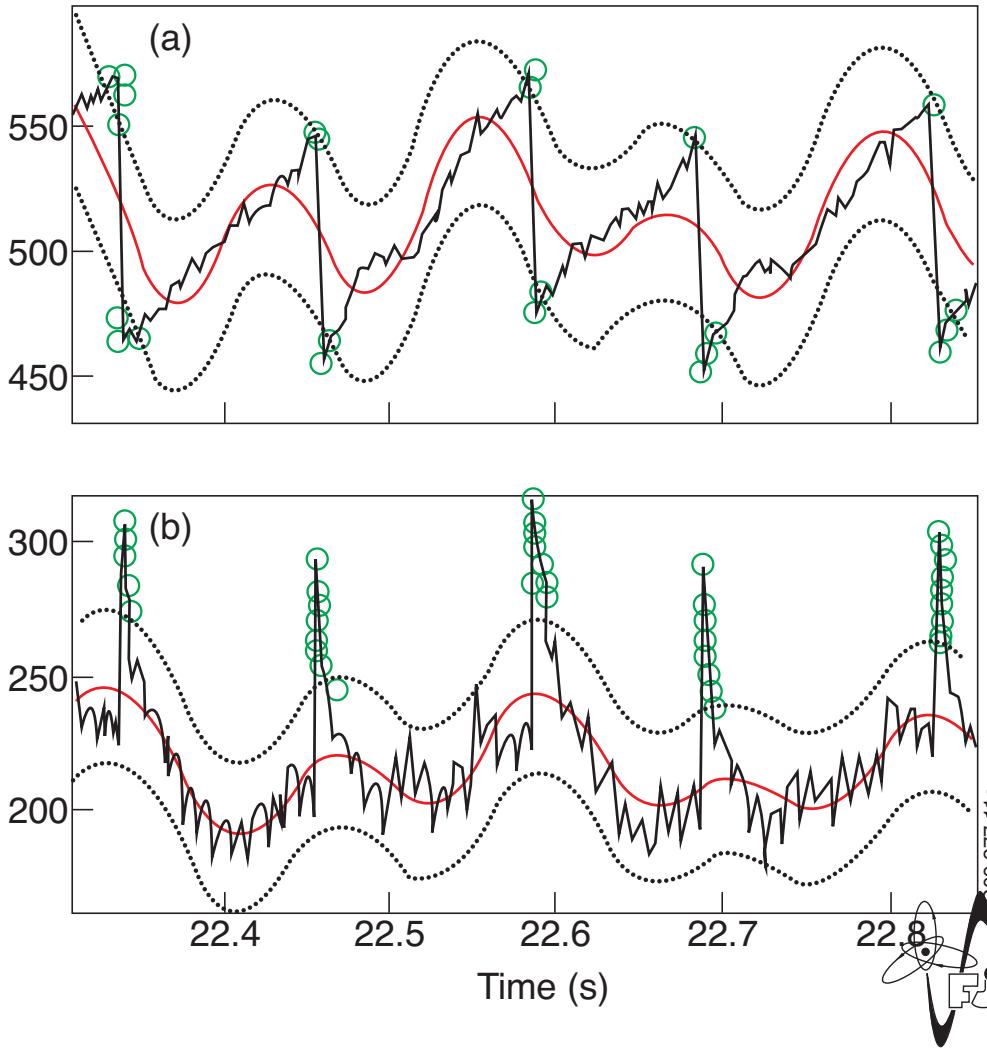


SVR - Detekce změny dat

SVR hledá nejrovnější řešení
→ Support vectors budou body změny

Využití:

- Detekce píků (ELMy)
- Detekce hran, skoků



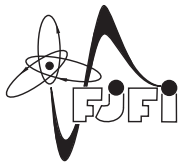
Ostatní metody

Jednoduché metody:

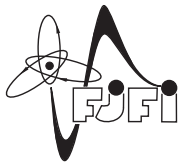
- Parzenovo okno - histogram, pomalá predikce
- k-NN, pomalá predikce
- k-mean - *k* skupin, určení těžiště

Další metody z TOP10:

- Pagerank
- CART - stromy
- Genetické algoritmy



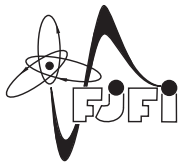
Outline



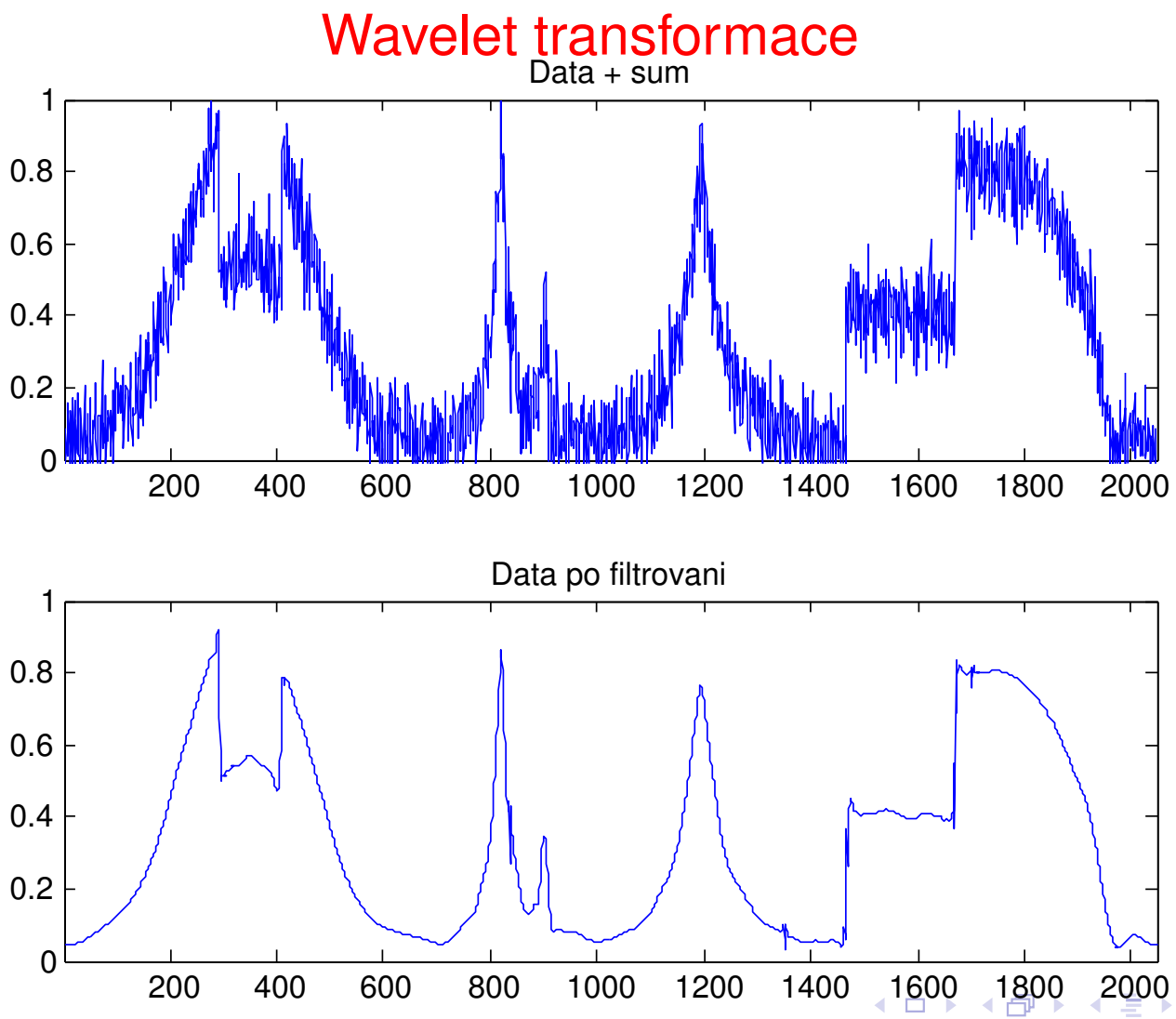
Preprocessing

Využití:

- Vyčištění dat od šumu
- Komprese dat => nižší náročnost na zpracování
- Normalizace dat
- Zvýraznění jevů (skoky, různé profily)



Wavelet transformace



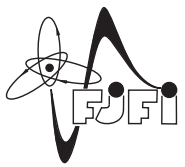
Praktická ukázka

Open source knihovna LIBSVM

Příkaz:
easy.py [zdrojová data] [testovaná data]

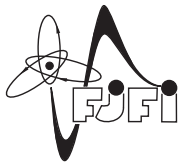
Formát dat:

[Y]	1:[$x^{(1)}$]	2:[$x^{(2)}$]	...			
0	1:0	2:400	3:400	4:10	5:	6:3.44



Závěr

- Základní učící se algoritmy jsou SVM a NN
- Velmi široké využití pro data mining
- Po učící se fázi jsou algoritmy velmi rychlé



Další informace I


C.M.Bishop
Pattern Recongnition and Machine Learning


V.Vapnik
The Nature of Statistical Learning Theory