Internet a klasifikační metody, 2. přednáška

Základní koncepty týkající se klasifikace

volitelný předmět pro magisterské studium

Martin Holeňa

O čem to bude?

- Klasifikace do 2 a do více tříd
  - specifické koncepty týkající se binární klasifikace
- Charakterizace kvality klasifikace (obecné i binární)
- Konstrukce klasifikátorů z existujících dat – učení
- Lineární separabilita, obecná metoda jejího dosažení
- Ohraničení klasifikace vůči regresi a shlukování

Klasifikace a klasifikátory

- Klasifikace: třídění objektů na základě příznaků
  - *třídy*: konečný počet, někdy nejednoznačná hranice
  - *příznaky* – data všech typů: *nominální* (bydliště,…), *ordinální* (vzdělání,…), *reálná* čísla + vektory + pole
- Abstraktní reprezentace klasifikace: zobrazení prostoru příznaků do množiny tříd – *klasifikátor*, $F: X \rightarrow \{C_1, \ldots, C_m\}$

Binární klasifikace

- Pozitivní | negativní třída $C_+ | C_-$, $F: \rightarrow \{C_+, C_-\}, \rightarrow \{1, -1\}, \rightarrow \{1, 0\}$
- *Falešná pozitivita*: $x \in C_-, F(x) = 1$ | *negativita*: $x \in C_+, F(x) = -1$
  - E-mail: FP = nedoručený ham, FN = nerozpoznaný spam
- Když je *nedosažitelná nízká FP+FN* současně,

  pomůže *3. třída* mezi C_+, C_- (interpretace často umělá)

spam | karanténa | ham

Jak měřit kvalitu klasifikace?

- Potřebujeme znát správnou příslušnost do tříd

n	Klasifikace	$C_1(n_{\cdot 1})$	...	$C_k(n_{\cdot k})$
Správná Třída	$C_1(n_{1\cdot})$	n_{11}	...	n_{1k}
	...	...	...	...
	$C_k(n_{k\cdot})$	n_{k1}	...	n_{kk}

matice záměn

- Nejsnazší: *přesnost* $= \frac{\text{počet souladů}}{n} = \frac{1}{n}\sum_{i=1}^{k} n_{ii}$
  - 1– přesnost = *chybovost* $= \frac{\text{počet chyb}}{n} = \frac{1}{n}\sum_{i\neq j=1}^{k} n_{ij}$

Různá cena různých chyb

- Tradiční přesnost a chybovost předpokládá, že

  nám *každá chyba vadí stejně. Nerealistické!*

  - nedoručený ham vadí > než nerozpoznaný spam
  - nerozpoznané napadení sítě ⋙ než falešný poplach

- Řešení: *cena* $c_{i,j}$ klasifikace C_j je-li C_i (tradičně $c_{ii} = 0, c_{ij} = 1$)
  - zobecnění chybovosti je střední hodnota ceny: $\frac{1}{n}\sum_{i,j=1}^{k} c_{ij} n_{ij}$

Matice záměn binární klasifikace

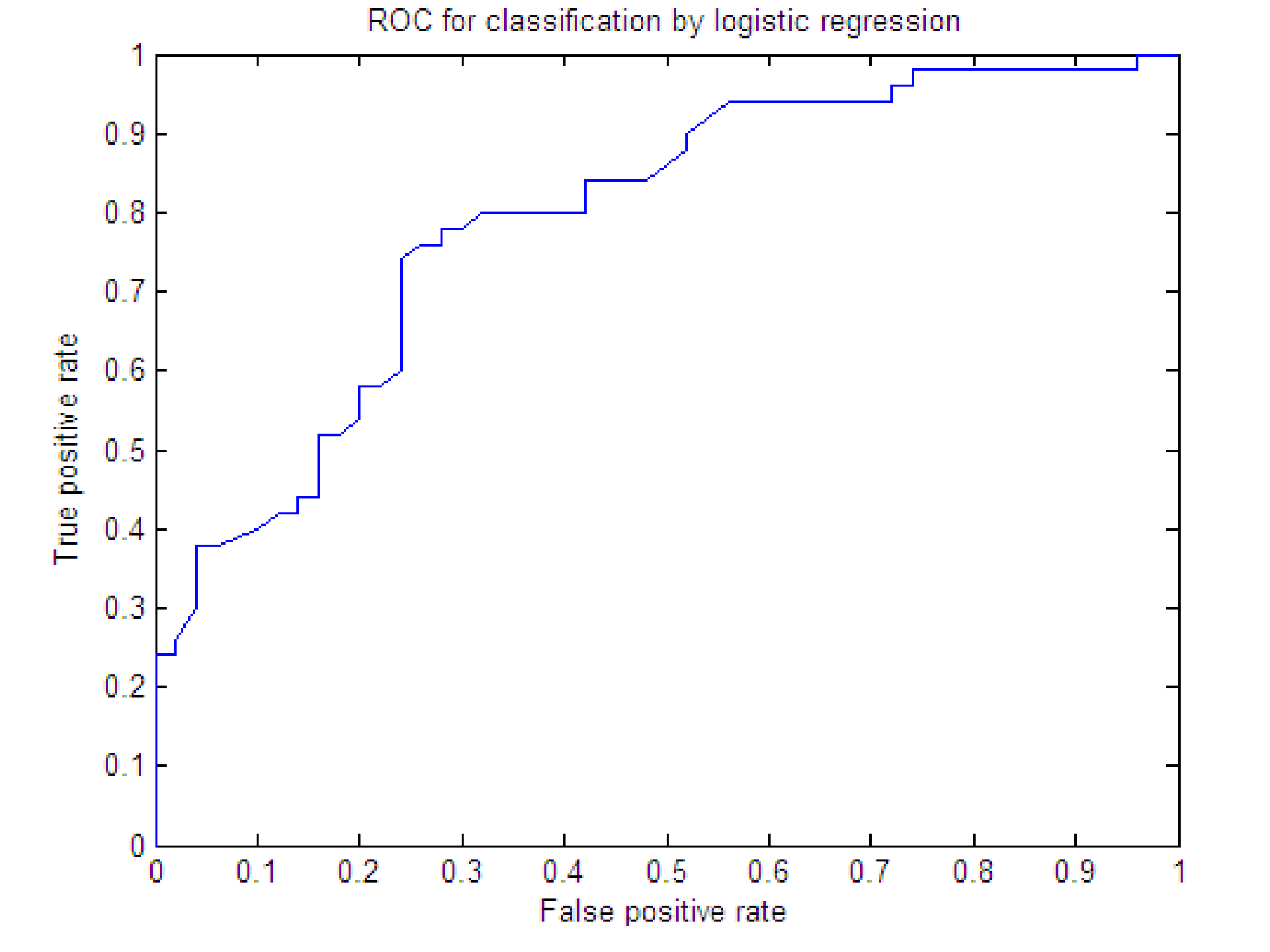
n	Klasifikace	$+(n_{+.})$	$-(n_{-.})$
Správná třída	$+(n_{+.})$	TP	FN
	$-(n_{-.})$	FP	TN

TP: pravá pozitivní

TN: pravá negativní

- Obvykle používán poměr (*ratio*): $\mathrm{TPr} = \frac{\mathrm{TP}}{n_{+.}}, \mathrm{FPr} = \frac{\mathrm{FP}}{n_{-.}}, \ldots$
  - pro množinu klasifikátorů (např. parametrizovanou): *křivka*

  $(\mathrm{FPr}, \mathrm{TPr})$ se označuje *ROC* (receiver operating characteristic)


ROC for classification by logistic regression
True positive rate
False positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

Míry kvality binární klasifikace

- *Přesnost* (accuracy) $\mathrm{AC} = \frac{\mathrm{TP}+\mathrm{TN}}{n}$
- *Správnost predikce* (precision, predictive value) $\mathrm{PR} = \frac{\mathrm{TP}}{n_{\cdot+}}$
- *Citlivost* (sensitivity), vybavování (recall), detekční poměr $= \mathrm{TPr}$
- Další: *specificita* $= \mathrm{TNr} = \frac{\mathrm{TN}}{n_{-\cdot}} = 1 - \mathrm{FPr}$, F-míra $\mathrm{FM} = 2\frac{\mathrm{PR}*\mathrm{TPr}}{\mathrm{PR}+\mathrm{TPr}}$

  AUC − Area Under Curve (= pod ROC)

- Střední hodnota ceny: $\frac{1}{n}(c_{++}\mathrm{TP}+c_{+-}\mathrm{FN} + c_{-+}\mathrm{FP} + c_{--}\mathrm{TN})$

Míry kvality ve spamových filtrech

1. *Bez zahrnutí ceny*: všechny právě uvedené

2. *S cenou* c_{-+}: nedoručený ham, c_{+-}: nerozpoznaný spam

   - bereme $c_{++} = c_{--} = 0 \Longrightarrow$ střední hodnota ceny: $\frac{c_{-+}\mathrm{FP}+c_{+-}\mathrm{FN}}{n}$
   - *TCR* (total cost ratio) : cena nefiltrování ($\mathrm{TP}' = \mathrm{FP}' = 0$, $\mathrm{TN}' = \mathrm{TN} + \mathrm{FP}, \mathrm{FN}' = \mathrm{FN} + \mathrm{TP}$) v poměru k ceně filtrování

$$\mathrm{TCR} = \frac{c_{+-}(\mathrm{FN}+\mathrm{TP})}{c_{-+}\mathrm{FP}+c_{+-}\mathrm{FN}} = \frac{\mathrm{FN}+\mathrm{TP}}{\rho\,\mathrm{FP}+\mathrm{FN}}\text{, kde } \rho = \frac{c_{-+}}{c_{+-}}$$

Učení klasifikátoru

- *Klasifikátor* vlastně není $F: X \to \{C_1, \ldots, C_m\}$, ale přesněji

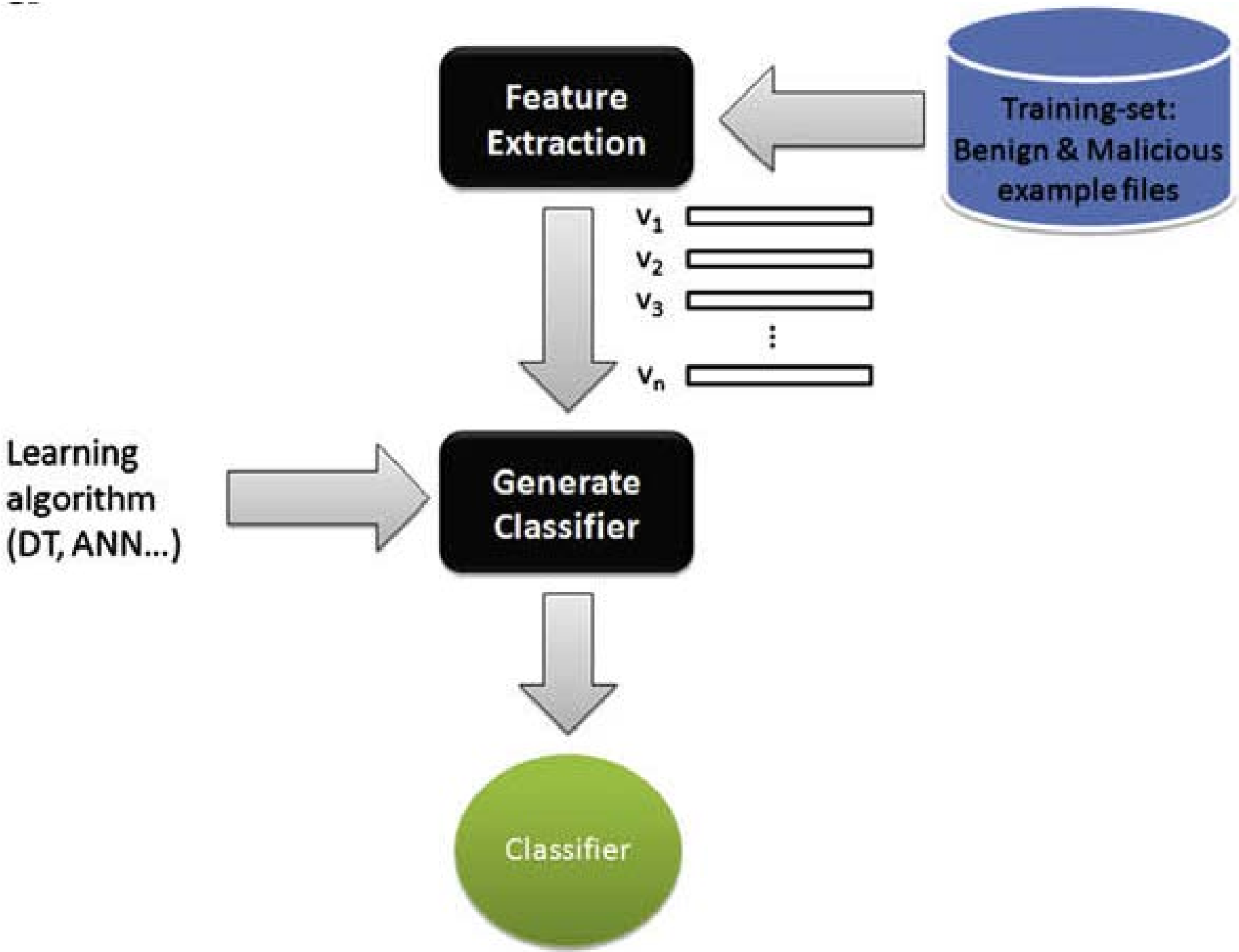
  $F: X \times \Theta \to \{C_1, \ldots, C_m\}$, Θ – přípustné kombinace *parametrů*

  - v každém okamžiku používáme $F(\cdot, \theta)$ s konkrétním θ

- Do *volby* $\theta \in \Theta$ zabudováváme zkušenost a znalost

  o kvalitě klasifikace $F(\cdot, \theta')$ při *předchozích volbách* $\theta' \in \Theta$

  - analogie k učení $\Longrightarrow$ používaný termín


Feature Extraction
Training-set: Benign & Malicious example files
v_1
v_2
v_3
v_n
Learning algorithm (DT, ANN...)
Generate Classifier
Classifier

Učení spamových filtrů

- Zdroje znalostí o kvalitě předchozích klasifikací:

1. *On-line feedback uživatele*: manuálně označí FN, FP
   - aktuální, ale na začátku nedostatečné množství
2. *Veřejně přístupná data*, většinou 0 ham (Spamarchive: 220000 mailů) nebo bez headerů (GenSpam: 41000)
   - zřídka vše (Spam track corpus: 92000, Spamassasin: 6000)

Přeučení trénovacími daty

- Cíl volby $\theta \in \Theta$ při učení: optimální klasifikace

  pro všechny možné *nové vstupy* klasifikátoru

- Učicí algoritmus používá jen předem dodaná

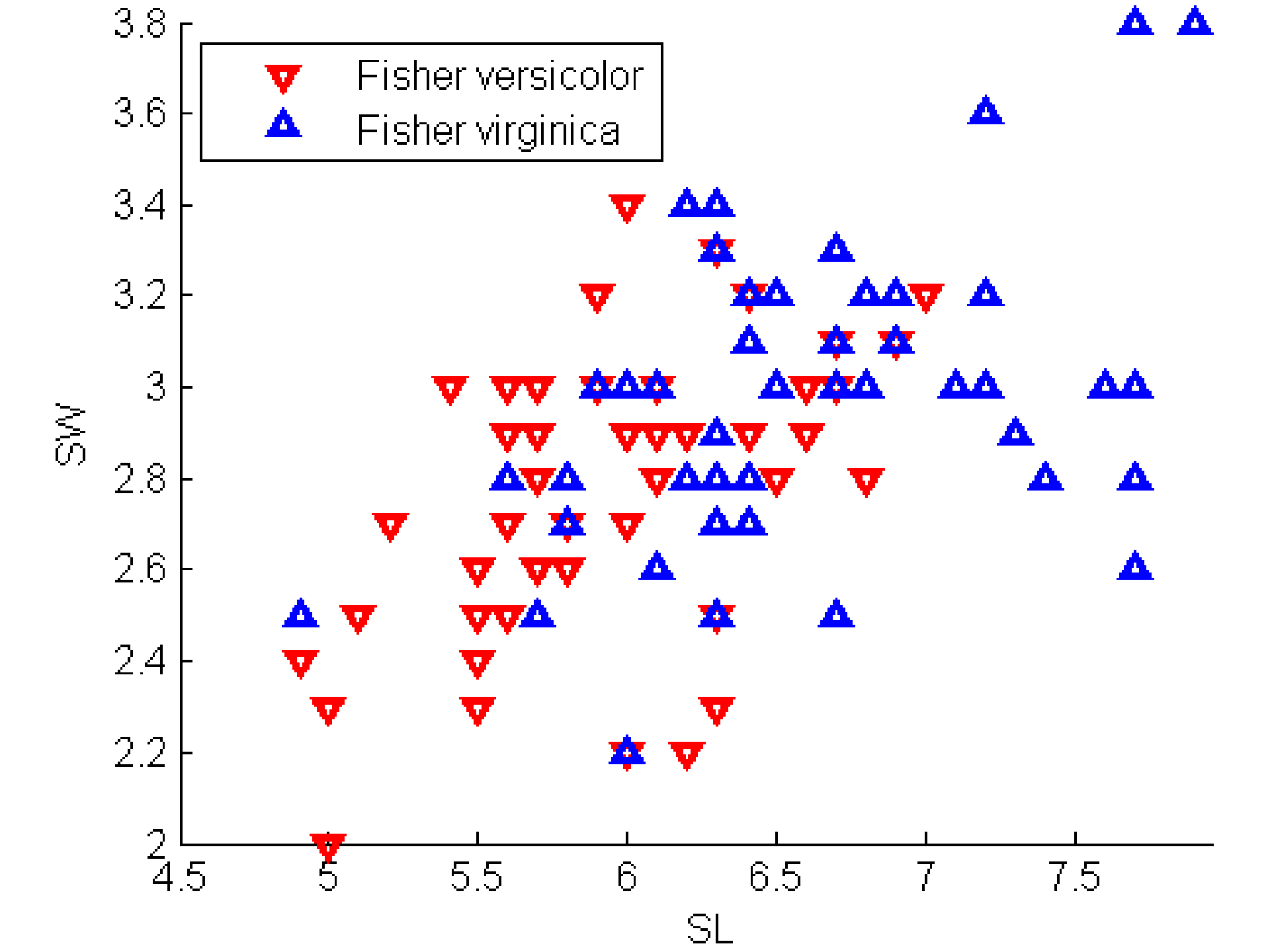
  učicí (trénovací) data $(x_1, c_1), \ldots, (x_q, c_q)$

  - $\Longrightarrow$ nebezpečí *přeučení*: vstupy podobné dodaným učicím

  klasifikuje přesně, nepodobné vstupy hodně nepřesně

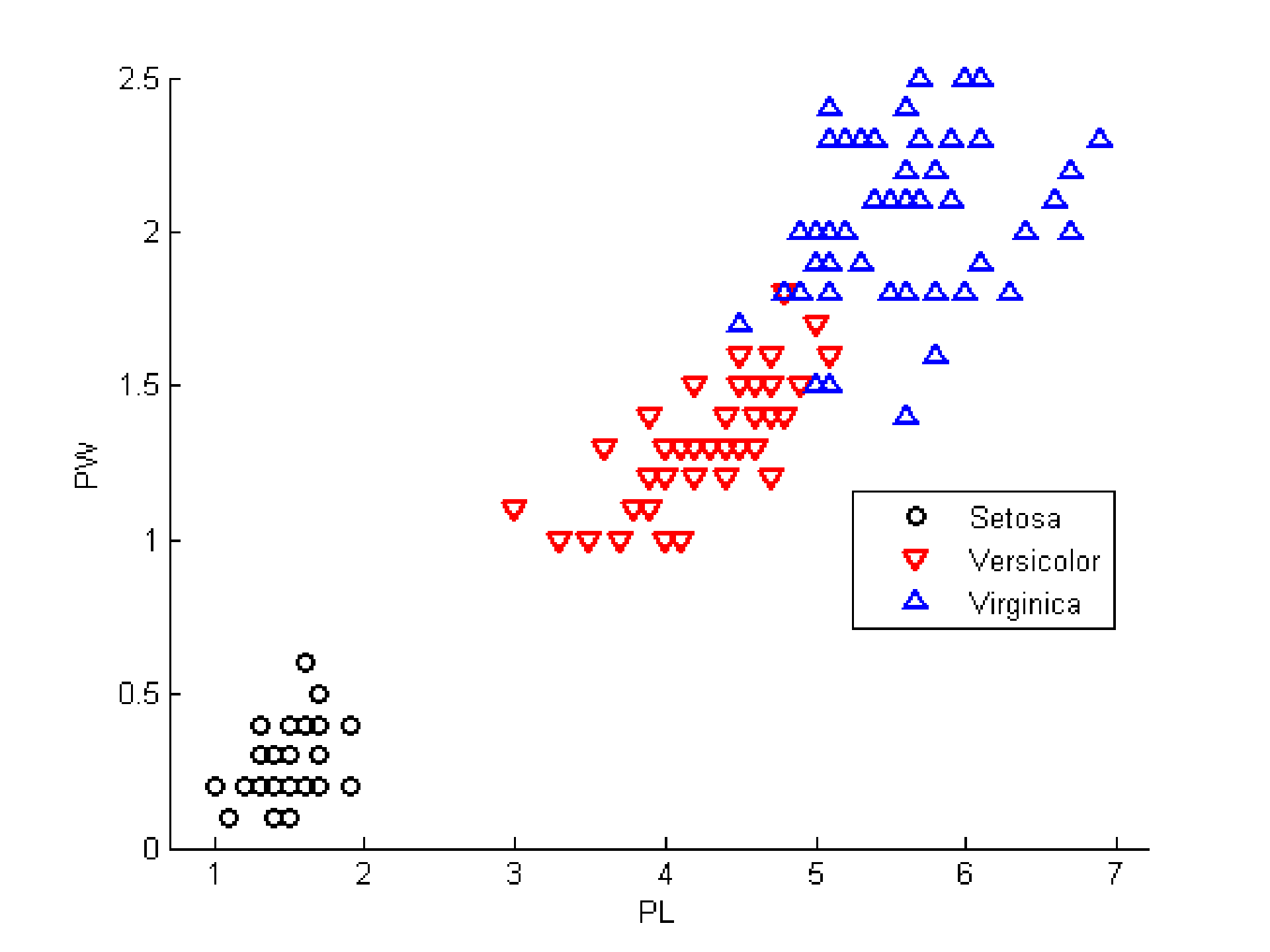
Tvar hranice mezi třídami

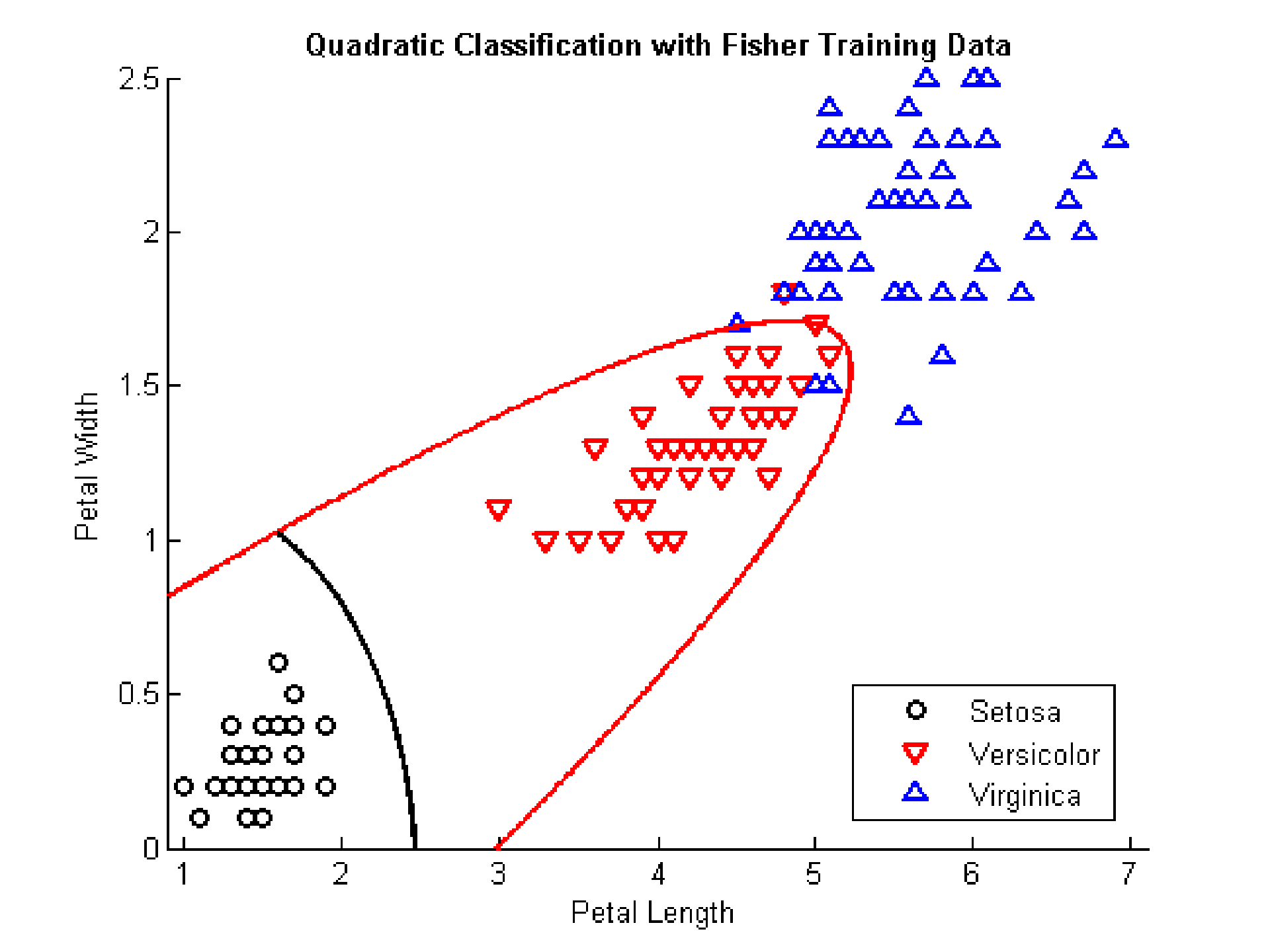
- Obvykle *složitý*, často vlivem *zašumění dat*

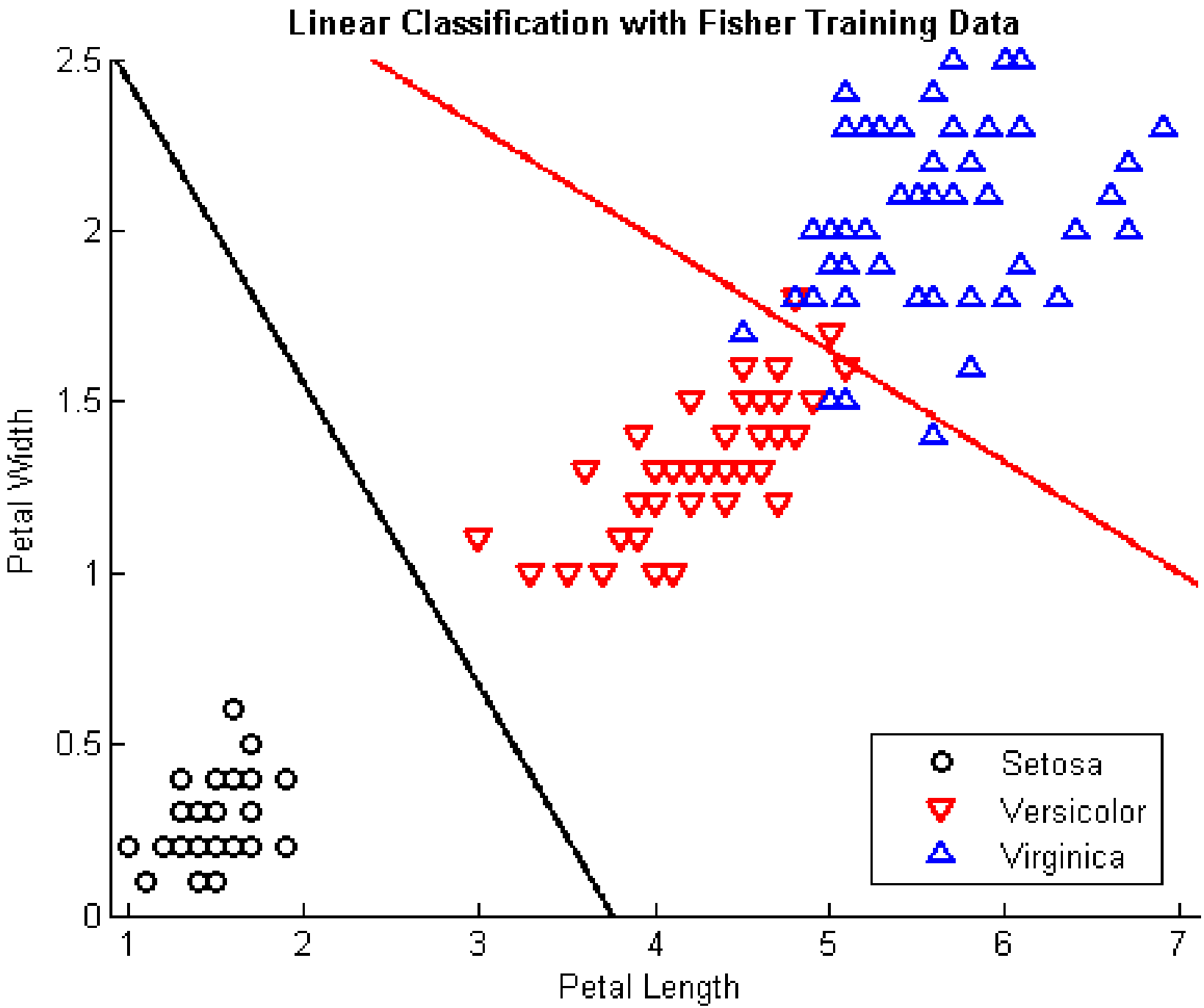

Fisher versicolor
Fisher virginica
SW
SL
3.8
3.6
3.4
3.2
3
2.8
2.6
2.4
2.2
2
4.5
5
5.5
6
6.5
7
7.5

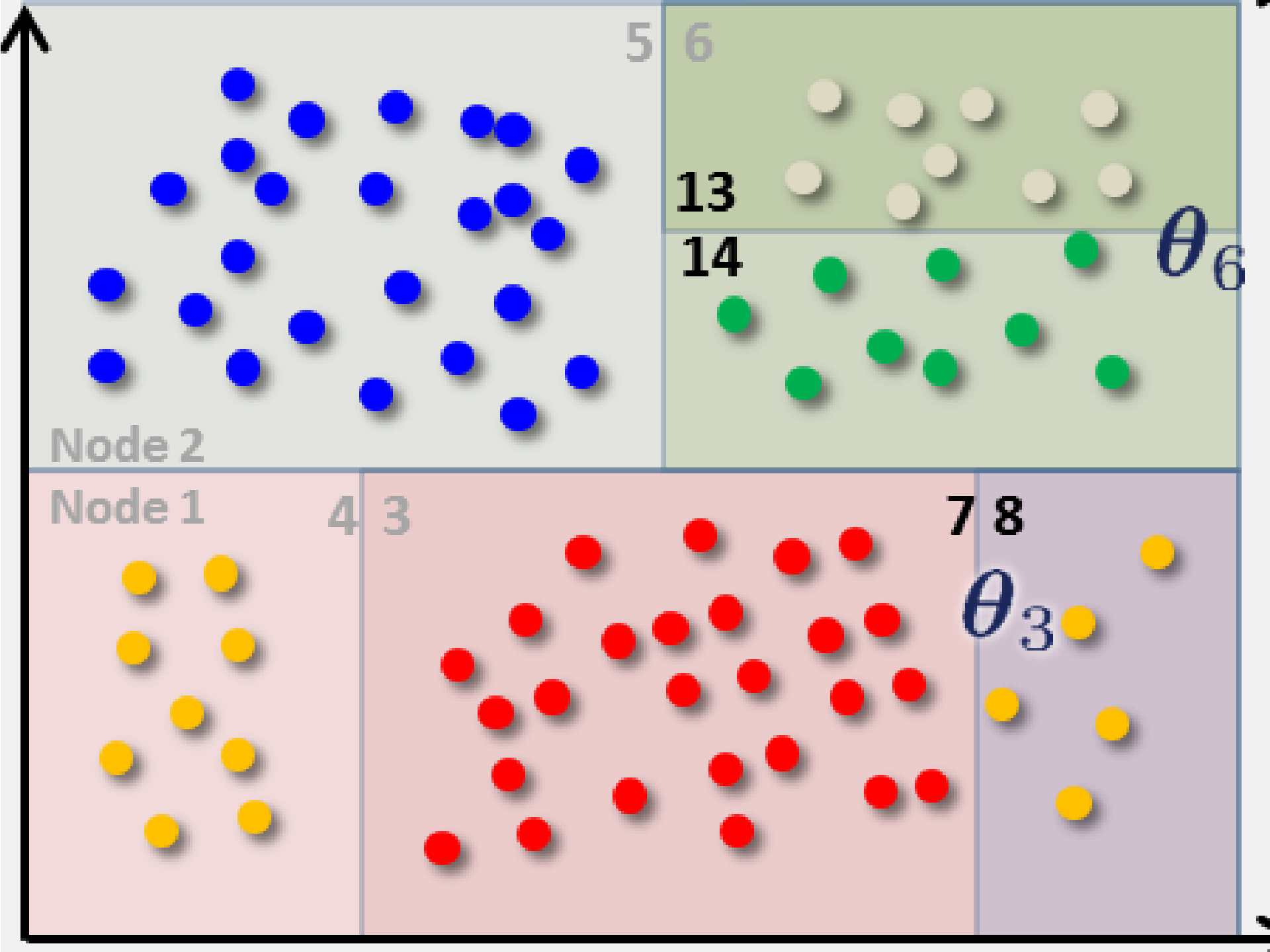
Tvar hranice mezi třídami

- Obvykle *složitý*, často vlivem *zašumění dat*
- Většina klasifikátorů předpokládá určitý tvar hranice
  - např. kvadratická plocha, po částech lineární, …
  - nejčastěji se předpokládá lineární plocha – nadrovina


2.5
2
1.5
1
0.5
0
PW
1
2
3
4
5
6
7
PL
Setosa
Versicolor
Virginica


Quadratic Classification with Fisher Training Data
Petal Width
Petal Length
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
Setosa
Versicolor
Virginica


Linear Classification with Fisher Training Data
Petal Width
Petal Length
2.5
2
1.5
1
0.5
0
1
2
3
4
5
6
7
Setosa
Versicolor
Virginica

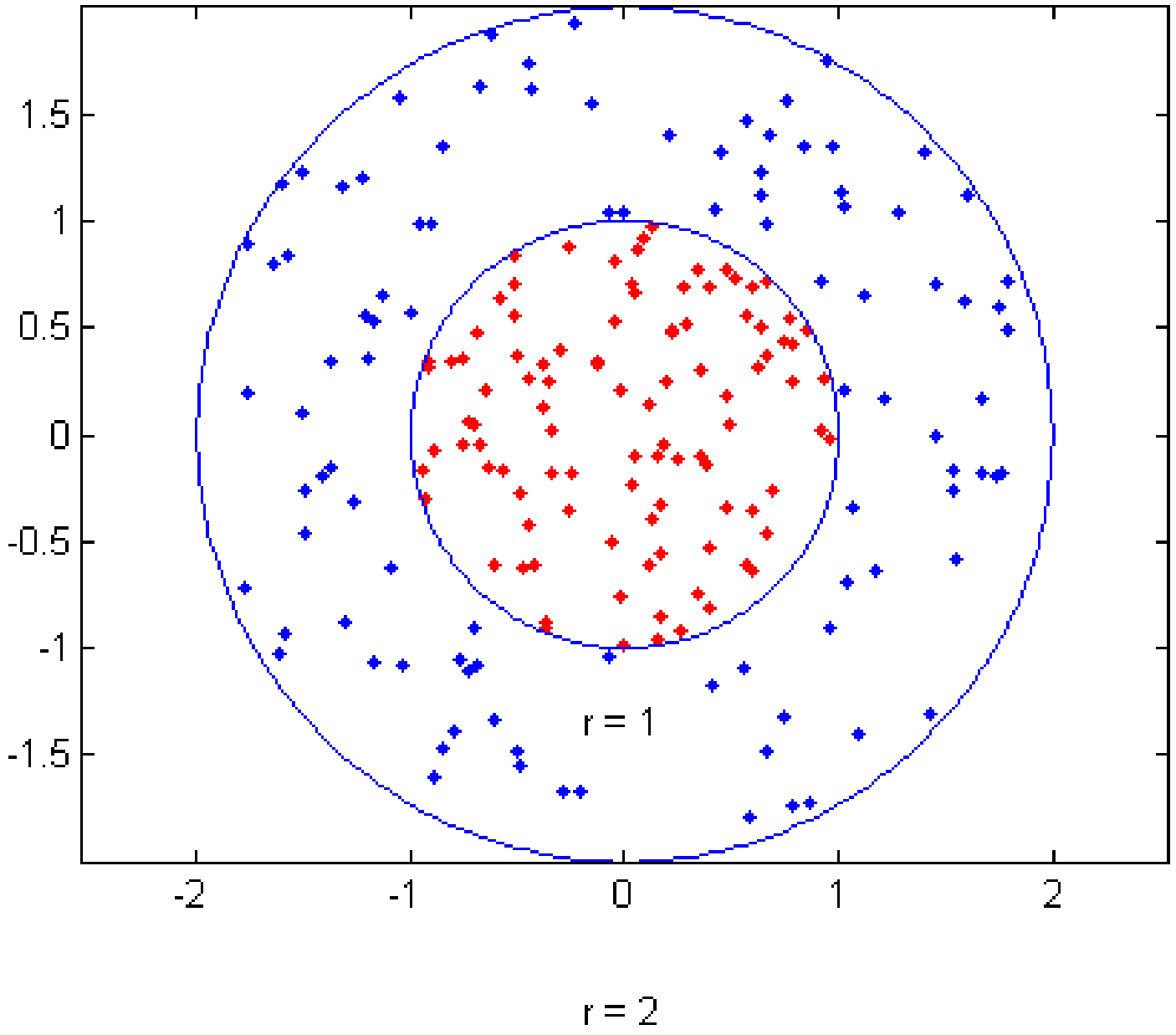

5
6
13
14
θ_6
Node 2
Node 1
4
3
7
8
θ_3

Tvar hranice mezi třídami

- *Obvykle složitý*, často vlivem *zašumění dat*
- Většina klasifikátorů předpokládá určitý tvar hranice
  - např. kvadratická plocha, po částech lineární, …
  - nejčastěji se předpokládá lineární plocha – nadrovina
- *Lineární separabilita*: $C_1 \cap \{x_1, \ldots x_q\}$ a $C_2 \cap \{x_1, \ldots x_q\}$ lze oddělit nadrovinou
  - u reálných dat vzácná, dosažitelná předzpracováním

Čím dosáhnout lineární separability?

- *Transformací* do prostoru vyšší dimenze
  - příklad: $C_1 = \left\{(x, y): \sqrt{x^2 + y^2} < r\right\}, C_2 = \left\{(x, y): \sqrt{x^2 + y^2} > r\right\}$


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

Čím dosáhnout lineární separability?

- *Transformací* do prostoru vyšší dimenze
  - příklad: $C_1 = \left\{(x,y): \sqrt{x^2+y^2} < r\right\}, C_2 = \left\{(x,y): \sqrt{x^2+y^2} > r\right\}$

  transformujeme na lineárně separabilní (x,y,z), $z = \sqrt{x^2+y^2}$

- *Potřebuji znát tvar hranice?* (kružnice, elipsa,…)
  - *NE*: má-li konečná množina M dimenzi $d \geq \#M - 1$, pak $C_1 \subset M$,

  $C_1 \neq M, \emptyset$ ($2^{\#M} - 2$ možností) oddělíme od $C_2 = M \setminus C_1$ nadrovinou

A jak postupovat v praxi?

- Mám *data* $x_1, \dots, x_n \epsilon X, \dim X \ll n$, např. $X = \mathbb{R}^m, m \ll n$
- Vezmu *jádro (kernel)*: *zobrazení* $\kappa: X^2 \to \mathbb{R}$, $\kappa(x, y) = \kappa(y, x)$,

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

  - např. $\kappa(x, y) = e^{-\|x-y\|^2}$, $\kappa(x, y) = (x^\intercal y + c)^q$
- Zavedu *množinu* $M = \{\varphi_1, \dots, \varphi_n\}, \varphi_i: X \to \mathbb{R}, \varphi_i(x) = \kappa(x_i, x)$
  - pak je dimenze $M = n = \#M$ (např. 10000)

Trik pro konstrukci nadroviny

- Konstruovat nadrovinu pro *prostor dimenze* n vyžaduje:

  1. vektorový prostor V dimenze n, 2. skalární součin na V

1. $V =$ *lineární obal* $M = \{\sum_{i=1}^{n} a_i \varphi_i \,|a_1, \dots, a_n \in \mathbb{R}\}$

2. $\langle \sum_{i=1}^{n} a_i \varphi_i \, , \sum_{i=1}^{n} b_i \varphi_i \rangle = \sum_{i,j=1}^{n} a_i b_j \kappa(x_i, x_j)$, tedy výpočtem

   pomocí x_i, x_j *dimenze* $\ll n$ zkonstruujeme nadrovinu

   v prostoru V vysoké dimenze n (kernel trick)

Odlišnost klasifikace od regrese

- *Regrese* $F: X \times \Theta \to \mathbb{R}$, *výstupy* jsou reálná čísla, lze je *uspořádat*, měřit jejich *velikost / vzdálenost*
  - Hlavní typy: lineární regrese, nelineární regrese.

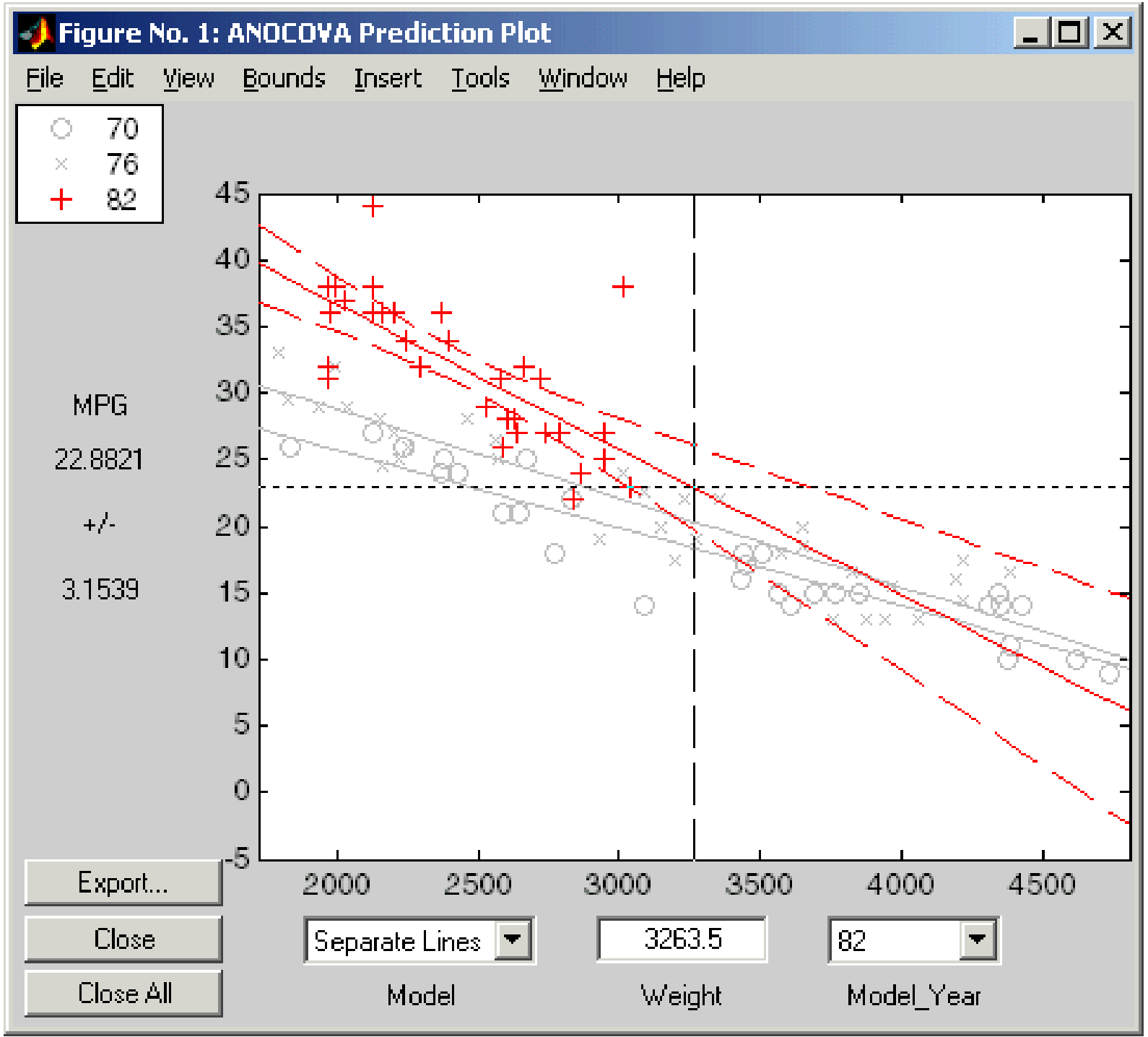
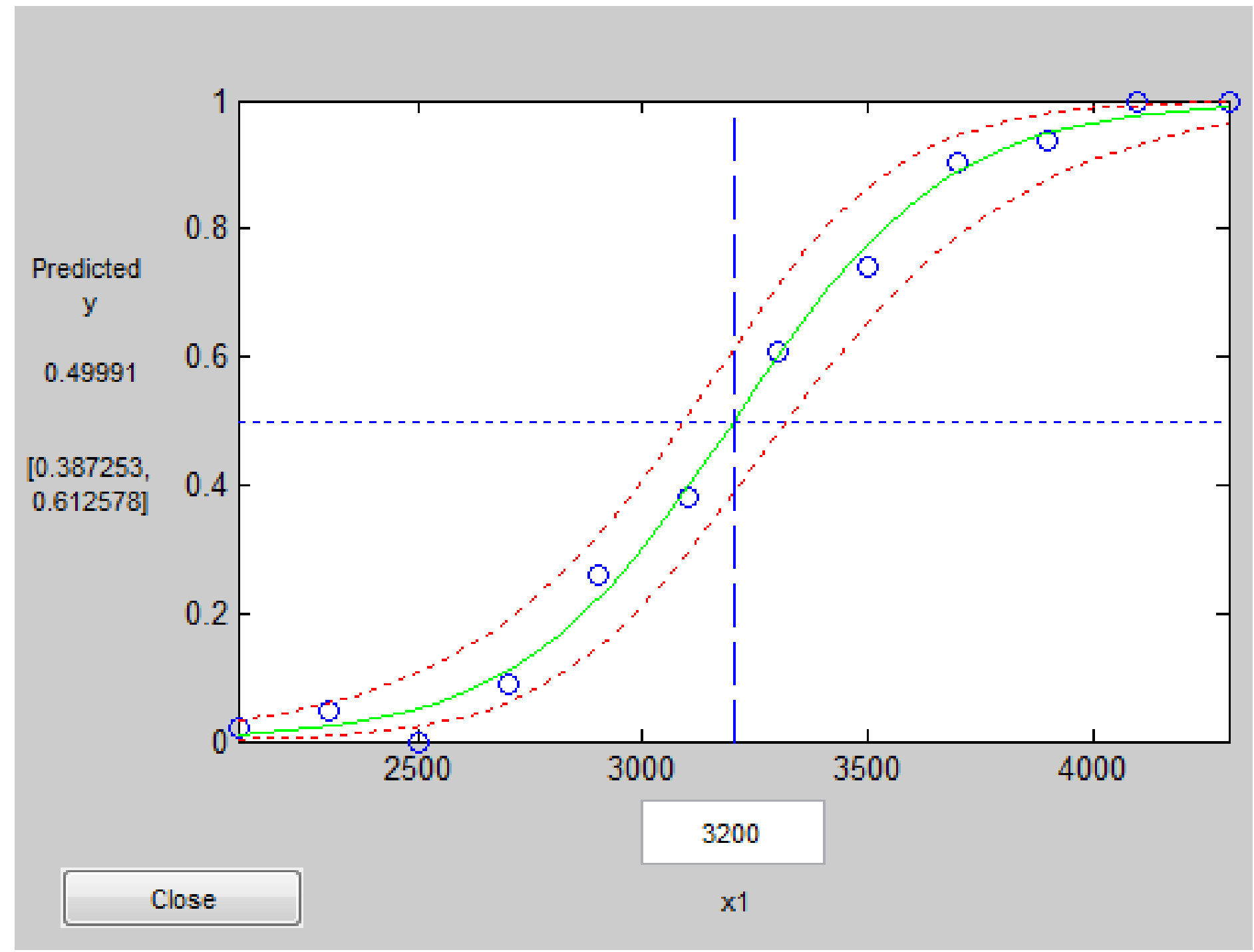


Figure No. 1: ANOCOVA Prediction Plot
File Edit View Bounds Insert Tools Window Help
70
76
82
MPG
22.8821
+/-
3.1539
45
40
35
30
25
20
15
10
5
0
-5
2000
2500
3000
3500
4000
4500
Export...
Close
Close All
Separate Lines
Model
3263.5
Weight
82
Model_Year


Predicted
y
0.49991
[0.387253,
0.612578]
1
0.8
0.6
0.4
0.2
0
2500
3000
3500
4000
3200
x1
Close

Odlišnost klasifikace od regrese

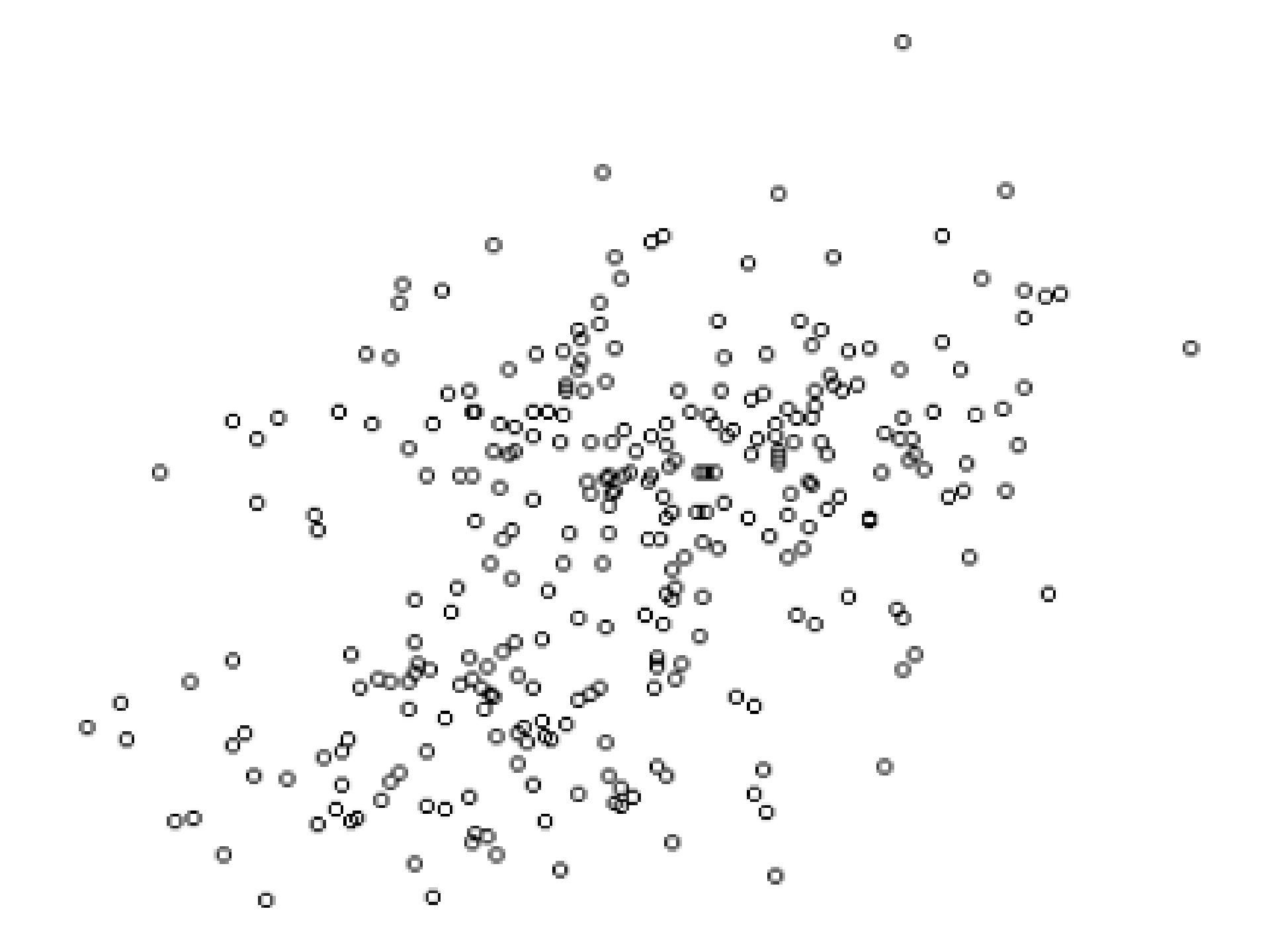
- *Regrese* $F: X \times \Theta \to \mathbb{R}$, *výstupy* jsou reálná čísla, lze je *uspořádat*, měřit jejich *velikost / vzdálenost*
  - Hlavní typy: lineární regrese, nelineární regrese.
- Klasifikace $F: X \times \Theta \to \{C_1, \ldots, C_m\}$, výstupy jsou třídy, nemají přirozené uspořádání ani velikost, vzdálenost
  - přechodná metoda – $C_1, \ldots, C_m$ uspořádané: ordinální regrese

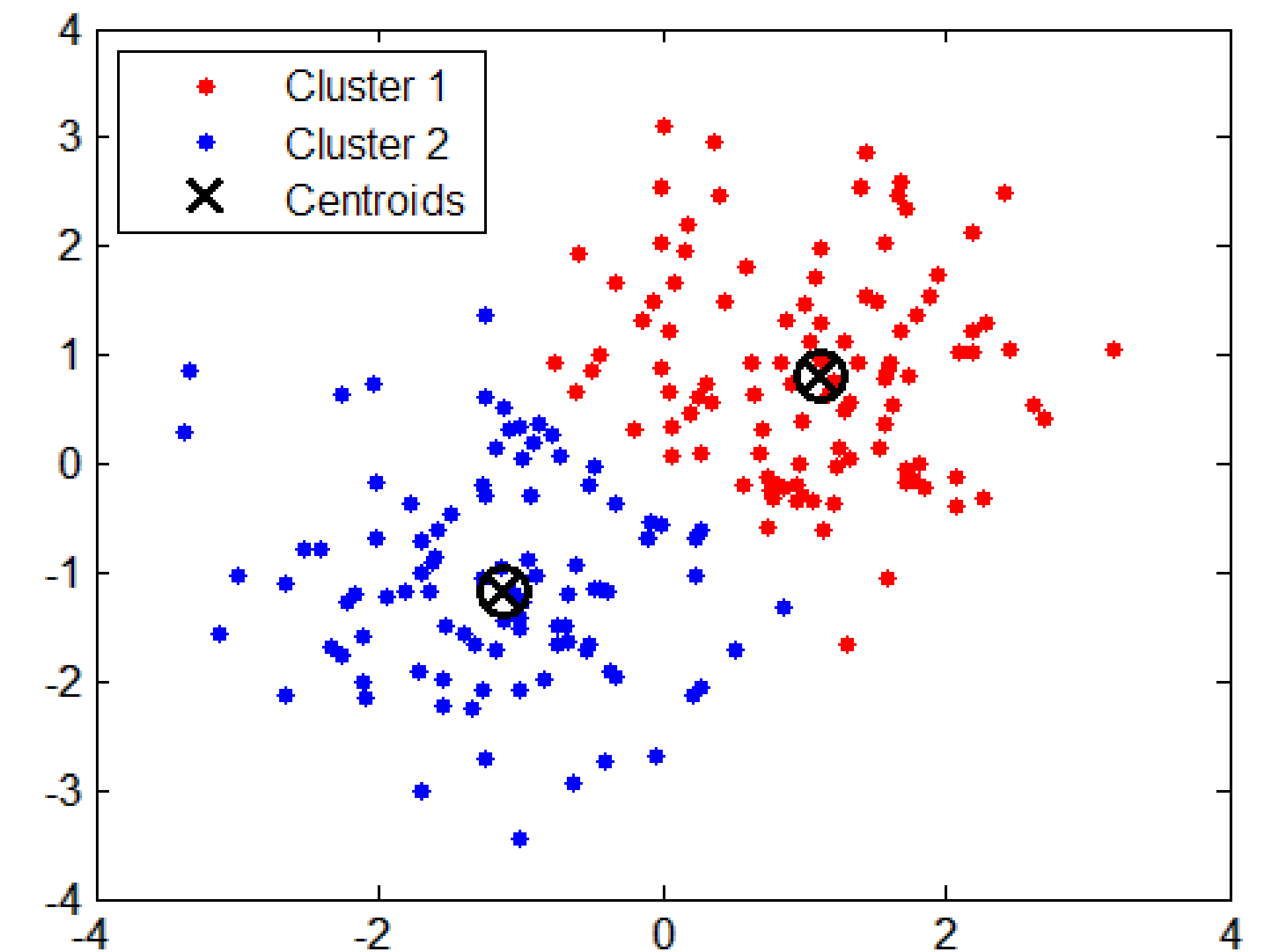
Odlišnost klasifikace od shlukování

- Shlukování také pracuje s *prostorem příznaků*

+ třídami (= *shluky*) $C_1, \ldots, C_m$, ale ty *neznáme předem*,

jsou *výsledkem* shlukování („unsupervised classification“)




Cluster 1
Cluster 2
Centroids
4
3
2
1
0
-1
-2
-3
-4
-4
-2
0
2
4

Odlišnost klasifikace od shlukování

- Shlukování také pracuje s *prostorem příznaků*

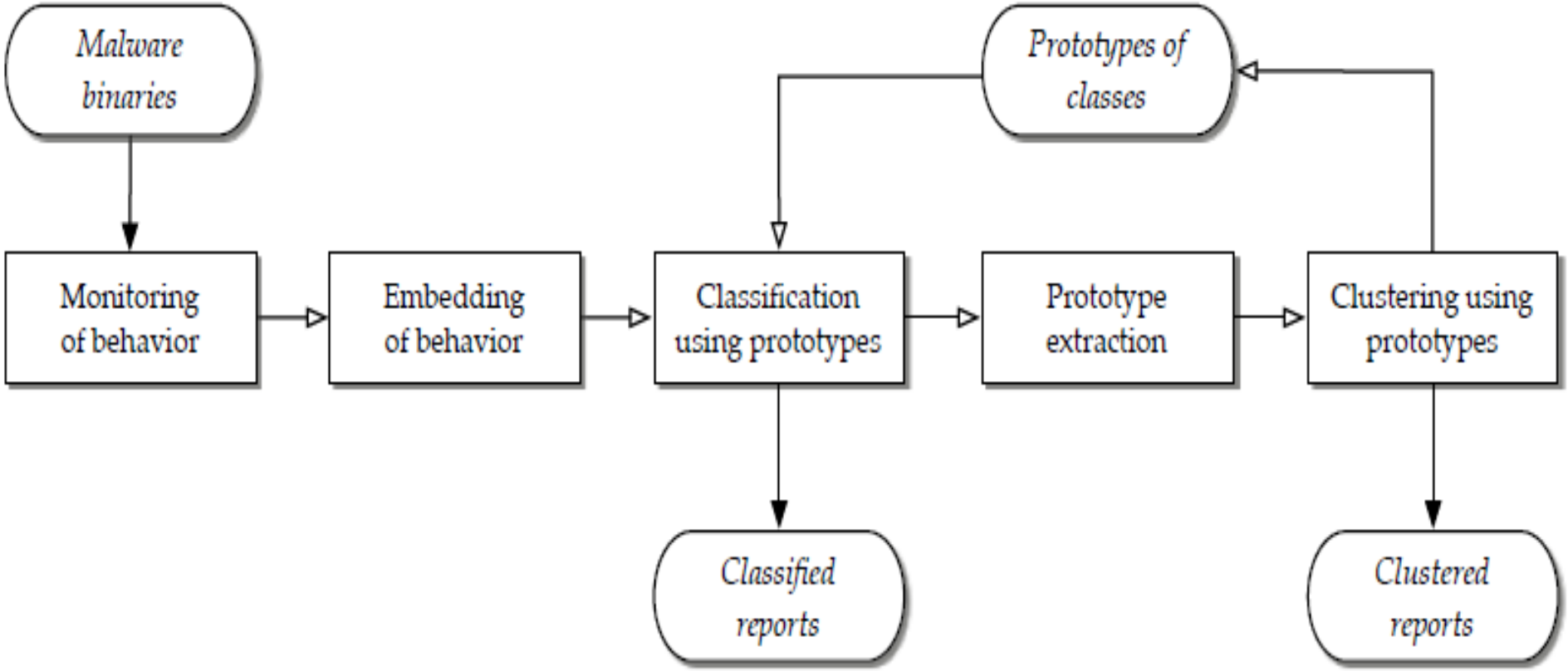
  + třídami (= *shluky*) $C_1, \ldots, C_m$, ale ty *neznáme předem*,

  jsou *výsledkem* shlukování („unsupervised classification“)

- Při klasifikaci jsou třídy známy předem

  - klasifikátor není výsledkem klasifikace, ale

    meziproduktem pro klasifikaci budoucích vstupů


Malware binaries
Prototypes of classes
Monitoring of behavior
Embedding of behavior
Classification using prototypes
Prototype extraction
Clustering using prototypes
Classified reports
Clustered reports

Regrese a shlukování při doporučování

- 2 důležité role regrese v doporučovacích systémech:

1. *Hodnocení* produktů uživateli − reálné nebo ordinální

2. *Měření podobnosti* mezi uživateli nebo produkty

  - obojí používané hlavně při kolaborativním filtrování

- Podobnost je základem pro shlukování ⟹ *shluky*

v doporučovacích systémech: *A. uživatelů, B. produktů*