EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ, MLÁDEŽE A TĚLOVÝCHOVY

Název projektu	Rozvoj vzdělávání na Slezské univerzitě v Opavě
Registrační číslo projektu	CZ.02.2.69/0.0./0.0/16_015/0002400

Dolování dat

Strojové učení

Jan Górecki

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Obsah přednášky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Co je to Strojového učení
- Typy učení
- Metody učení
- Učení jako prohledávání
- Učení jako aproximace

Strojové učení

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

The field of machine learning is concerned with the question of how to construct computer programs that automatically improve with experience.

(Mitchell, 1997)

Strojové učení

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Things learn when they change their behavior in a way that makes them perform better in a future.

(Witten, Frank, 1999)

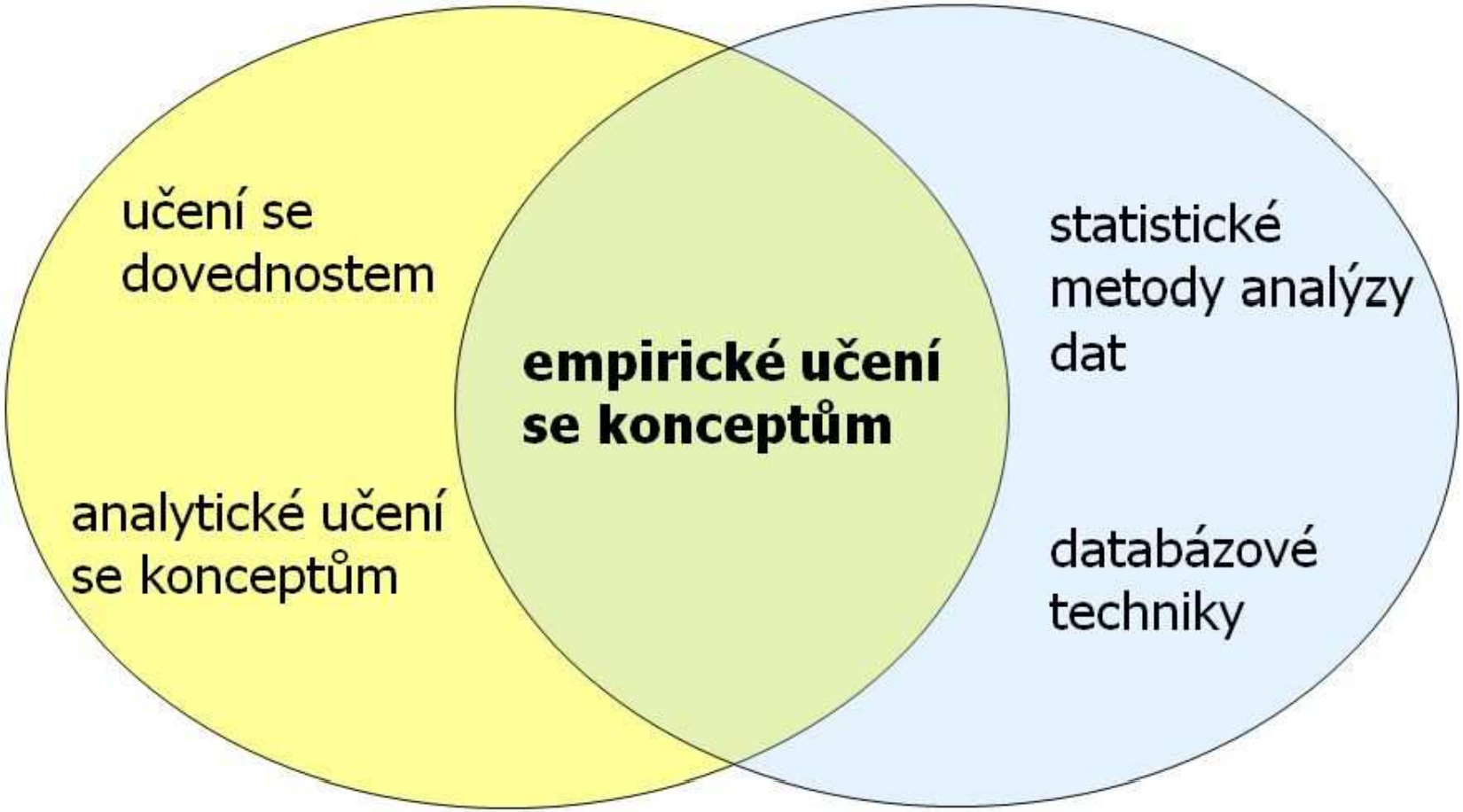
Typy učení

SLEZSKÁ
UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINÉ

- učení se konceptům (knowledge acquisition)
- učení se dovednostem (skill refinement).

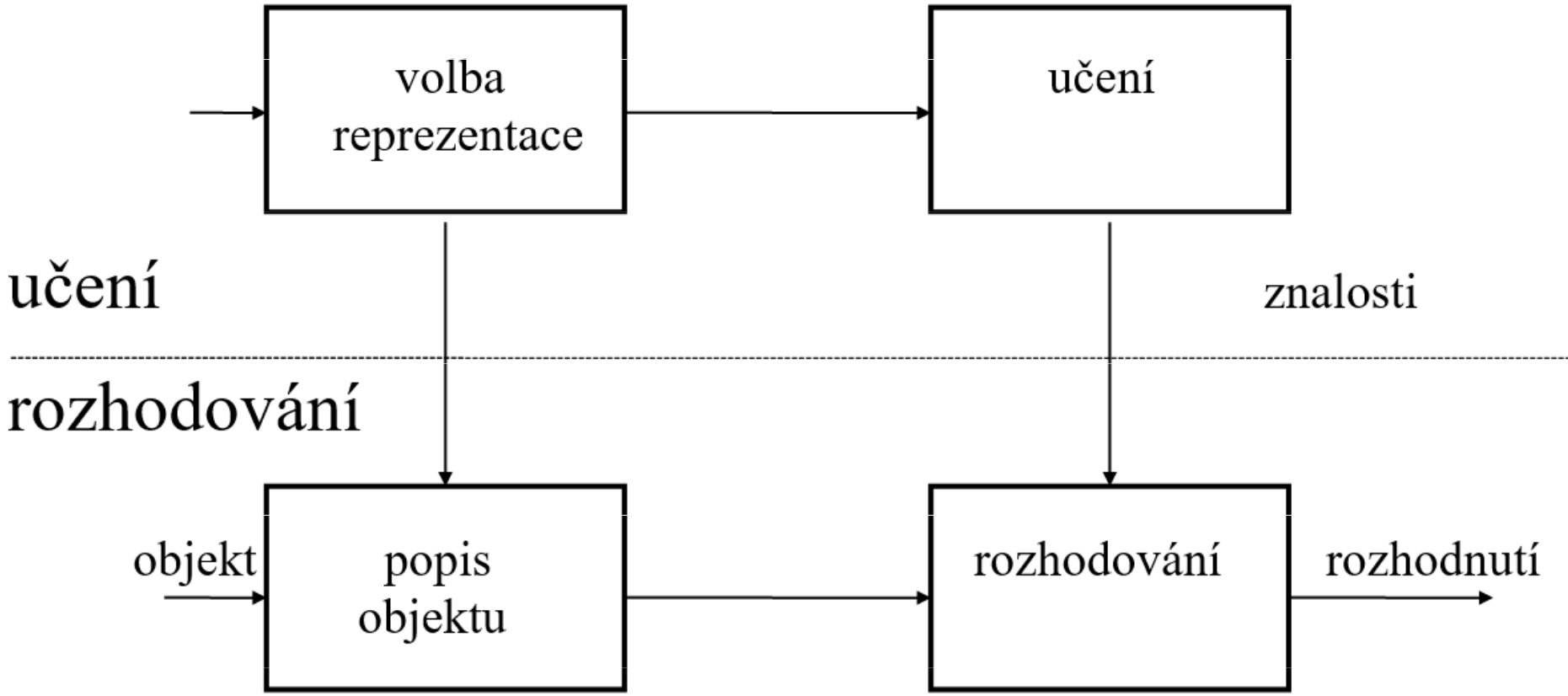
Vztah strojového učení a dobývání znalostí

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



Obecné schéma učícího se systému

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



Metody učení

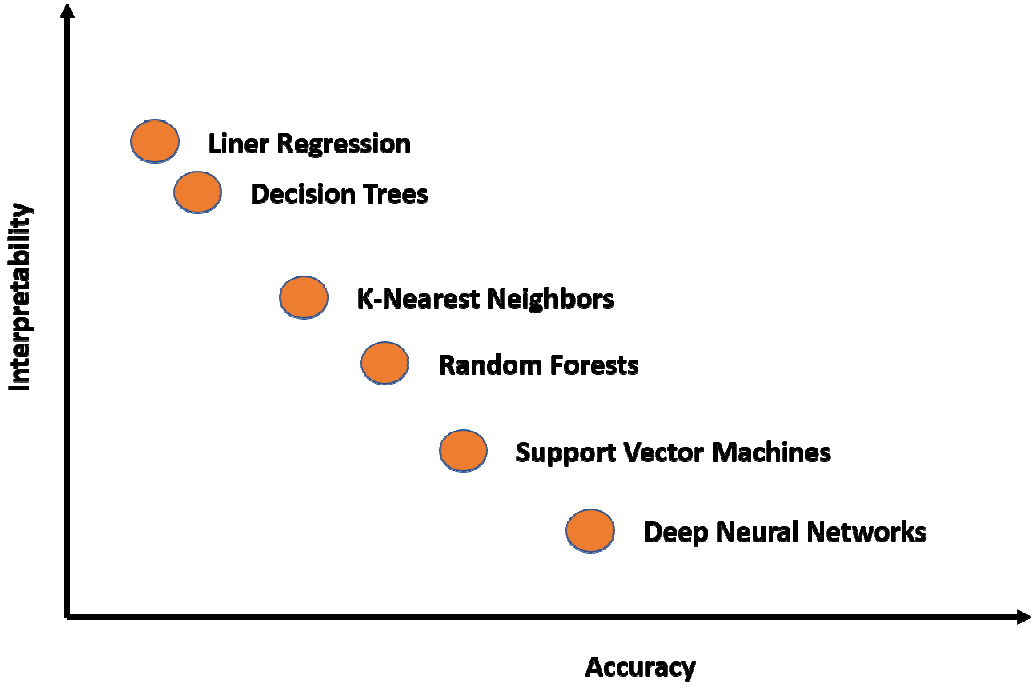
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

1. učení zapamatováním (rote learning neboli biflování),
2. učení se z instrukcí (learning from instruction, learning by being told),
3. učení se z analogie (learning by analogy, instance-based learning, lazy learning),
4. učení na základě vysvětlení (explanation-based learning),
5. učení se z příkladů (learning from examples),
6. učení se z pozorování a objevováním (learning from observation and discovery),

Metody učení

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- **statistické metody** - regresní metody, diskriminační analýza, shluková analýza,
- **symbolické metody umělé inteligence** - rozhodovací stromy a pravidla, případové usuzování (CBR),
- **subsymbolické metody umělé inteligence** - neuronové sítě, bayesovské sítě nebo genetické algoritmy.



Informace o správnosti učení

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Může mít podobu:

- příkladů zařazených do tříd (**učení s učitelem - supervised learning**)
- odměny za správné chování a tresty za chování nesprávné (**reinforcement learning**)
- nepřímé náznaky odvozené s chování učitele (**apprenticeship learning**)
- žádné (**učení bez učitele - unsupervised learning**)

Metody učení

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- **empirické** – vychází se z velkého množství příkladů a žádných (nebo jen mála) počátečních znalostí
- **analytické** – vychází se z velkého množství počátečních znalostí a jen několika (ilustračních) příkladů

Principy empirického učení z dat – 1.

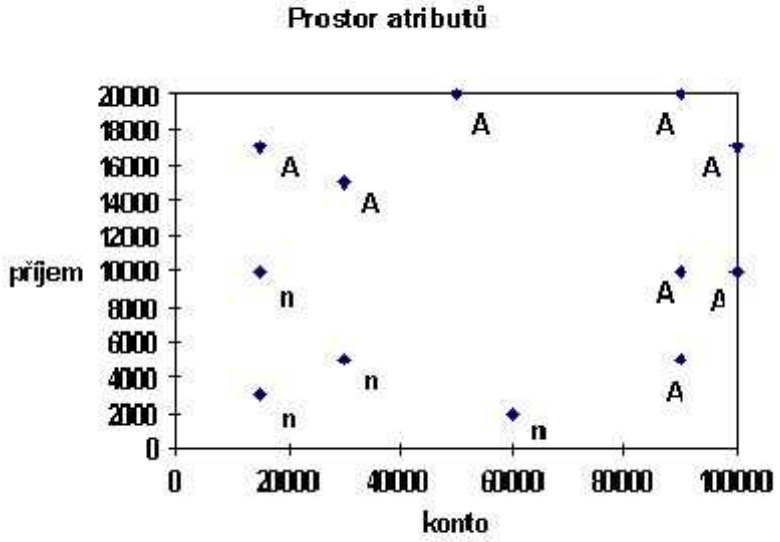
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- objekty, patřící do téže třídy mají podobné charakteristiky (**učení na základě podobnosti, similarity-based learning**)
  - příklady téže třídy vytvářejí shluky v prostoru atributů
  - cílem učení je tyto shluky nalézt a popsat

Principy empirického učení z dat – 2.

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

klient	konto	příjem	půjčit
01	15000	3000	ne
02	15000	10000	ne
03	15000	17000	Ano
04	30000	5000	ne
05	30000	15000	Ano
06	50000	20000	Ano
07	60000	2000	ne
08	90000	5000	Ano
09	90000	10000	Ano
10	90000	20000	Ano
11	100000	10000	Ano
12	100000	17000	Ano

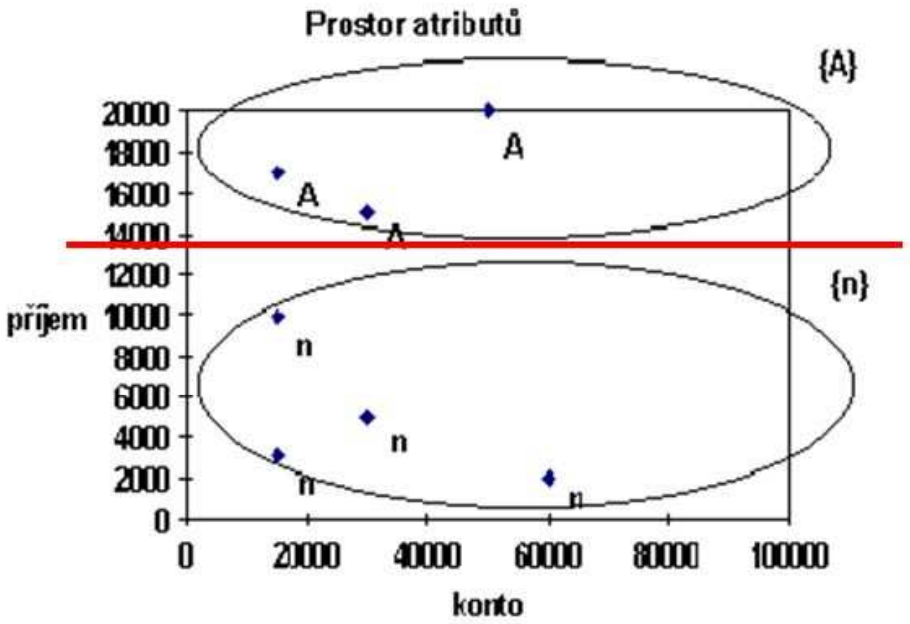


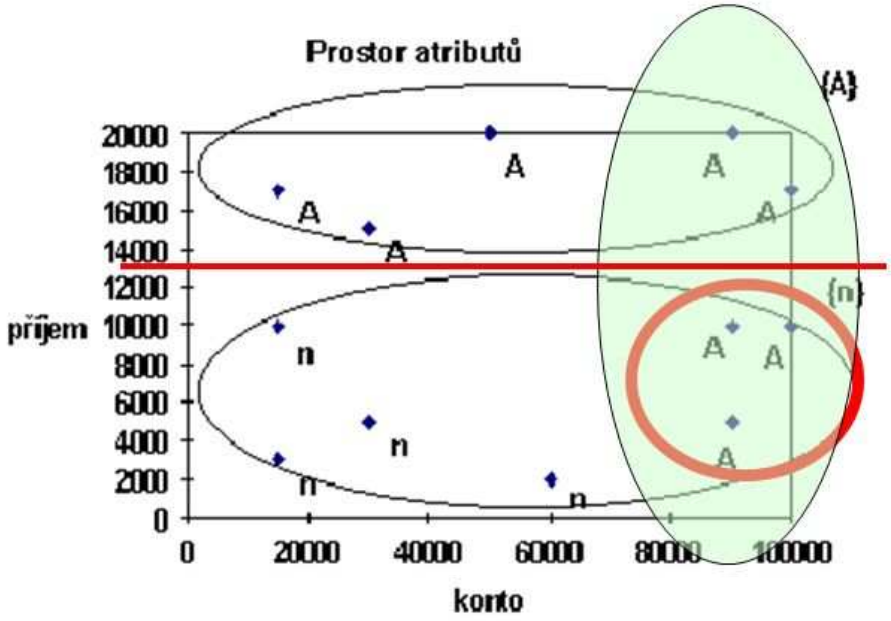
- Nebezpečí „garbage in, garbage out“
- Důležitost přípravy a předzpracování dat

Principy empirického učení z dat – 3.

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- z konečného počtu příkladů odvozujeme obecné znalosti (**induktivnost**)





Principy empirického učení z dat – 4.

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Příklady rozděleny do 2 (někdy 3) množin:
  - **trénovací data** pro vytvoření modelu
  - (**validační data** pro doladění parametrů)
  - **testovací data** pro otestování modelu

Obecná definice strojového učení

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Analyzovaná data:

$$D = \begin{bmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1m} \\ x_{21} & x_{22} & \cdots\cdots & x_{2m} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots\cdots & x_{nm} \end{bmatrix}$$

Řádky tabulky reprezentují sledované **objekty**

Sloupce datové tabulky odpovídají **atributům**

Obecná definice strojového učení (s učitelem)

Přidáme-li cílový atribut do datové tabulky, získáme data vhodná pro použití některé metody učení s učitelem (tzv. **trénovací data**).

$$\mathbf{D}_{TR} = \begin{bmatrix} x_{11} & x_{12} & \cdots\cdots & x_{1m} & y_1 \\ x_{21} & x_{22} & \cdots\cdots & x_{2m} & y_2 \\ \vdots & \vdots & & \vdots & \vdots \\ x_{n1} & x_{n2} & \cdots\cdots & x_{nm} & y_n \end{bmatrix}$$

Objekt (trénovací příklad) z této tabulky budeme značit

$$\mathbf{o_i} = [\, \mathbf{x_i}\, ,\, y_i \,]$$

Klasifikační úloha: hledáme **znalosti** (reprezentované rozhodovací funkcí *f*), které by umožňovaly k hodnotám vstupních atributů nějakého objektu přiřadit vhodnou hodnotu atributu cílového

$$f: \mathbf{x} \rightarrow y.$$

Obecná definice strojového učení (s učitelem)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

V průběhu klasifikace se tedy pro hodnoty vstupních atributů ***x*** nějakého objektu odvodí hodnota cílového atributu:

$$\hat{y} = f(\mathbf{x}).$$

Odvozená hodnota $\hat{y}$ se pro objekty z trénovacích dat může lišit od skutečné hodnoty y. Můžeme tedy pro každý objekt $\mathbf{o}_i \in D_{TR}$ vyčíslit *chybu klasifikace* $Q_f(\boldsymbol{o}_i, \hat{y}_i)$. V případě numerického atributu C může být touto chybou například čtverec rozdílu skutečné a odvozené hodnoty cílového atributu

$$Q_f(\mathbf{o_i}, \hat{y}_i) = (y_i - \hat{y}_i)^2,$$

v případě kategoriálního atributu C může být touto chybou informace o tom že se odvozená a skutečná hodnota vzájemně liší,

$$Q_f(\mathbf{o_i}, \hat{y}_i) = \begin{cases} 1 \text{ pro } y_i \neq \hat{y}_i \\ 0 \text{ pro } y_i = \hat{y}_i \end{cases}$$

Chyba na trénovacích datech

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Pro celou trénovací množinu *DTR* pak můžeme vyčíslit souhrnnou chybu $Err(f,D_{TR})$, například jako střední chybu

$$Err(f,D_{TR}) = \frac{1}{n}\sum_{i=1}^{n} Q_f(\mathbf{o_i}, \hat{y}_i)$$

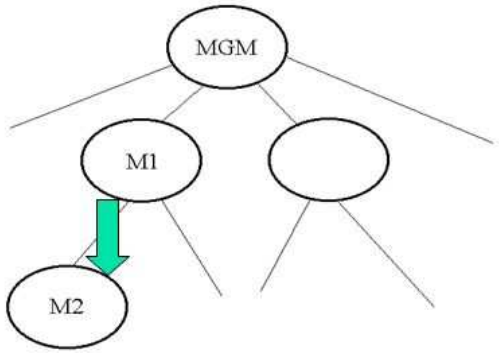
Cílem učení je nalézt takové znalosti f^*, které by minimalizovaly tuto chybu

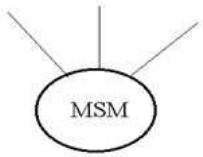
$$Err(f^*,D_{TR}) = \min_f Err(f,D_{TR}).$$

Učení jako prohledávání

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- hledáme strukturu i parametry modelu
- např. Rozhodovací stromy, Rozhodovací pravidla





Modely jako popisy shluků:

- MGM - nejobecnější model (jeden shluk pro všechno)
- MSM - nejspeciálnější model(y) (co příklad to shluk)
- M1 obecnější než M2, M2 je speciálnější než M1

Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

příjem	konto	auto	bydlení	úvěr
vysoký	vysoké	ano	vlastní	Ano
vysoký	vysoké	ano	vlastní	Ano
nizký	nízké	ano	nájemní	Ne
vysoký	vysoké	ne	nájemní	Ano

Nejvíce obecné pravidlo (MSM):
If cokoli **then** úvěr = Ano

První specializace MSM (M1) :
If příjem = vysoký **then** úvěr = Ano

…

Evaluace:
„Kolik objektů (řádků) v datech porušuje dané pravidlo.“

Nejvíce speciální pravidlo (MGM):
If příjem = vysoký & konto = vysoké & auto = ano & bydlení = vlastní **then** úvěr = Ano

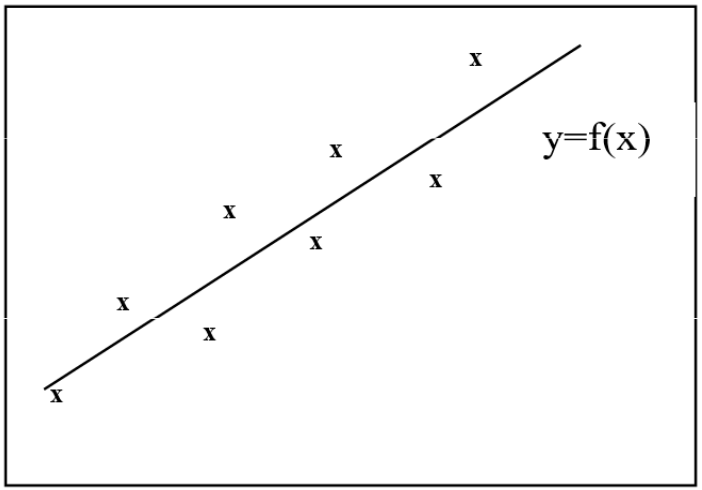
Učení jako aproximace

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Hledáme „pouze“ parametry modelu

Příklad:

na základě hodnot funkce v konečném počtu bodů snažíme zrekonstruovat její obecnou podobu



$$f(x) = q_1 x + q_0$$

Metoda nejmenších čtverců

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Hledání minima celkové odchylky

$$\min \sum_i (y_i - f(x_i))^2$$

se převádí na řešení rovnice

$$\frac{d}{dq} \sum_{i=1}^{n} \left(y_i - f(\mathbf{x_i}) \right)^2 = 0$$

Řešení

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

1) **analytické** (známe typ funkce)

řešení soustavy rovnic pro parametry funkce

$f(x) = q_1 x + q_0$ ⇒

$$\frac{\partial}{\partial q_0} \sum_{i=1}^{n} \left(y_i - (q_1 x_i + q_0)\right)^2 = -2 \sum_{i=1}^{n} y_i + 2q_1 \sum_{i=1}^{n} x_i + 2q_0 n$$

$$\frac{\partial}{\partial q_1} \sum_{i=1}^{n} \left(y_i - (q_1 x_i + q_0)\right)^2 = -2 \sum_{i=1}^{n} x_i y_i + 2q_1 \sum_{i=1}^{n} x_i^2 + 2q_0 \sum_{i=1}^{n} x_i$$

⇒

$$q_0 = \frac{(\sum_i y_i)(\sum_i x_i^2) - (\sum_i x_i y_i)(\sum_i x_i)}{n(\sum_i x_i^2) - (\sum_i x_i)^2}$$

$$q_1 = \frac{n(\sum_i x_i y_i) - (\sum_i x_i)(\sum_i y_i)}{n(\sum_i x_i^2) - (\sum_i x_i)^2}$$

2) **numerické** (neznáme typ funkce)

gradientní metody

Gradientní metody

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

$$\nabla \mathrm{Err}(\mathbf{q}) = \left[\frac{\partial \mathrm{Err}}{\partial q_0}, \frac{\partial \mathrm{Err}}{\partial q_1}, \ldots, \frac{\partial \mathrm{Err}}{\partial q_Q} \right].$$

Modifikace znalostí $\mathbf{q} = [q_0, q_1, \ldots, q_Q]$ pak probíhá podle algoritmu

$$q_j \leftarrow q_j + \Delta q_j$$

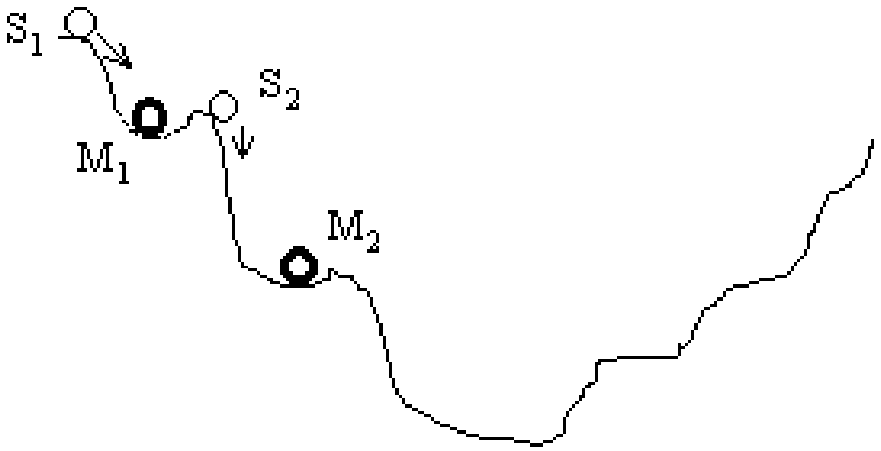
kde

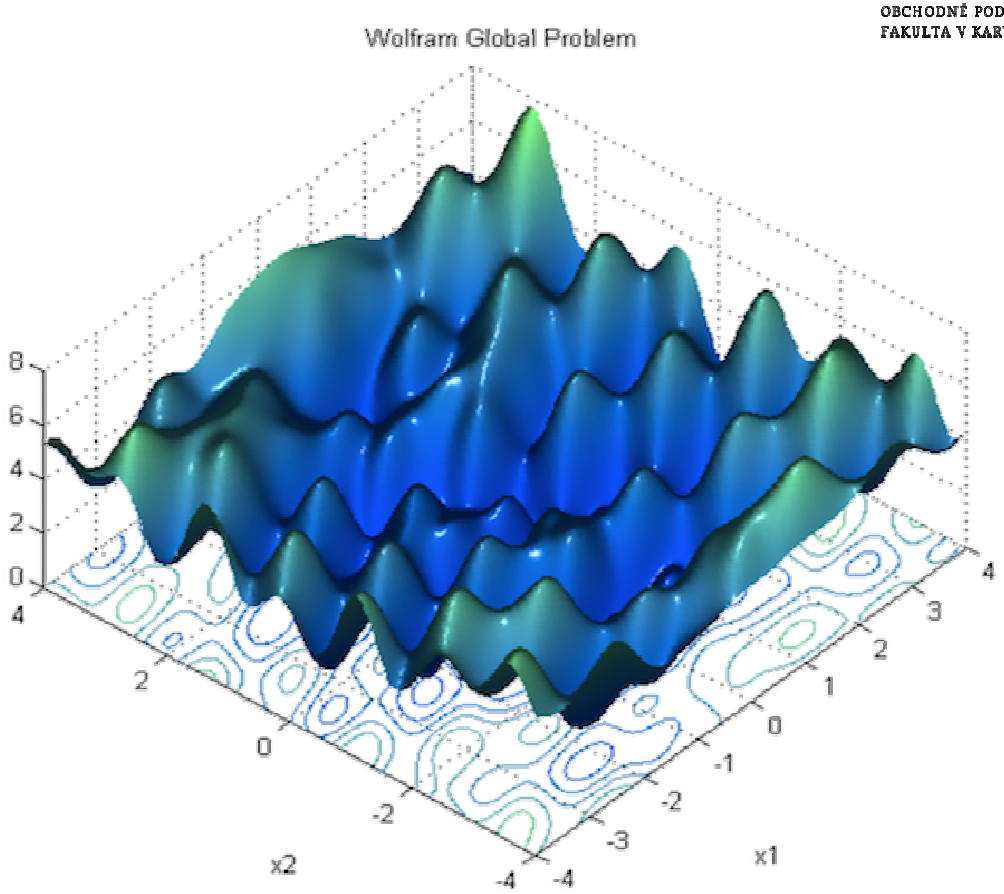
$$\Delta q_j = -\eta \frac{\partial \mathrm{Err}}{\partial q_j}$$

a η je parametr vyjadřující „velikost kroku" kterým se přibližujeme k minimu funkce *Err*.

Problém uváznutí v lokálním minimu

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ





Děkuji za pozornost

Některé snímky převzaty od:

prof. Ing. Petr Berka, CSc. berka@vse.cz