



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

Rozhodovací stromy

Jan Górecki



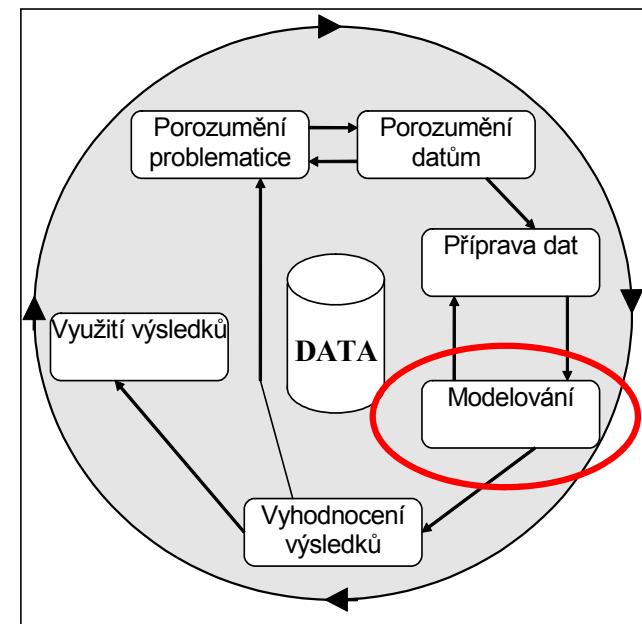
**SLEZSKÁ
UNIVERZITA**
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

Obsah přednášky



**SLEZSKÁ
UNIVERZITA**
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

- Co jsou Rozhodovací stromy
- Obecný algoritmus a omezení
- Příklad na bankovních datech
- Gini index
- Převod stromu na pravidla
- Prořezávání
- Práce s numerickými atributy

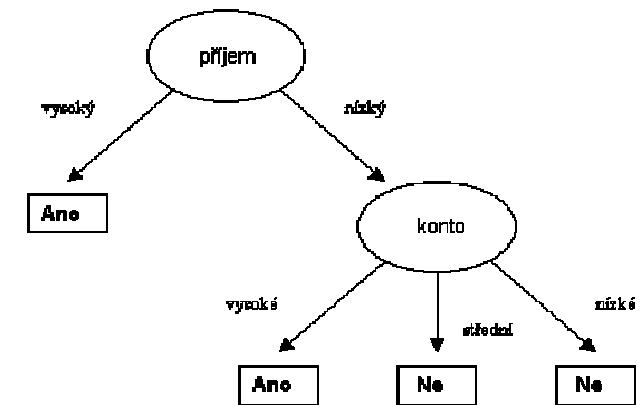


Rozhodovací stromy



SLEZSKÁ
UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

- Úloha klasifikace objektů do tříd.
- Top down induction of decision trees (TDIDT) - metoda **divide and conquer** (rozděl a panuj)
- Metoda specializace v prostoru hypotéz – stromů (postup shora dolů, počínaje prázdným stromem).
- Cílem je nalézt nějaký strom konsistentní s trénovacími daty.
- Dává se přednost menším stromům (Occamova břitva).



Obecný algoritmus pro tvorbu rozhodovacích stromů



**SLEZSKÁ
UNIVERZITA**
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

TDIDT algoritmus

1. vezmi jeden atribut jako kořen dílčího stromu,
2. rozděl data na podmnožiny podle hodnot tohoto atributu,
3. nepatří-li všechna data v podmnožině do téže třídy, pro tuto podmnožinu opakuj postup od bodu 1.

Omezení algoritmu:

1. Jen kategorické atributy
2. Data bez šumu (pro stejné kombinace hodnot vstupních atributů je stejná třída)

klient	příjem	konto	pohlaví	nezaměstnaný	úvěr
k1	vysoký	vysoké	žena	ne	ano
k2	vysoký	vysoké	muž	ne	ano
k3	nízký	nízké	muž	ne	ne
k4	nízký	vysoké	žena	ano	ano
k5	nízký	vysoké	muž	ano	ano
k6	nízký	nízké	žena	ano	ne
k7	vysoký	nízké	muž	ne	ano
k8	vysoký	nízké	žena	ano	ano
k9	nízký	střední	muž	ano	ne
k10	vysoký	střední	žena	ne	ano
k11	nízký	střední	žena	ano	ne
k12	nízký	střední	muž	ne	ano

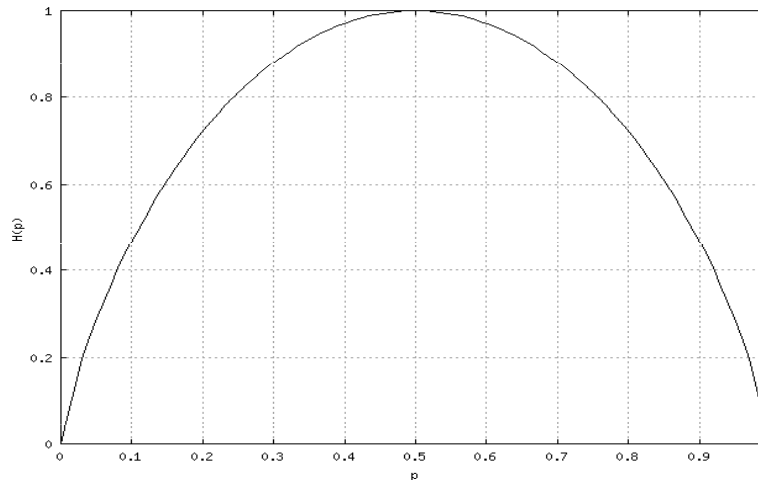
Volba atributu (krok 1 algoritmu)



Entropie:

$$H(p) = - \sum_{t=1}^T p_t \log_2 p_t$$

- $p = (p_1, \dots, p_T)$ a p_t je pravděpodobnost výskytu třídy t (v našem případě relativní četnost třídy t počítaná na určité množině příkladů)
- T je počet tříd



Entropie

Volba atributu (krok 1 algoritmu)



Hodnoty
zvažovaného
vstupního
atributu

Třídy

	Y ₁	Y ₂	Y _S	Σ
X ₁	a ₁₁	a ₁₂	a _{1S}	r ₁
X ₂	a ₂₁	a ₂₂	a _{2S}	r ₂
⋮	⋮	⋮		⋮	⋮
⋮	⋮	⋮		⋮	⋮
X _R	a _{R1}	a _{R2}	a _{RS}	r _R
Σ	s ₁	s ₂		s _S	n

$$H(A) = \sum_{i=1}^R \frac{r_i}{n} \left(- \sum_{j=1}^S \frac{a_{ij}}{r_i} \log_2 \frac{a_{ij}}{r_i} \right)$$

Hledáme atribut
s **minimální** hodnotou
kritéria (střední entropie
H)!

Příklad



**SLEZSKÁ
UNIVERZITA**
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

klient	příjem	konto	pohlaví	nezaměstnaný	úvěr
k1	vysoký	vysoké	žena	ne	ano
k2	vysoký	vysoké	muž	ne	ano
k3	nízký	nízké	muž	ne	ne
k4	nízký	vysoké	žena	ano	ano
k5	nízký	vysoké	muž	ano	ano
k6	nízký	nízké	žena	ano	ne
k7	vysoký	nízké	muž	ne	ano
k8	vysoký	nízké	žena	ano	ano
k9	nízký	střední	muž	ano	ne
k10	vysoký	střední	žena	ne	ano
k11	nízký	střední	žena	ano	ne
k12	nízký	střední	muž	ne	ano

$$H(\text{příjem}) = \frac{5}{12}H(\text{příjem}(\text{vysoký})) + \frac{7}{12}H(\text{příjem}(\text{nízký}))$$

Příklad

$$H(\text{příjem}) = 5/12 * H(\text{příjem}(\text{vysoký})) + 7/12 * H(\text{příjem}(\text{nízký}))$$

- $H(\text{příjem}(\text{vysoký})) = -5/5 * \log_2 5/5 - 0/5 * \log_2 0/5 = 0 + 0 = 0$

- $H(\text{příjem}(\text{nízký})) = -3/7 * \log_2 3/7 - 4/7 * \log_2 4/7 = 0.9852$

$$H(\text{příjem}) = 0.5747$$

Příklad



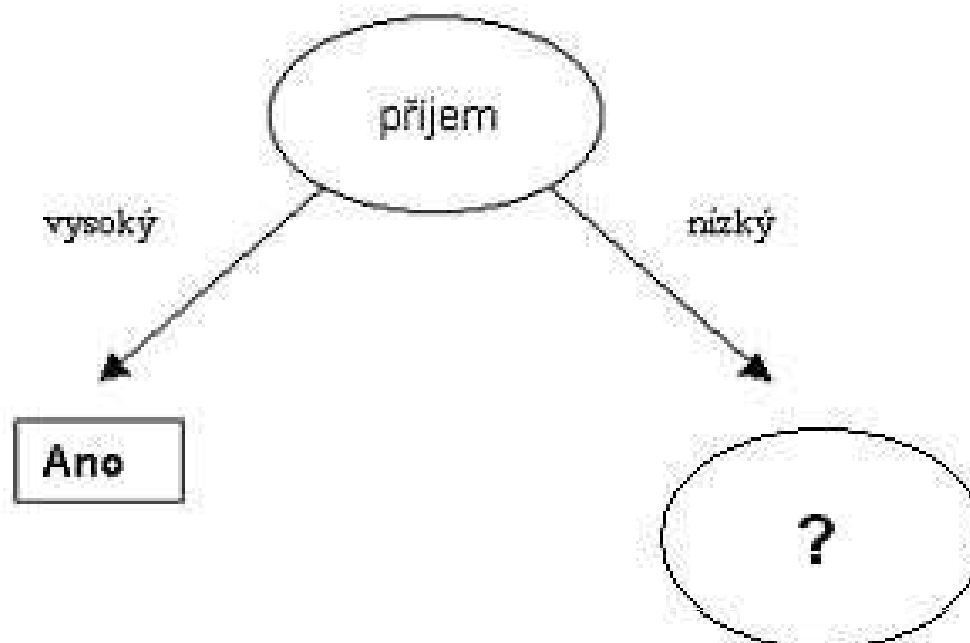
SLEZSKÁ
UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

$$H(\text{konto}) = 4/12 * H(\text{konto}(\text{vysoké})) + 4/12 * H(\text{konto}(\text{střední})) + 4/12 * H(\text{konto}(\text{nízké})) = 1/3 * 0 + 1/3 * 1 + 1/3 * 1 = \mathbf{0.6667}$$

$$H(\text{pohlaví}) = 6/12 * H(\text{pohlaví}(\text{muž})) + 6/12 * H(\text{pohlaví}(\text{žena})) = 1/2 * 0.9183 + 1/2 * 0.9183 = \mathbf{0.9183}$$

$$H(\text{nezaměstnaný}) = 6/12 * H(\text{nezaměstnaný}(\text{ano})) + 6/12 * H(\text{nezaměstnaný}(\text{ne})) = 1/2 * 1 + 1/2 * 0.6500 = \mathbf{0.8250}$$

Příklad



Příklad



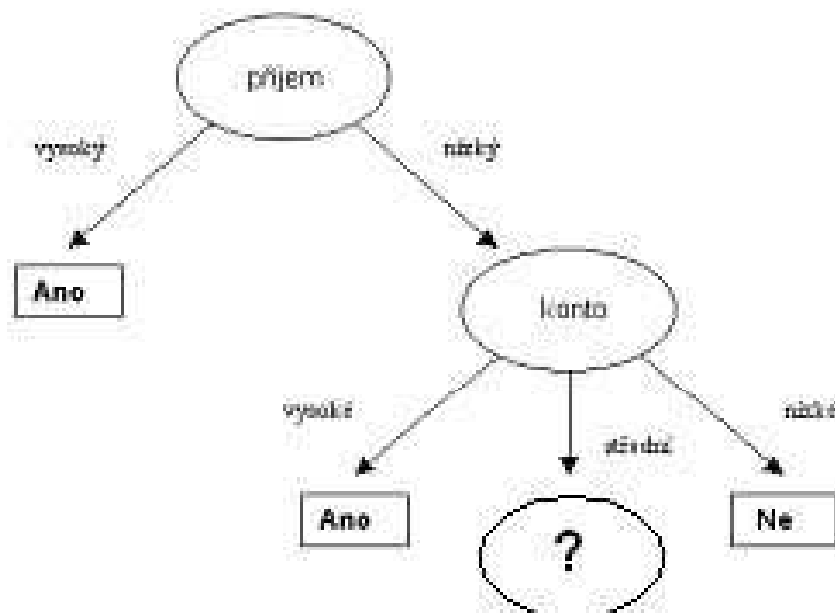
SLEZSKÁ
UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

$$H(\text{konto}) = 2/7 * H(\text{konto}(\text{vysoké})) + 3/7 * H(\text{konto}(\text{střední})) + 2/7 * H(\text{konto}(\text{nízké})) = 2/7 * 0 + 3/7 * 0.9183 + 2/7 * 0 = 0.3935$$

$$H(\text{pohlaví}) = 4/7 * H(\text{pohlaví}(\text{muž})) + 3/7 * H(\text{pohlaví}(\text{žena})) = 4/7 * 1 + 3/7 * 0.9183 = 0.9650$$

$$H(\text{nezaměstnaný}) = 5/7 * H(\text{nezaměstnaný}(\text{ano})) + 2/7 * H(\text{nezaměstnaný}(\text{ne})) = 5/7 * 0.9709 + 2/7 * 1 = 0.9792$$

Příklad



Příklad

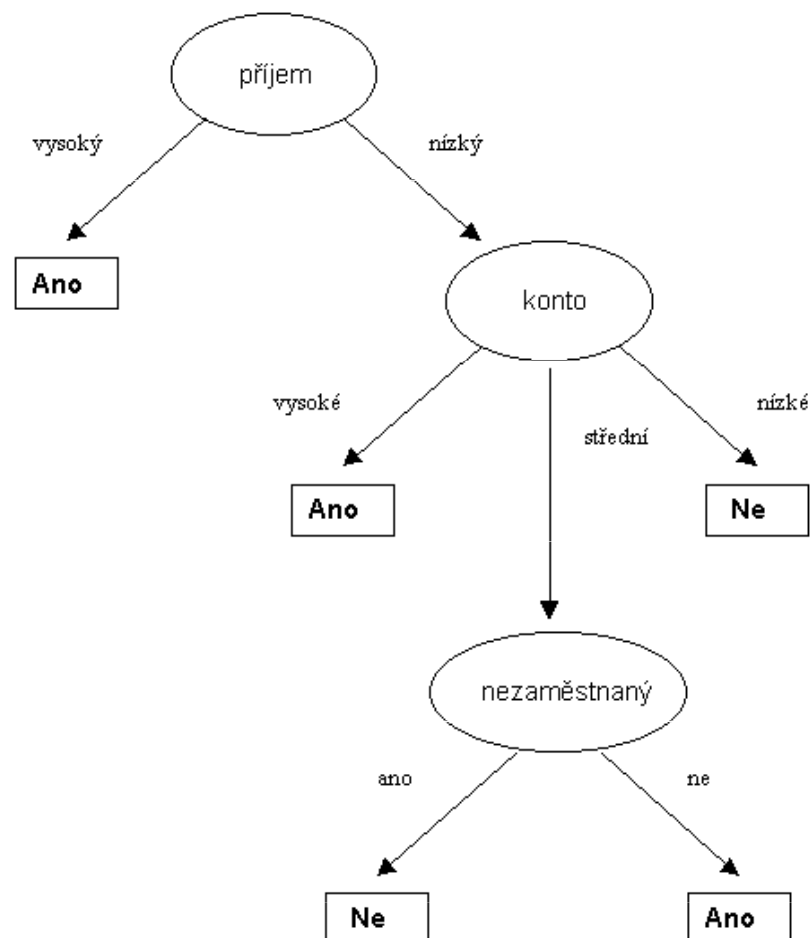


$$H(\text{pohlaví}) = 2/3 * H(\text{pohlaví}(\text{muž})) + 1/3 * H(\text{pohlaví}(\text{žena})) = 2/3 * 1 + 1/3 * 0 = 0.6667$$

$$H(\text{nezaměstnaný}) = 2/3 * H(\text{nezaměstnaný}(\text{ano})) + 1/3 * H(\text{nezaměstnaný}(\text{ne})) = 2/3 * 0 + 1/3 * 0 = 0$$

Pozn: V případě kategoriálních atributů se každý atribut může pro větvení stromu vybrat v jedné větvi **nejvýše** jednou.

Příklad



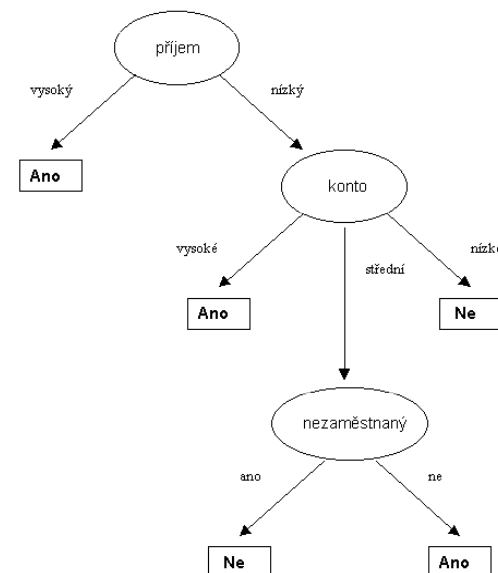
Shrnutí

Tedy, tvorba rozhodovacích stromů je založena na prohledávání prostoru stromů:

- Shora dolů
- Heuristické

Dále:

- Jednoduché použití
- Má schopnost generalizovat,
např. pro [příjem(nízký), konto(nízké), pohlaví(muž), nezaměstnaný(ano)]
dává úvěr = ne

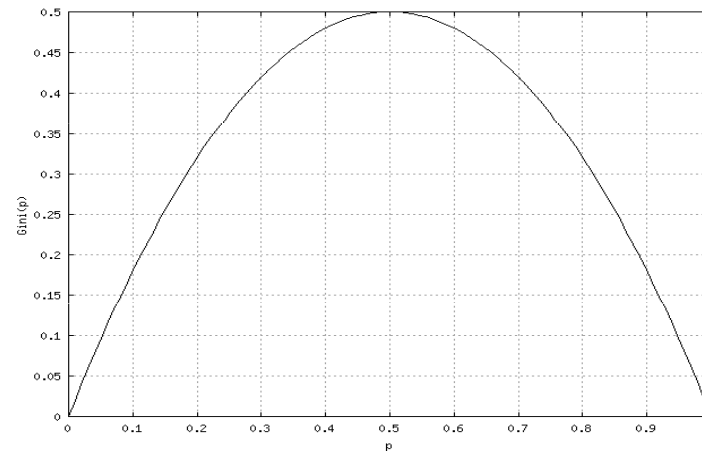


Gini index



SLEZSKÁ
UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

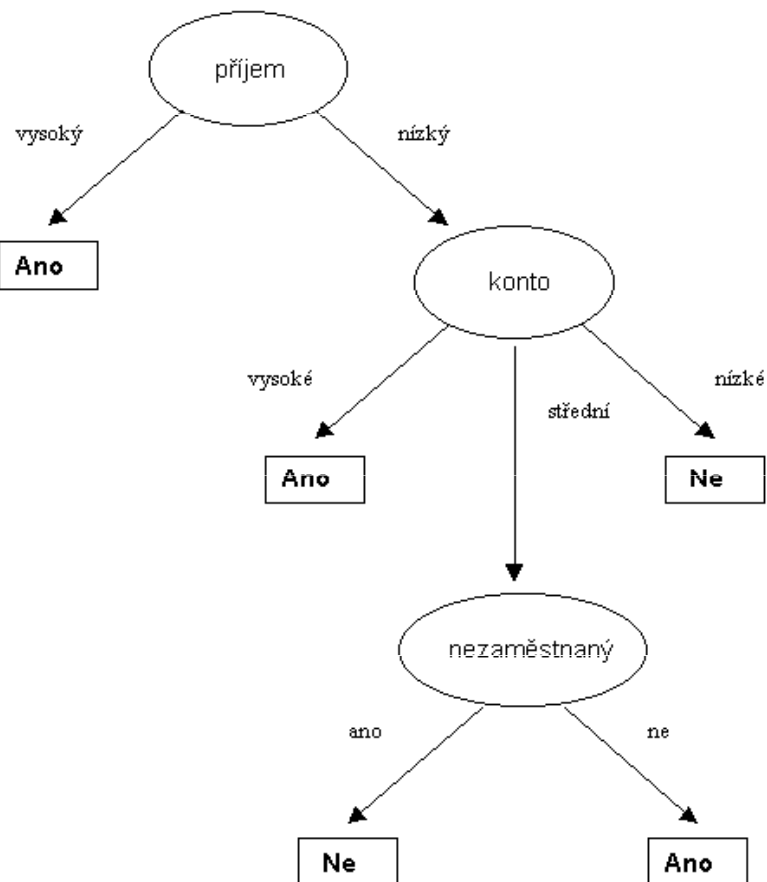
$$\text{Gini} = \sum_{t=1}^T p_t (1 - p_t) = 1 - \sum_{t=1}^T p_t^2$$



$$\text{Gini}(A) = \sum_{i=1}^R \frac{r_i}{n} \left(1 - \sum_{j=1}^S \left(\frac{a_{ij}}{r_i} \right)^2 \right)$$

Hledáme atribut s minimální hodnotou kritéria (střední Gini index)!

Převod stromů na pravidla



1. If příjem(vysoký) then úvěr(ano)
2. If příjem(nízký) \wedge konto(vysoké) then úvěr(ano)
3. If příjem(nízký) \wedge konto(střední) \wedge nezaměstnaný(ano) then úvěr(ne)
4. If příjem(nízký) \wedge konto(střední) \wedge nezaměstnaný(ne) then úvěr(ano)
5. If příjem(nízký) \wedge konto(nízké) then úvěr(ne)

Prořezávání stromů



SLEZSKÁ
UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

Důvody:

- Bezchybná klasifikace trénovacích dat nezaručuje kvalitní klasifikaci dat testovacích (overfitting)
- Úplný strom může být příliš veliký

Redukce stromu, aby v listovém uzlu „převažovaly“ příklady jedné třídy.

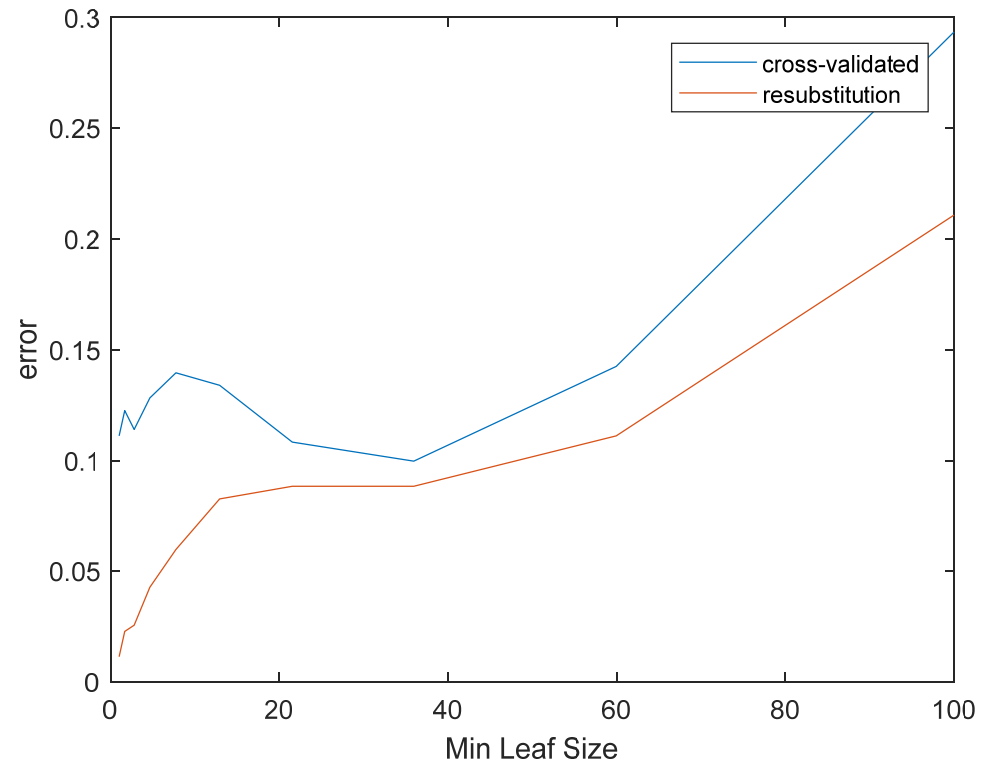
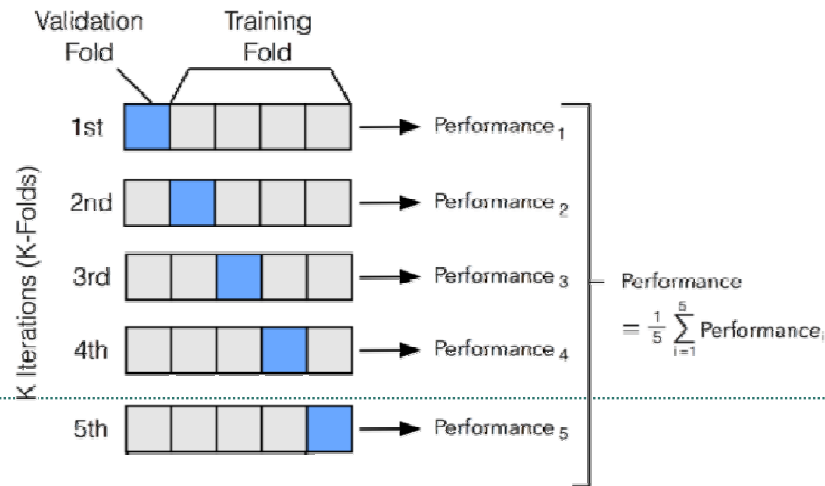
- **pre-pruning** – modifikuje se zastavovací kritérium (krok 3 algoritmu) = větvit se nebude pokud počet příkladů v uzlu klesne pod danou hodnotu nebo pokud relativní počet příkladů jedné třídy překročí danou hodnotu
- **post-pruning** – vytvoří se úplný strom, který se následně redukuje – ukazuje se jako úspěšnější než pre-pruning, protože předem lze těžko poznat, jak nastavit kritéria zastavení

Lze kombinovat pre-pruning s post-pruningem.

Pre-pruning



1. Pro každou zvažovanou hodnotu hodnoty Min Leaf Size generuj:
 - a. Jeden strom a spočti chybu klasifikace na celých datech (**resubstitution error**).
 - b. Pět stromů tak, že data se rozdělí na pět částí a vždy čtyři z nich se použije pro trénování stromu a pátá část se použije změření chyby klasifikace (**cross-validated error**).



Post-pruning



- Na rozdíl od pre-pruningu, není třeba vytvořit celý strom pro každou volbu parametrů – celý strom se vytvoří pouze jednou a ten se pak ořezává

Dvě strategie:

Ořezávej větve stromu, které nejvíce snižují chybu na testovacích datech (s využitím křížové validace) dokud:

- a) je možno chybu ořezáváním snížit - strategie *Minimální chyba (Minimum error)*
 - b) je chyba menší než minimální chyba (z předchozí strategie) + standardní odchylka minimální chyby - strategie *Nejmenší strom (Smallest tree)* – tato strategie je schopna produkovat menší stromy než předchozí strategie za cenu mírně vyšší chyby
-



Algoritmus pracuje s kategoriálními atributy, numerické je třeba diskretizovat:

1. **off-line** v rámci přípravy a předzpracování dat
 2. **on-line** v rámci běhu modifikovaného algoritmu
 - binarizace na základě entropie
-

Algoritmus diskretizace



SLEZSKÁ
UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

1. Seřad' vzestupně hodnoty diskretizovaného atributu A ,
2. Pro každou možnou hodnotu dělicího bodu θ spočítej entropii $H(A_\theta)$
3. Vyber dělicí bod θ , který dá nejmenší hodnotu $H(A_\theta)$

$$H(A_\theta) = \frac{n(A(<\theta))}{n} H(A(<\theta)) + \frac{n(A(>\theta))}{n} H(A(>\theta))$$

První člen součtu se týká příkladů, které mají hodnotu atributu menší než θ ($H(A(<\theta))$ je entropie na těchto příkladech, $n(A(<\theta))/n$ je relativní četnost těchto příkladů), druhý člen součtu se analogicky týká příkladů, které mají hodnotu atributu větší než θ .

Příklad



**SLEZSKÁ
UNIVERZITA**
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

Klient	příjem	Konto	úvěr
K101	3000	15000	ne
K102	10000	15000	ne
K103	17000	15000	ano
K104	5000	30000	ne
K105	15000	30000	ano
K106	20000	50000	ano
K107	2000	60000	ne
K108	5000	90000	ano
K109	10000	90000	ano
K110	20000	90000	ano
K111	10000	100000	ano
K112	17000	100000	ano

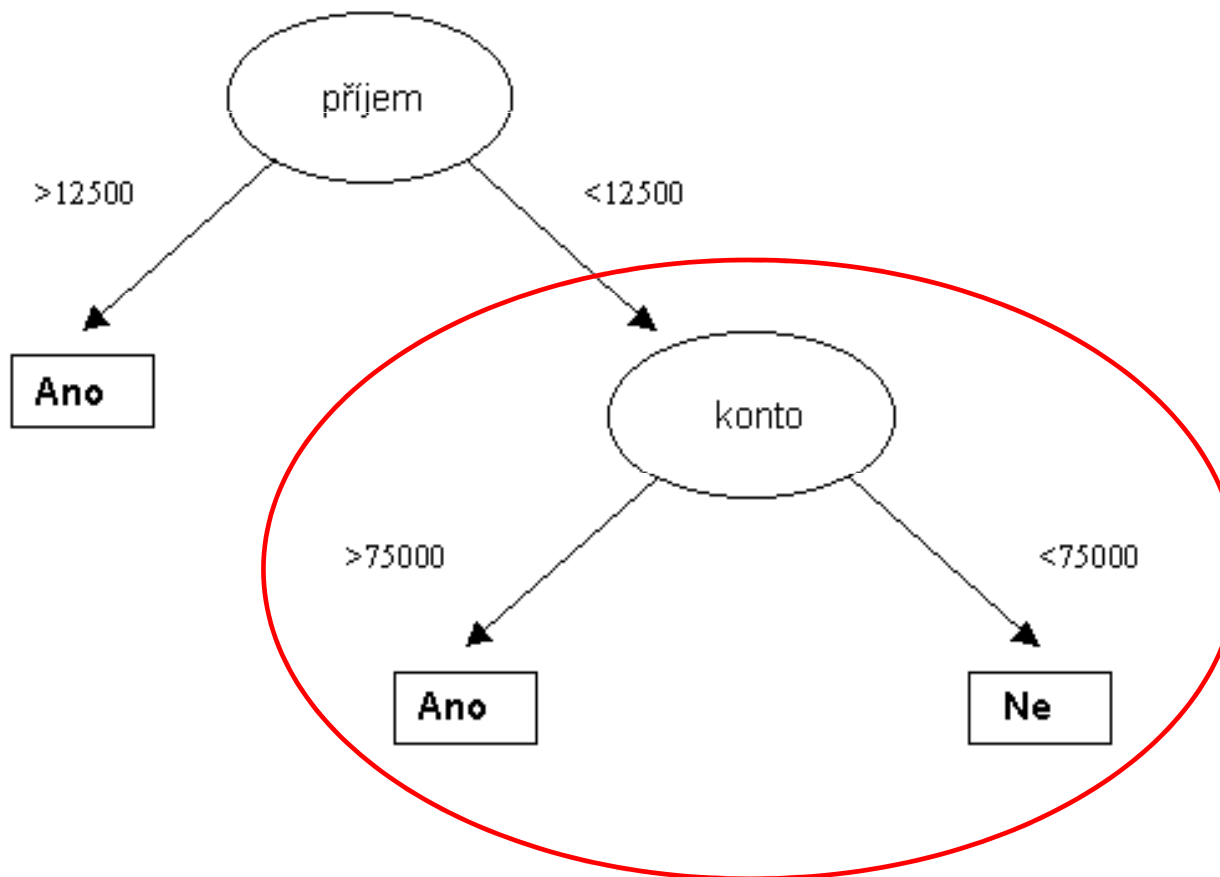
$$H(\text{konto}_{22500}) = 3/12 * H(\text{konto}(<22500)) + 9/12 * H(\text{konto}(>22500)) = \\ 1/4 * 0.9183 + 3/4 * 0.5640 = 0.6526$$

$$H(\text{konto}_{40000}) = 5/12 * H(\text{konto}(<40000)) + 7/12 * H(\text{konto}(>40000)) = \\ 5/12 * 0.9706 + 7/12 * 0.5917 = 0.7497$$

$$H(\text{konto}_{55000}) = 6/12 * H(\text{konto}(<55000)) + 6/12 * H(\text{konto}(>55000)) = \\ 1/2 * 1 + 1/2 * 0.6500 = 0.8250$$

$$H(\text{konto}_{75000}) = 7/12 * H(\text{konto}(<75000)) + 5/12 * H(\text{konto}(>75000)) = \\ 7/12 * 0.9852 + 5/12 * 0 = \mathbf{0.5747}$$

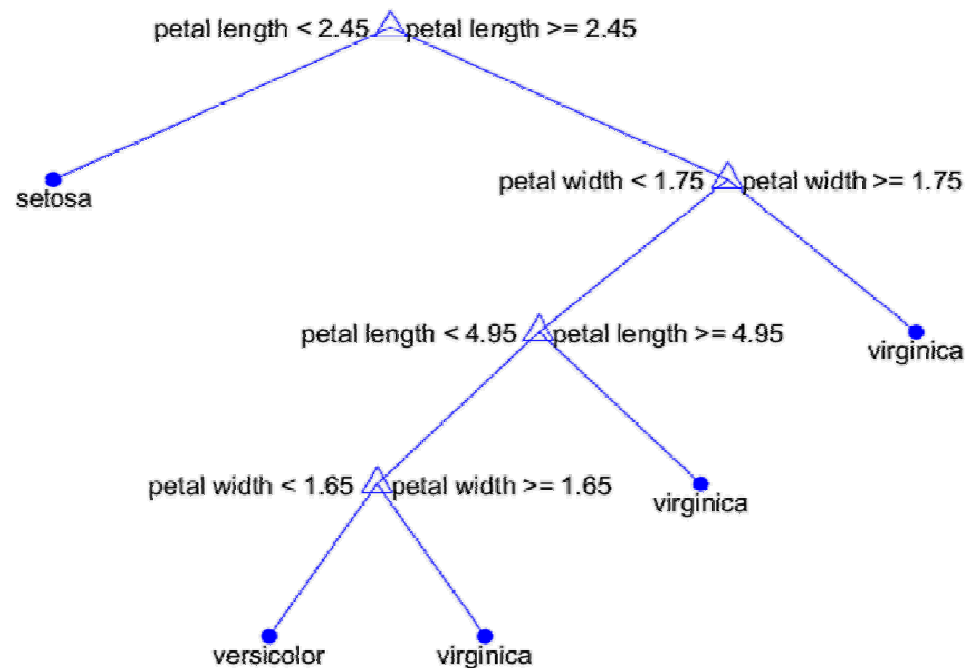
Příklad



Kategoriální vs numerické atributy



- na rozdíl od kategoriálních atributů se mohou v jedné větvi numerické atributy opakovat



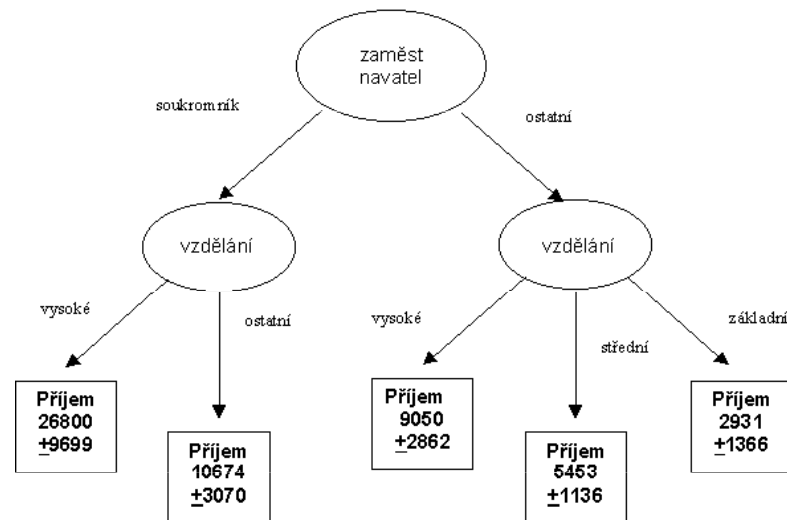
Regresní stromy



- Úloha odhadu hodnoty nějakého numerického atributu.

Volba atributu (krok 1):

- kritérium **redukce směrodatné odchylky**



Použití rozhodovacích stromů



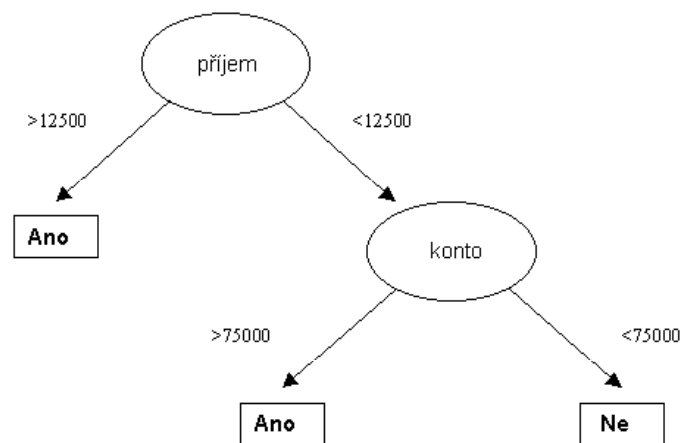
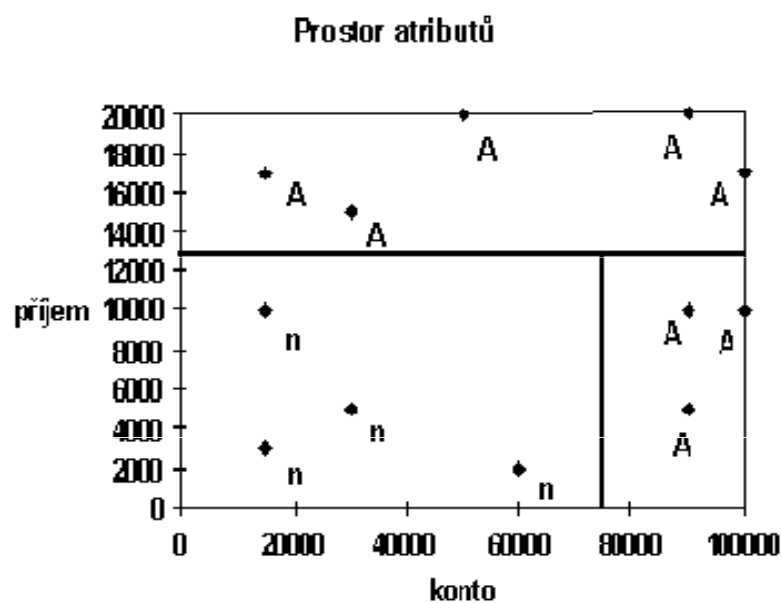
**SLEZSKÁ
UNIVERZITA**
OBCHODNĚ PODNIKATELSKÁ
FAKULTA V KARVINĚ

- příklady jsou reprezentovány hodnotami atributů,
 - úkolem je klasifikovat příklady do konečného (malého) počtu tříd,
 - trénovací data mohou být zatížena šumem,
 - trénovací data mohou obsahovat chybějící hodnoty
-

Vyjadřovací síla rozhodovacích stromů



- Rozhodovací stromy dělí prostor atributů na (mnoharozměrné) hranoly rovnoběžné s osami souřadné soustavy:



Děkuji za pozornost

Některé snímky převzaty od:
prof. Ing. Petr Berka, CSc. berka@vse.cz