České vysoké učení technické v Praze

Fakulta elektrotechnická

Katedra kybernetiky
Katedra počítačů

Intelligent Data Analysis
RESEARCH GROUP

COMPUTATIONAL INTELLIGENCE GROUP

Vytěžování dat – cvičení V

Klasifikace a regrese pomocí k-NN a lineárních modelů

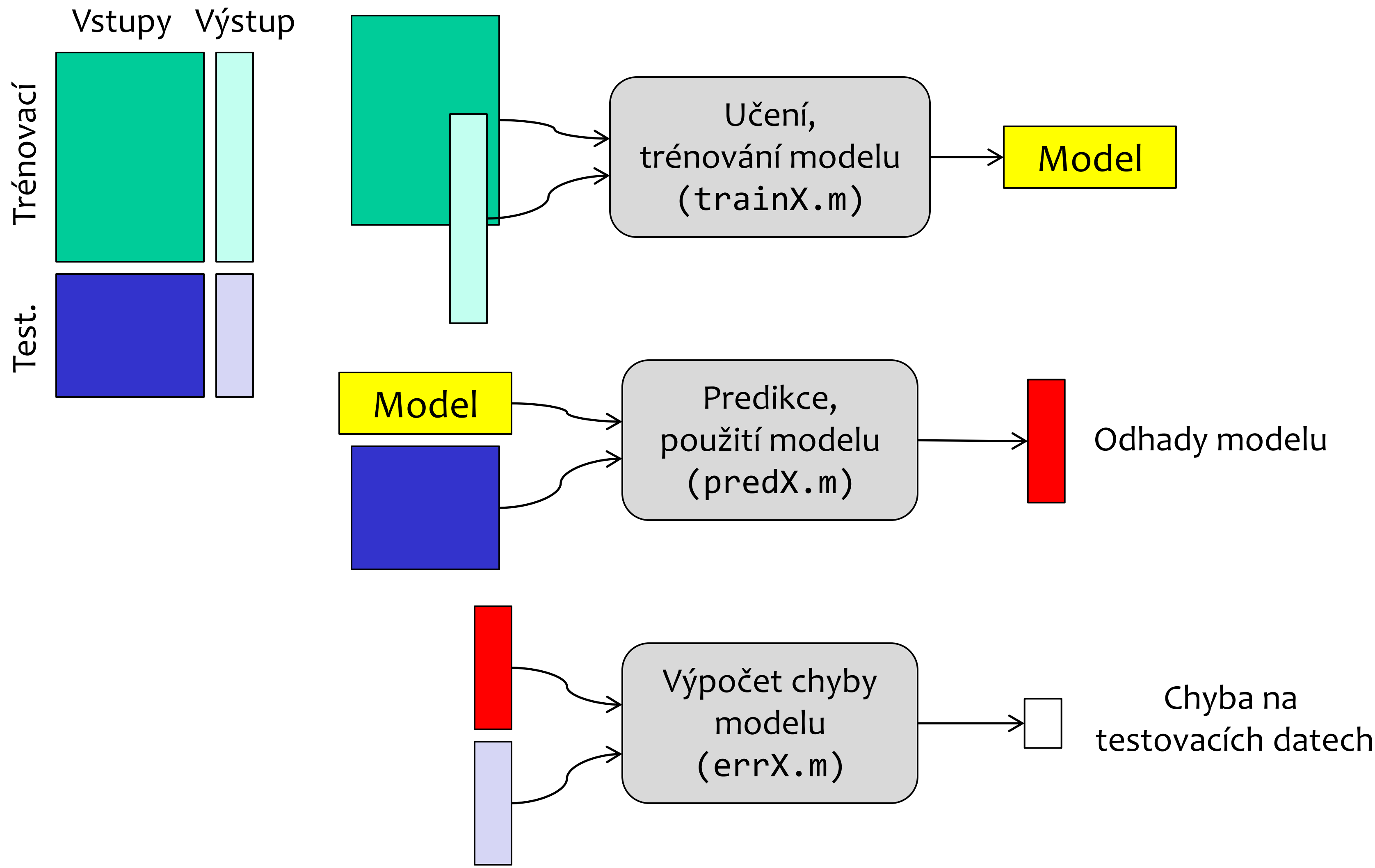
Křivka učení, křížová validace

Petr Pošík: posik@labe.felk.cvut.cz
Pavel Kordík: kordikp@fel.cvut.cz

Program cvičení

- Klasifikace pomocí k-NN
  - Vizualizace, křivka učení, chyby vs. „ohebnost" modelu
- Klasifikace pomocí lineárního modelu (perceptron)
  - Vizualizace, křivka učení, chyby vs. „ohebnost" modelu
- Regrese pomocí k-NN
  - Vizualizace, křivka učení, chyby vs. „ohebnost" modelu
- Regrese pomocí lineárního modelu
  - Vizualizace, křivka učení, chyby vs. „ohebnost" modelu
- Křížová validace, rozšíření báze

Typický postup učení



Načtení dat

- V dnešním cvičení budeme opět používat databázi aut.
- Načtěte soubor auto-mpg.data-mod-names.csv do objektu dataset a definujte jména jednotlivých atributů

Načtení dat

- V dnešním cvičení budeme opět používat databázi aut.
- Načtěte soubor auto-mpg.data-mod-names.csv do objektu dataset a definujte jména jednotlivých atributů

```
auta = dataset('file', 'auto-mpg.data-mod-names.csv',...
   'ReadVarNames', false, 'ReadObsNames', false,...
   'delimiter', ',', ...
   'VarNames', {'mpg', 'cyl', 'disp', ...
   'hp', 'wgt', 'acc', 'year', 'org', 'name'});
```

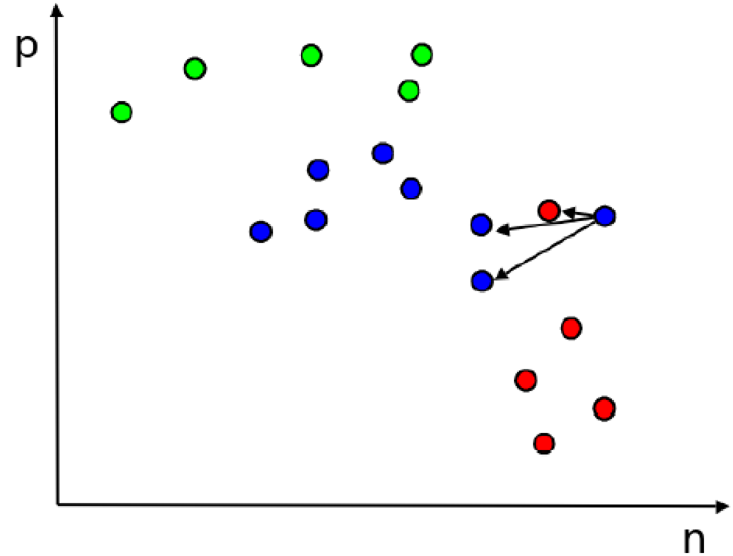
Normalizace dat

- `auta_norm = datasetfun( @minmax, auta(:,1:5), 'UniformOutput', false );`
- `auta_norm = [auta_norm{:}];`
- `auta = replacedata( auta, auta_norm, 1:5);`

KLASIFIKACE

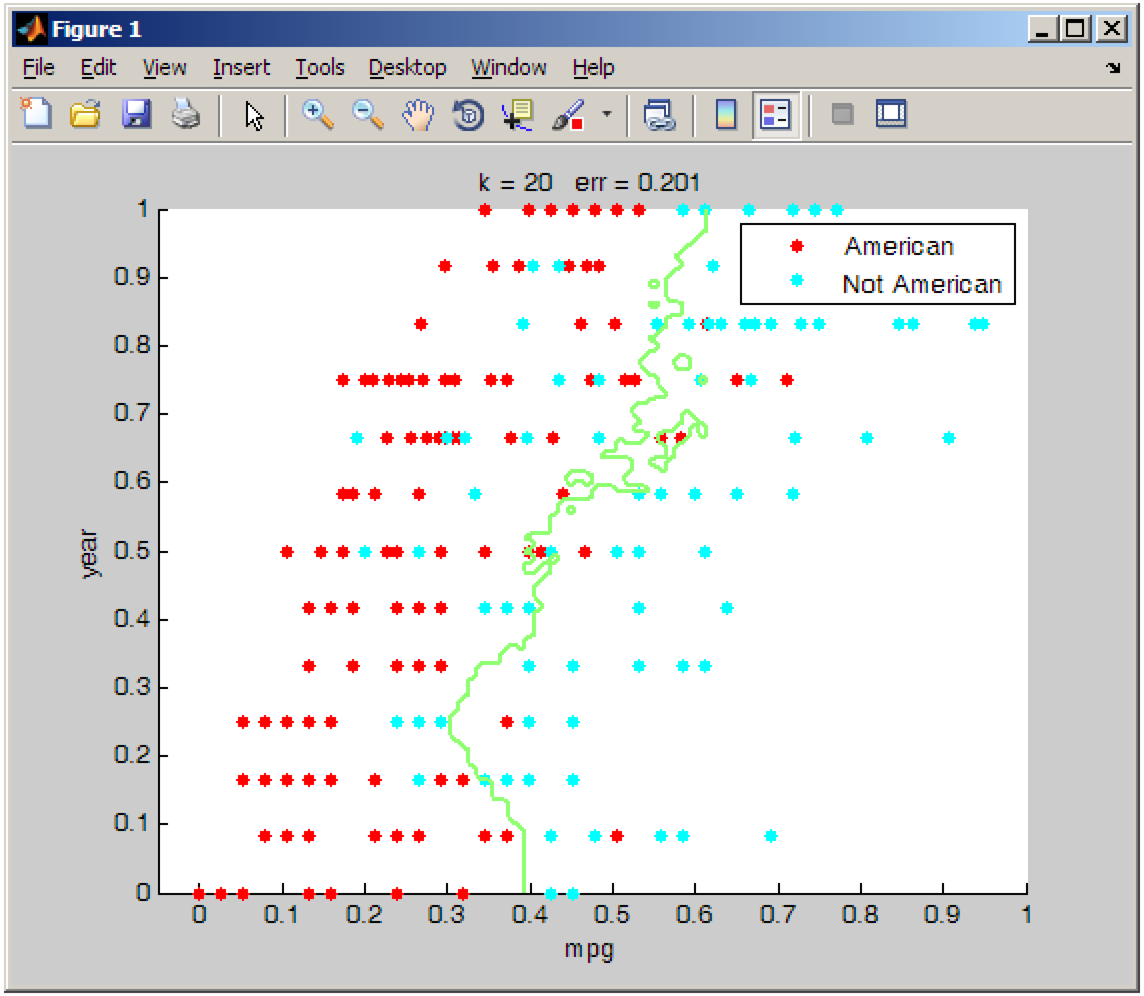
K-NN: Připomenutí

- Když je potřeba oklasifikovat novou instanci, naleznu v trénovací množině nejbližší instanci (**k** nejbližších instancí) a podle jejich tříd určím výslednou třídu nové instance.



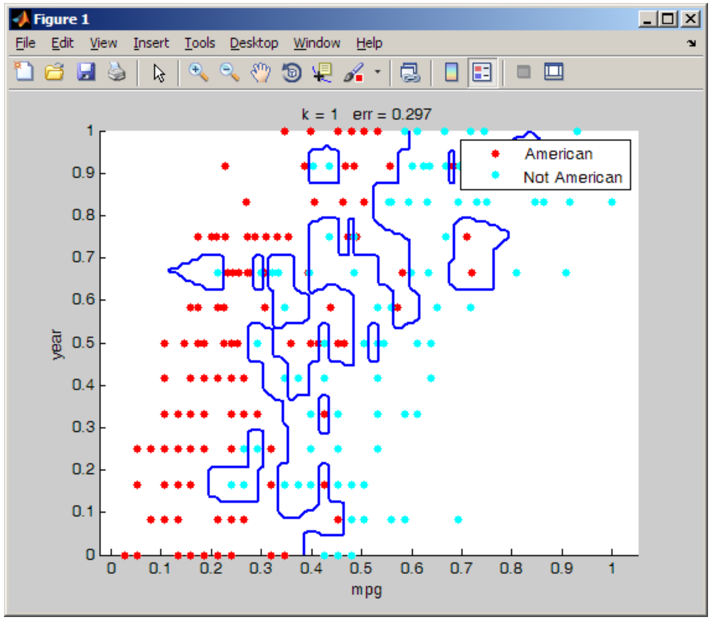
K-NN:
Vizualizace

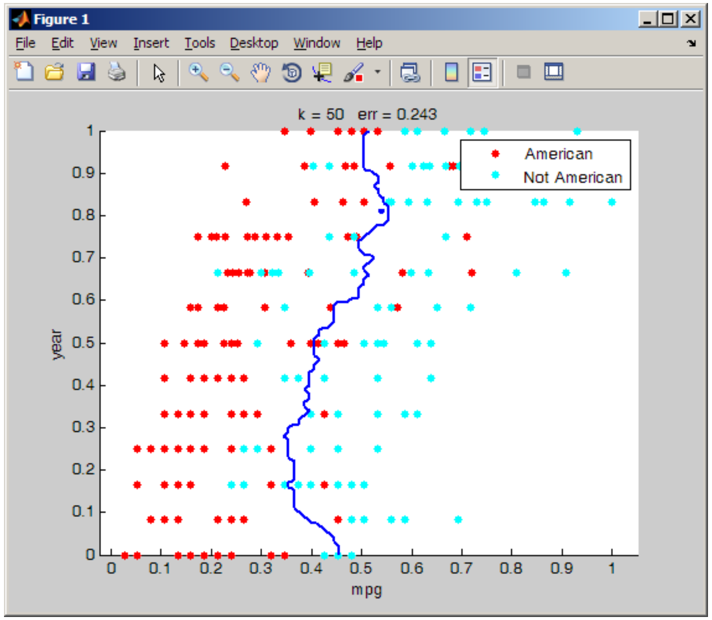
- viz **`scrVizClassKNN.m`**



K-NN:
Chyba vs. „ohebnost“ modelu

- Parametr *k* určuje ohebnost modelu.
  - Jak? Jak se to projeví z hlediska chyb?
  - viz **`scrVizClassKNN.m`**



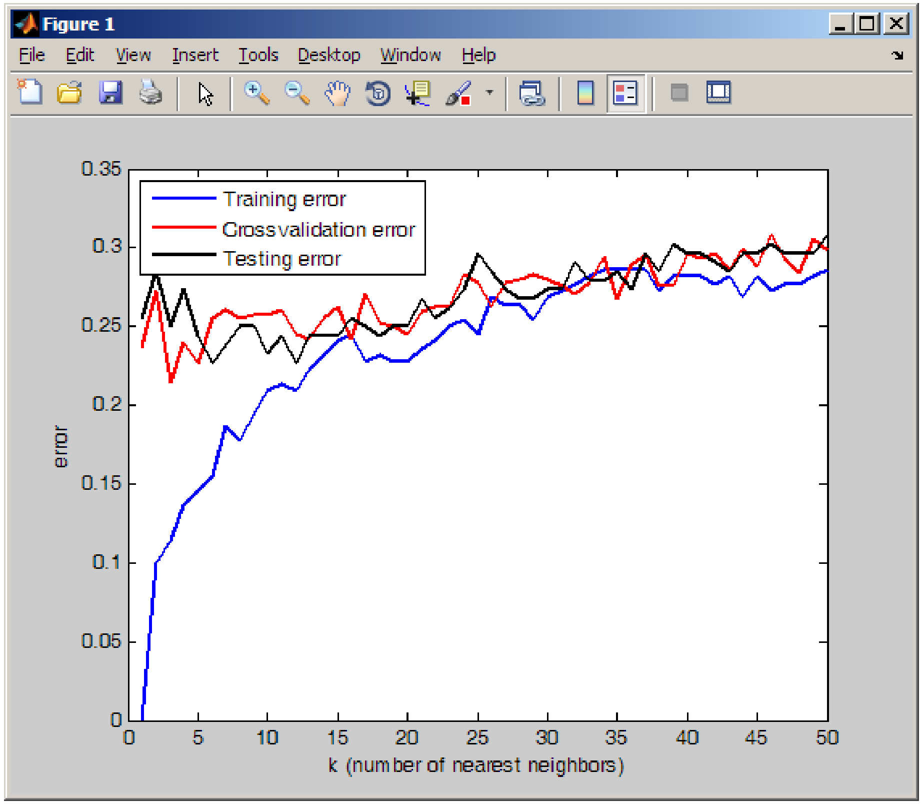


COMPUTATIONAL INTELLIGENCE GROUP

IDA Intelligent Data Analysis RESEARCH GROUP

K-NN:
Chyba vs. „ohebnost" modelu II

- Závislost chyby kNN klasifikátoru na parametru *k*
  - viz **`scrClassTTErrorKNN.m`**
  - Jak to, že trénovací chyba pro 1NN je nulová?



COMPUTATIONAL INTELLIGENCE GROUP

IDA Intelligent Data Analysis RESEARCH GROUP

Křížová validace

- Znáte z přednášky
  - K čemu slouží?
  - Jak funguje?

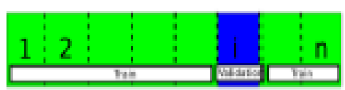
Křížová validace

- Umožňuje odhadnout testovací chybu a potřebuje k tomu jen trénovací data
- Slouží k výběru (vhodné struktury a parametrů) modelu

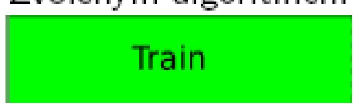
Úplný postup pro výběr algoritmu křížovou validací, získání klasifikátoru a odhad jeho kvality

1. Rozdělíme data na Train / Test

2. Křížovou validací na Train vybereme algoritmus



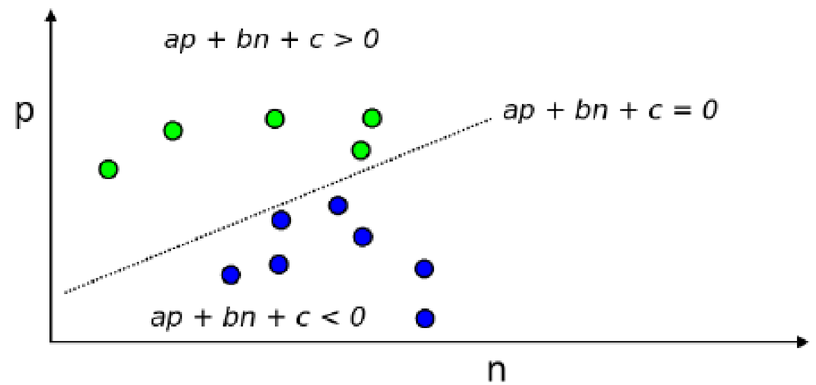
3. Zvoleným algoritmem sestrojíme klasifikátor na Train



4. Jeho kvalitu odhadneme na Test

Lineární klasifikátor:
Princip

- Lineární funkce: $f(\mathbf{x}) = w_1x_1+w_2x_2+\ldots+w_0$
- Stejná funkce, jiný zápis:
  $\mathbf{x} = (x_1,x_2,\ldots,x_D,1)$, $\mathbf{w} = (w_1,w_2,\ldots,w_D,w_0)$
  $f(\mathbf{x}) = \mathbf{w}*\mathbf{x}$ ... skalární součin
- Klasifikační pravidlo:
  - když $f(\mathbf{x})>0$,
    $\mathbf{x}$ je ze třídy 1,
  - když $f(\mathbf{x})<0$,
    $\mathbf{x}$ je ze třídy 2.



- Hranice mezi třídami: $f(\mathbf{x}) = 0$

Lineární klasifikátor:
Učení

- Známe-li vektor **w**, klasifikace je jednoduchá
- Jak vektor **w** zjistit? (Jak sestrojit lineární kl.?)
  - stanovit „ručně“
  - naučit na základě trénovacích dat
- Existuje mnoho metod učení lin. klasifikátoru
  - perceptron
  - lineární diskriminační analýza
  - ...

Lineární klasifikátor: Perceptronový algoritmus

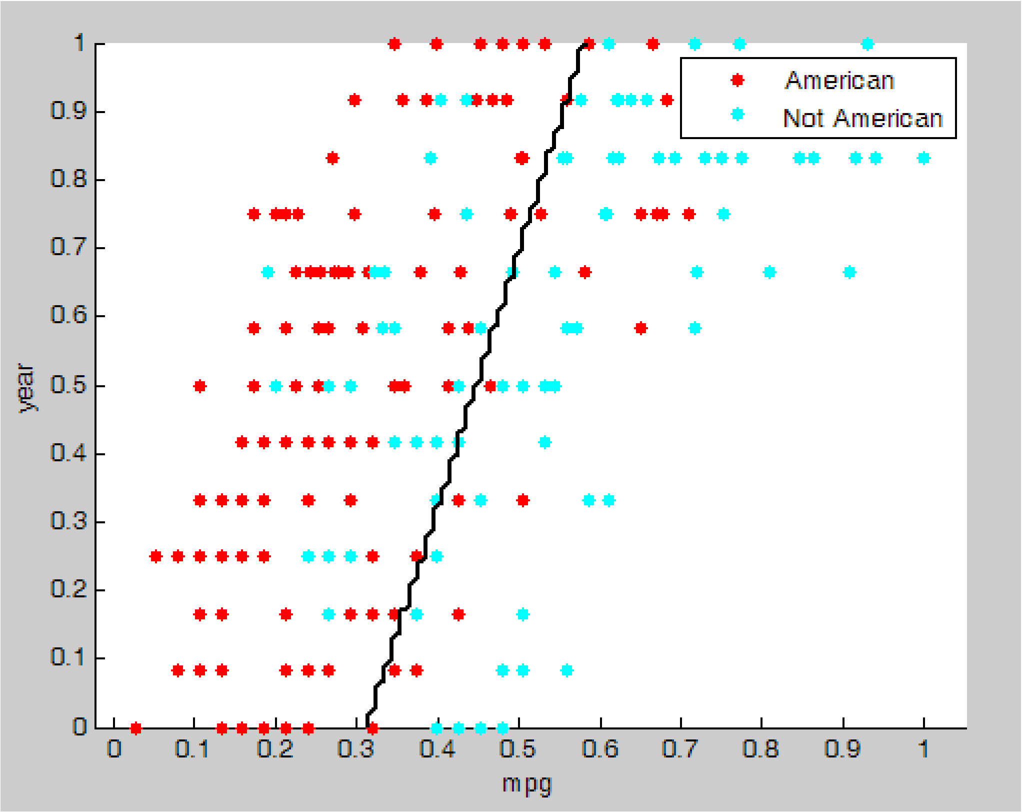
Perceptronový algoritmus

Input: $\vec{w}$, $\eta(.)$, θ
Output: $\vec{w}$
repeat
  $k \leftarrow k + 1$
  $\vec{w} \leftarrow \vec{w} + \eta(k) \sum_{\vec{x}_i \in E} \vec{x}_i$
until $|\eta(k) \sum_{\vec{x}_i \in E} \vec{x}_i| < \theta$;
return $\vec{w}$

- body **x** jsou v homogenních souřadnicích, tedy „s přidanou jedničkou na konci."
- body ve třídě 2 jsou invertovány
- množina E špatně zaklasifikovaných bodů se mění s každou iterací cyklu

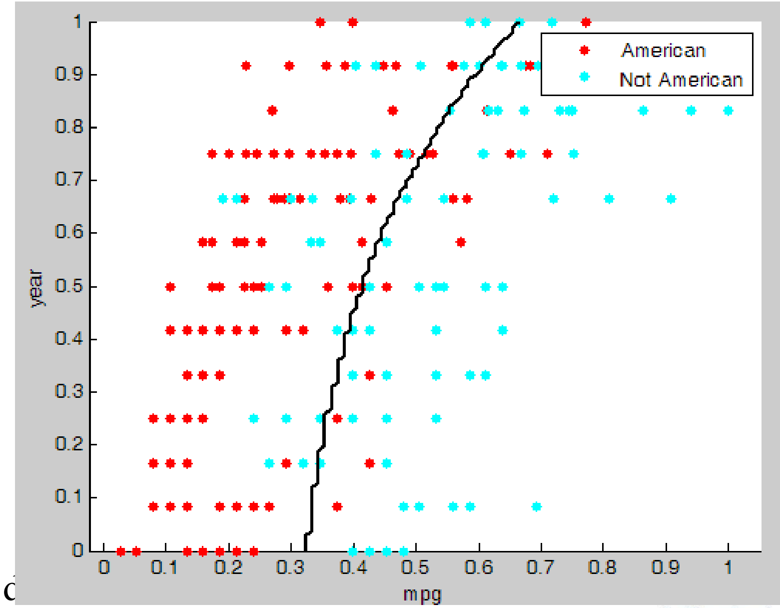
Lineární klasifikátor:
Vizualizace učení perceptronem

- viz `scrVizClassLinear.m`



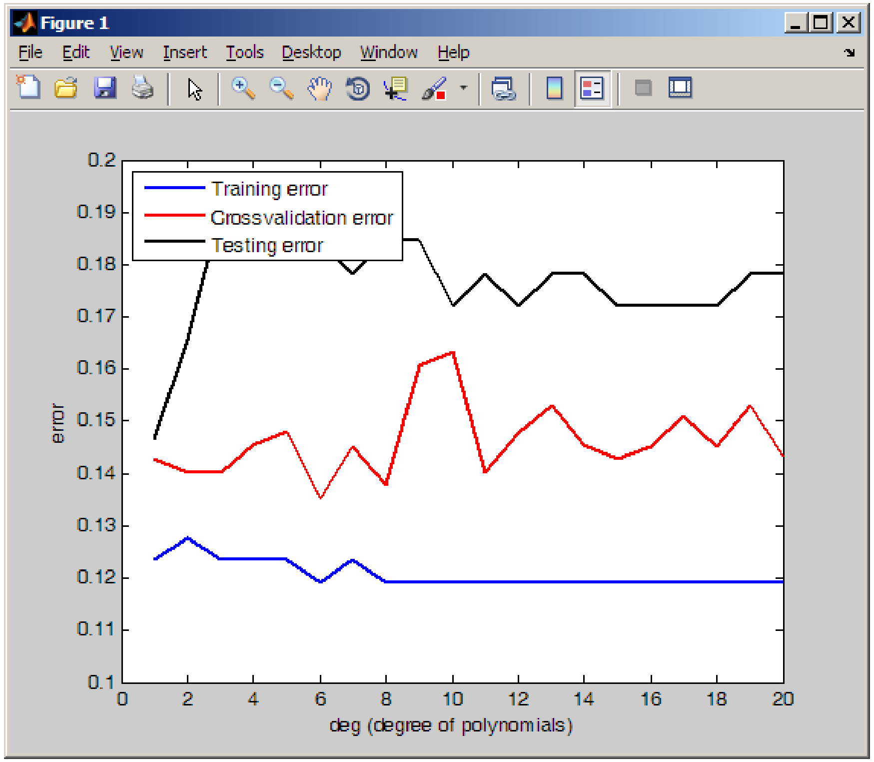
Rozšíření báze

- Převádí lineární separaci na nelineární v prostoru o vyšším rozměru
- $\mathbf{x} = (x_1,x_2) \rightarrow \mathbf{z} = (z_1,z_2,z_3) = (x_1^2, x_1x_2, x_2^2)$
- Najdeme-li lin. funkci $f(\mathbf{z}) = w_1z_1+w_2z_2+w_3z_3$, najdeme i nelin.fci $f(\mathbf{x}) = w_1x_1^2+w_2x_1x_2+w_3x_2^2$
- $\mathbf{w}$ zjistíme metodami pro učení lineárního klasifikátoru
- `scrVizClassLinear.m`
- nastavit deg>1



Lineární klasifikátor:
Chyba vs. „ohebnost“ modelu

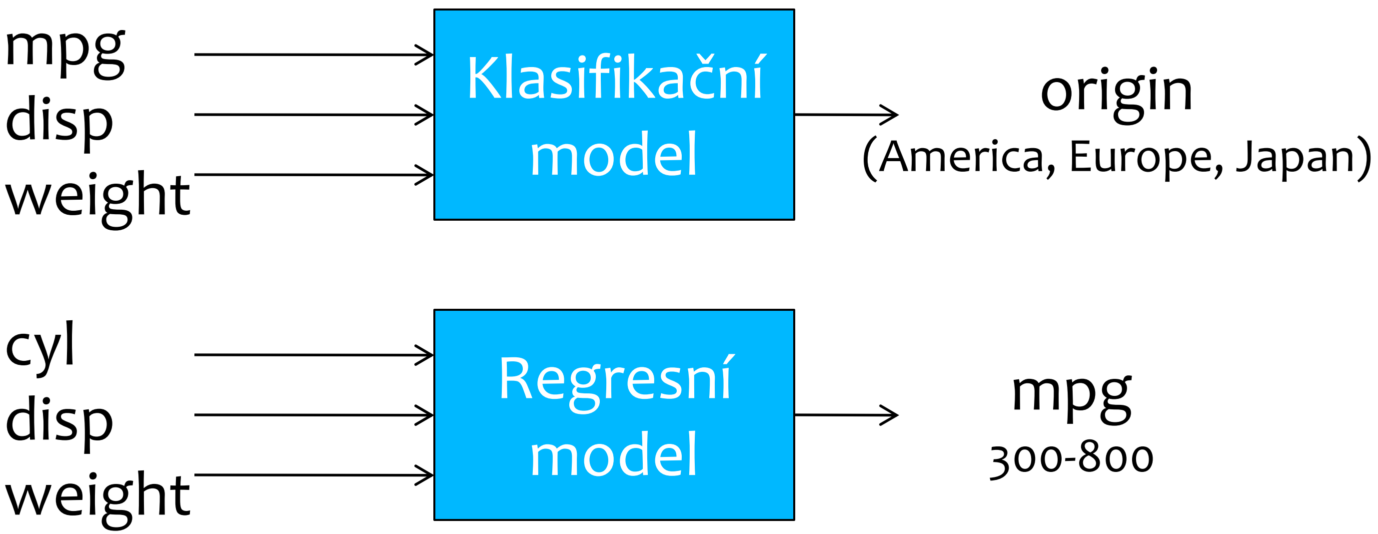
- „Ohebnost“ zajištěna rozšířením báze
- viz `scrClassTTErrorLinear.m`



REGRESE

Regrese: Úvod

- Klasifikační i regresní model:
  y = f(**x**)
- Klasifikace: y je nominální (název třídy)
- Regrese: y je spojitá veličina (teplota, výška)



Chyba modelu

- Klasifikační model:
  - procento nesprávných předpovědí
- Regresní model:
  - součet absolutních hodnot odchylek $err = \sum_i \left| y_i - f(x_i) \right|$
  - **součet čtverců odchylek** $err = \sum_i \left(y_i - f(x_i) \right)^2$
  - průměrný čtverec odchylky $err = \frac{1}{N} \sum_i \left(y_i - f(x_i) \right)^2$
  - **odmocnina průměrného čtverce odchylky (RMSE)**

$$RMSE = \sqrt{\frac{1}{N} \sum_i \left(y_i - f(x_i) \right)^2}$$

K-NN pro regresi

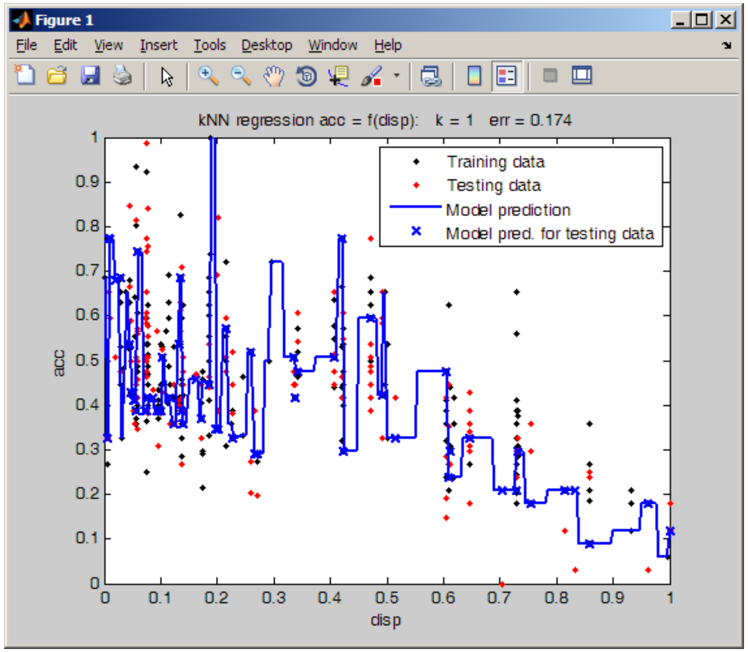
- Jak byste použili k-NN pro regresi (pro predikci spojité veličiny)?

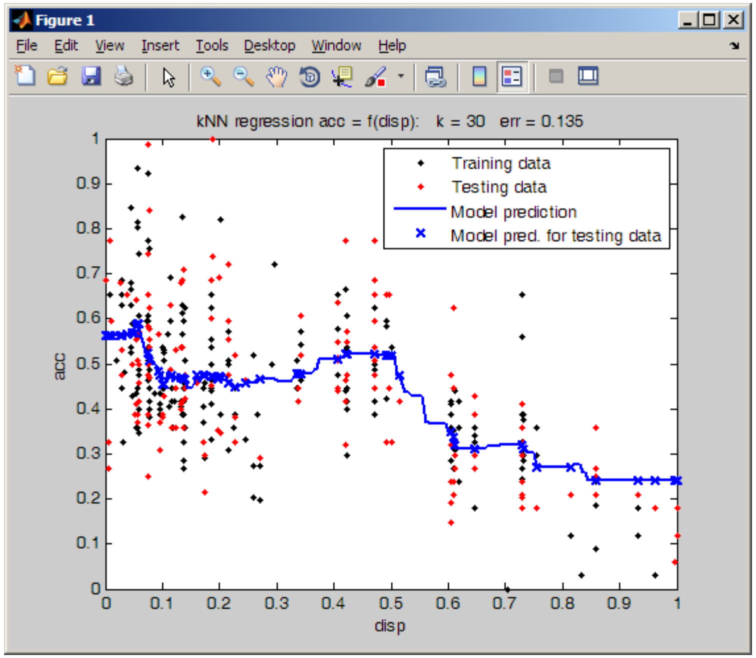
cyl	disp	wgt	mpg
2	1800	2000	35
2	1900	2500	30
4	1800	1500	33
4	2400	2200	25
6	2000	2500	16

cyl	disp	wgt	mpg
4	2000	2800	????

K-NN regrese

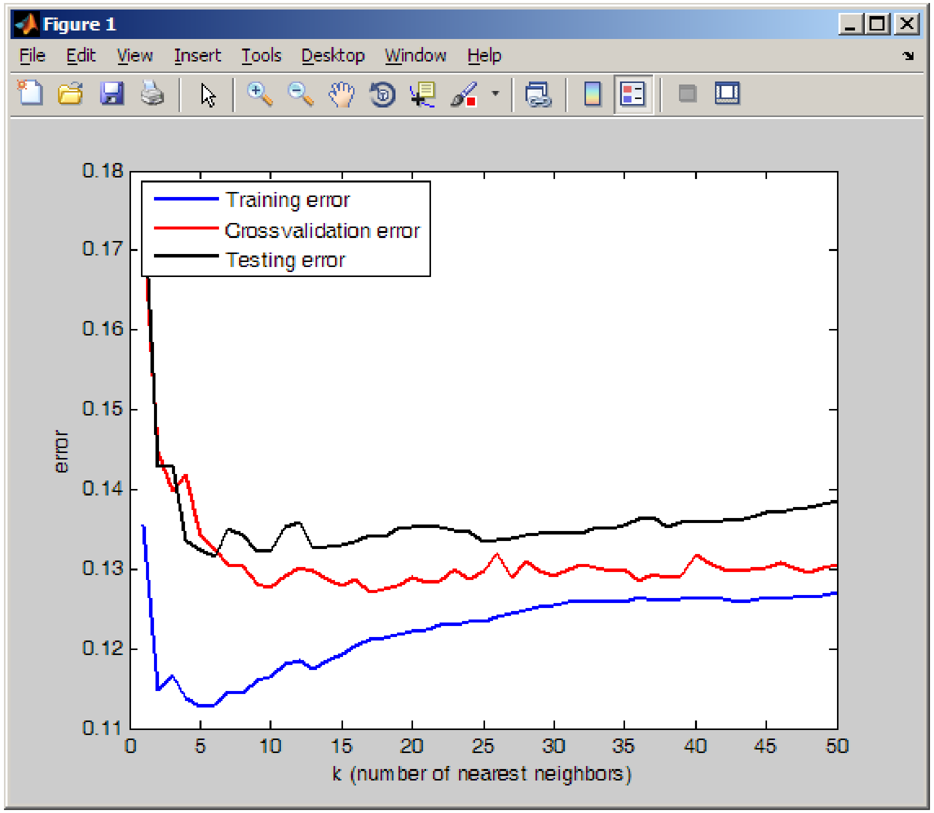
- viz `scrVizRegrKNN.m`
- acc = f(disp)
- Experimentujte s parametrem k





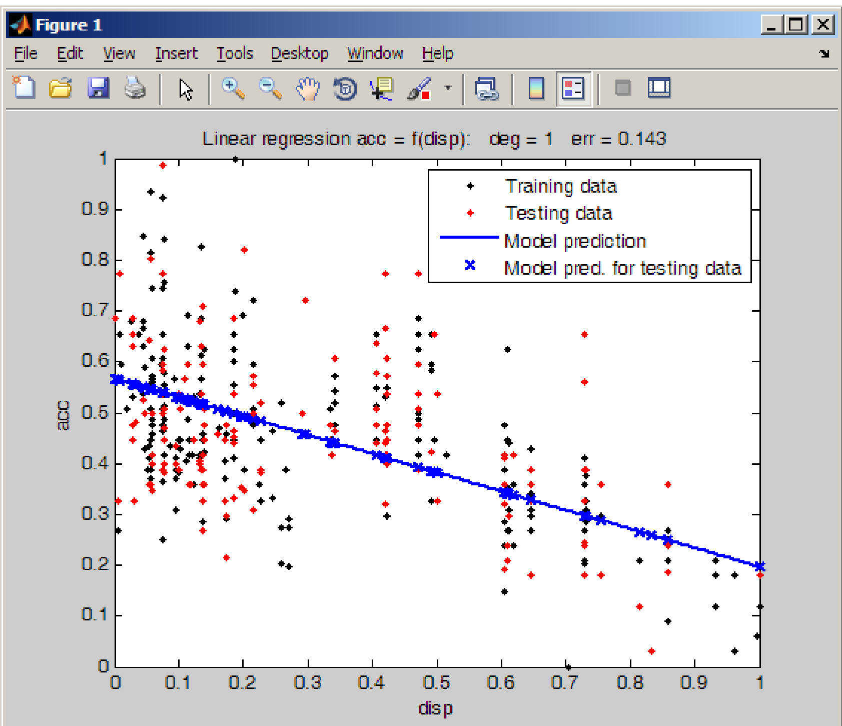
K-NN regrese:
Chyba vs. „ohebnost“ modelu II

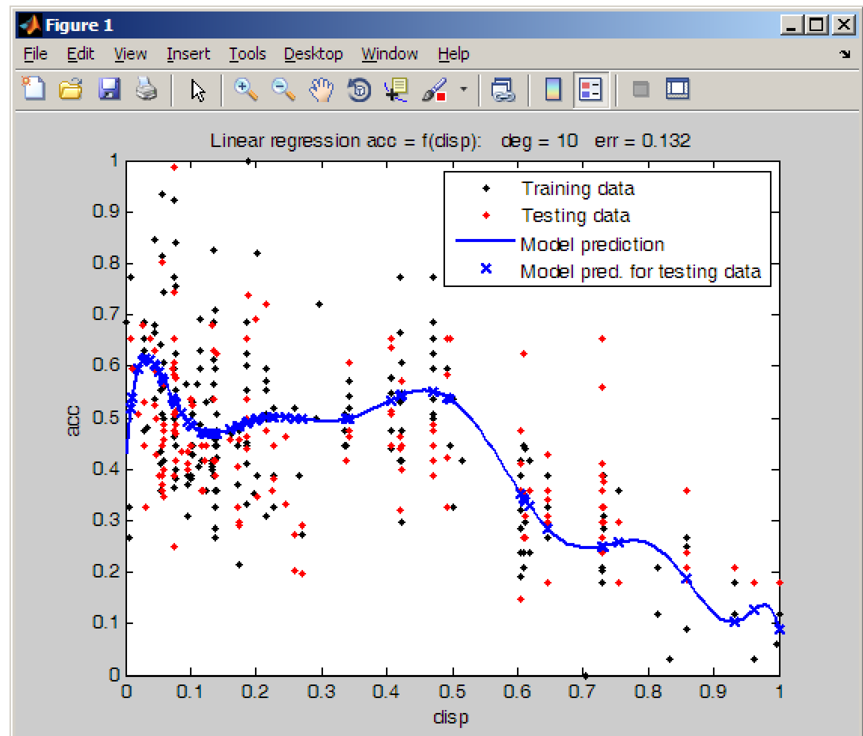
- Závislost chyby kNN modelu na parametru *k*
  - viz **`scrRegrTTErrorKNN.m`**
  - Jak to, že trénovací chyba pro 1NN není nulová?



Lineární regrese

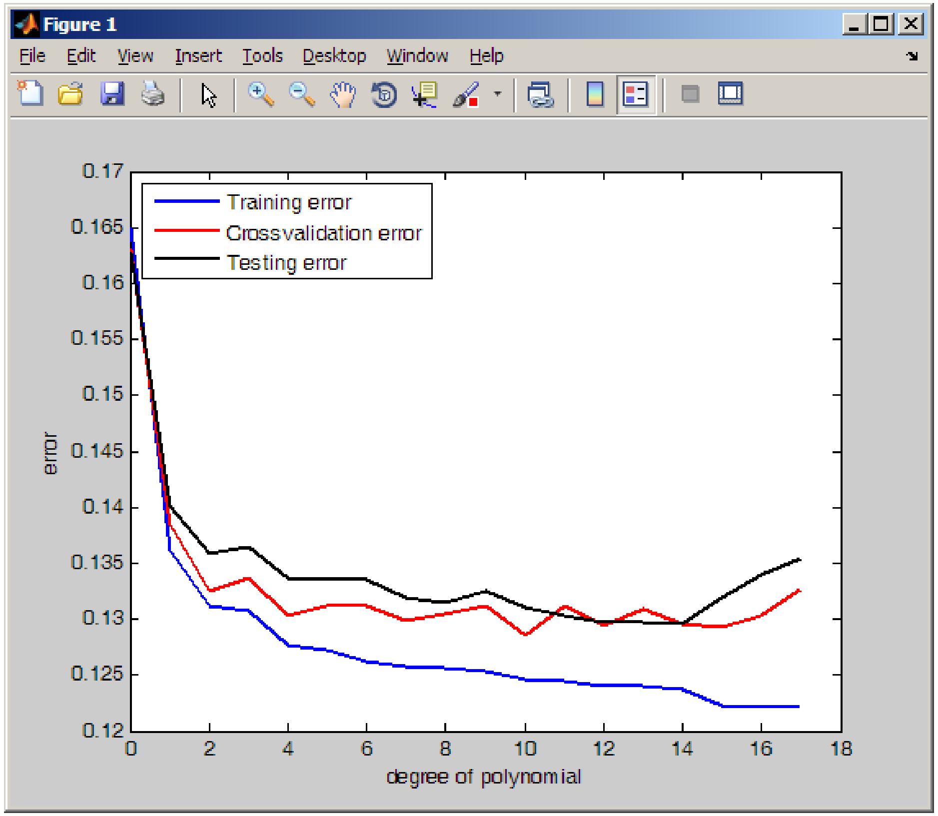
- viz `scrVizRegrLinear.m`
- acc = f(disp)
- Experimentujte s parametrem deg





Lineární regrese:
Chyba vs. „ohebnost“ modelu

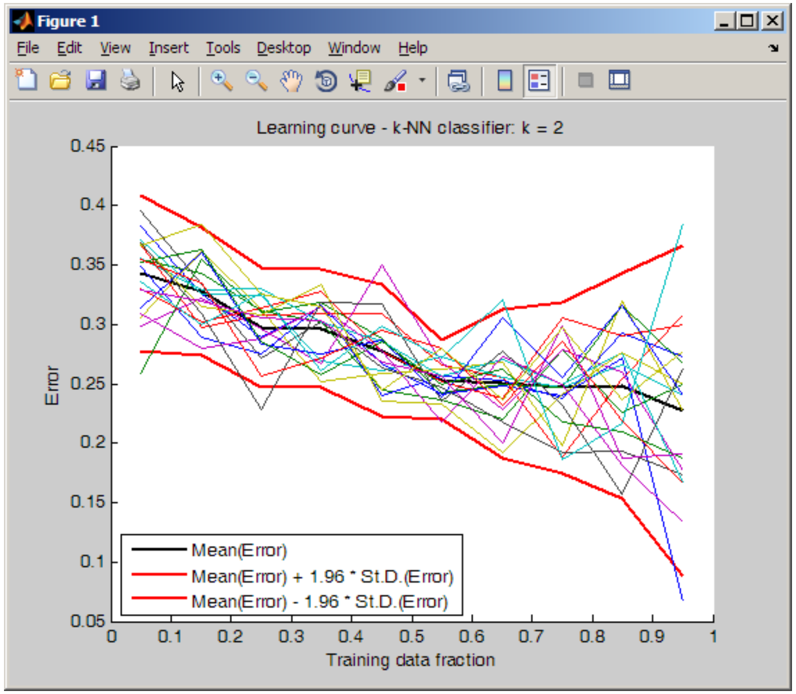
- Závislost chyby lin. modelu na parametru deg
  - viz **`scrRegrTTErrorLinear.m`**



KŘIVKY UČENÍ

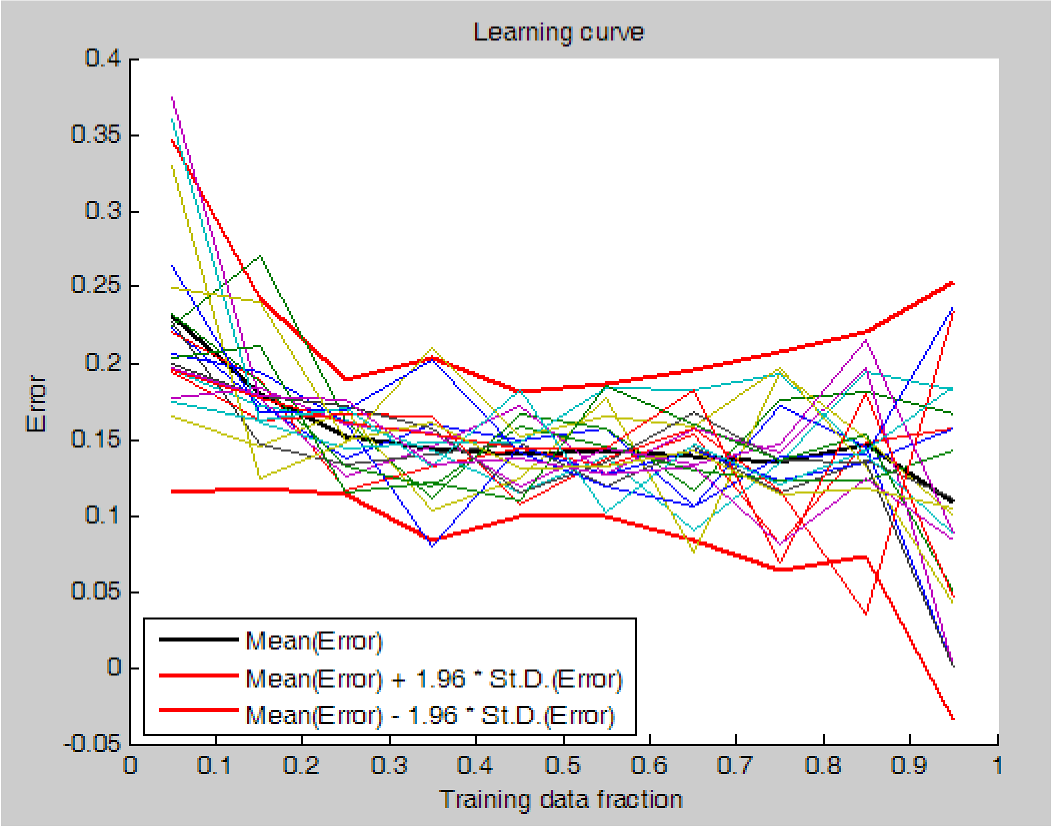
K-NN klasifikátor:
Křivka učení

- Opakování z přednášky:
  - přesnost (chyba) modelu (na testovacích datech) v závislosti na velikosti trénovacích dat
  - viz **`scrClassLearningCurveKNN.m`**



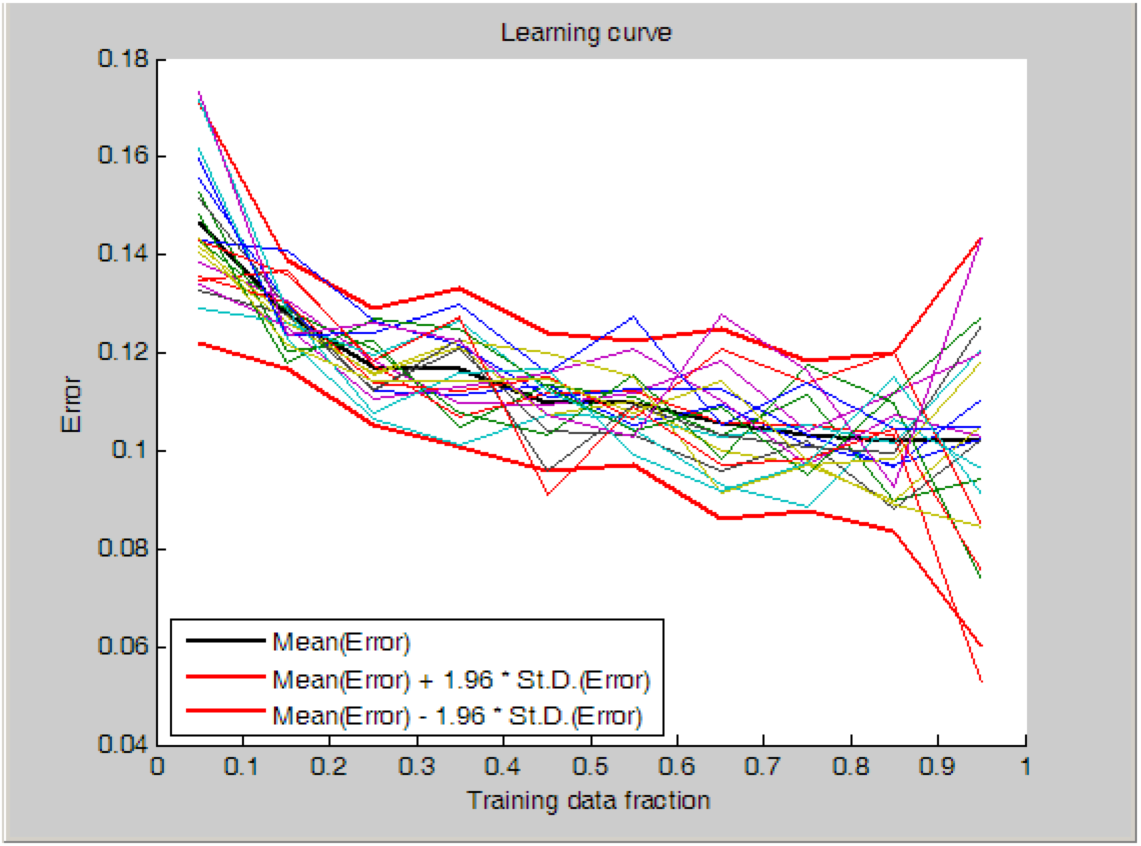
Lineární klasifikátor:
Křivka učení

- viz `scrClassLearningCurveLinear.m`



K-NN regresní model:
Křivka učení

- viz **`scrRegrLearningCurveKNN.m`**



Lineární regresní model:
Křivka učení

- viz `scrRegrLearningCurveLinear.m`