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

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

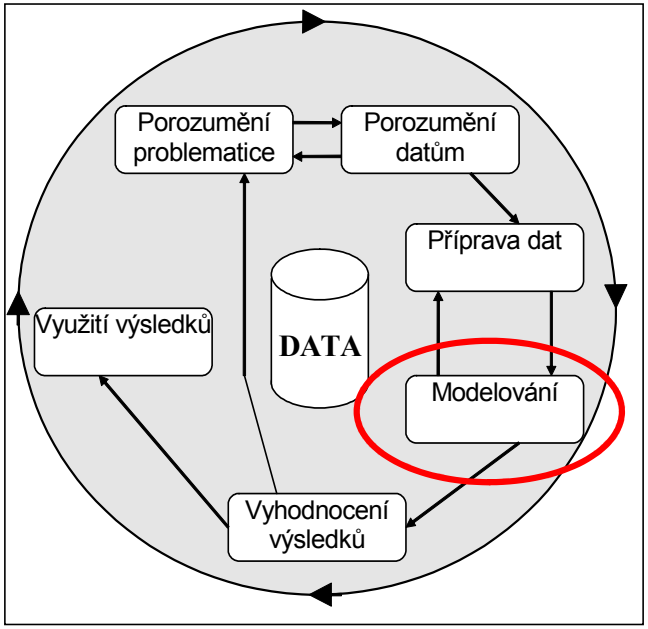
Rozhodovací stromy

Jan Górecki

Obsah přednášky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

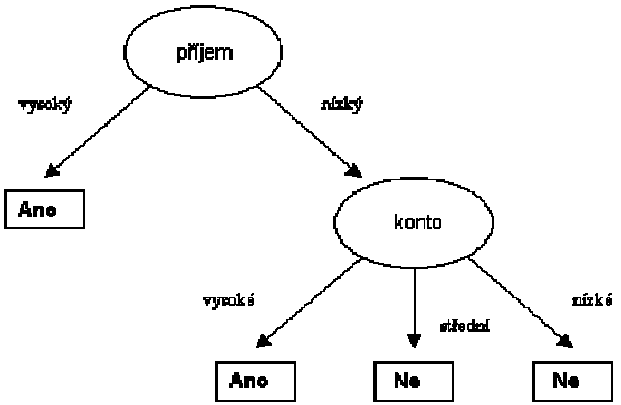
- Co jsou Rozhodovací stromy
- Obecný algoritmus a omezení
- Příklad na bankovních datech
- Gini index
- Převod stromu na pravidla
- Prořezávání
- Práce s numerickými atributy



Rozhodovací stromy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- 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 KARVINÉ

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.

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

Omezení algoritmu:

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

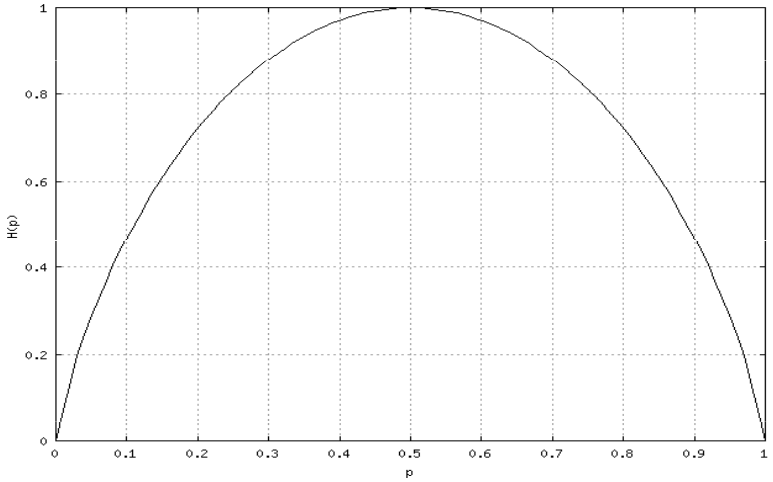
Volba atributu (krok 1 algoritmu)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Entropie:

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

- $p = (p_1, \ldots, p_T)$ *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)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Třídy

Hodnoty zvažovaného vstupního atributu

	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_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 KARVINÉ

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

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

- $H(\text{příjem(vysoký)}) = -5/5*\log_2 5/5 - 0/5*\log_2 *0/5 = 0 + 0 = \mathbf{0}$
- $H(\text{příjem(nízký)}) = - 3/7*\log_2 3/7 - 4/7*\log_2 4/7 = \mathbf{0.9852}$

H(příjem) = 0.5747

Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

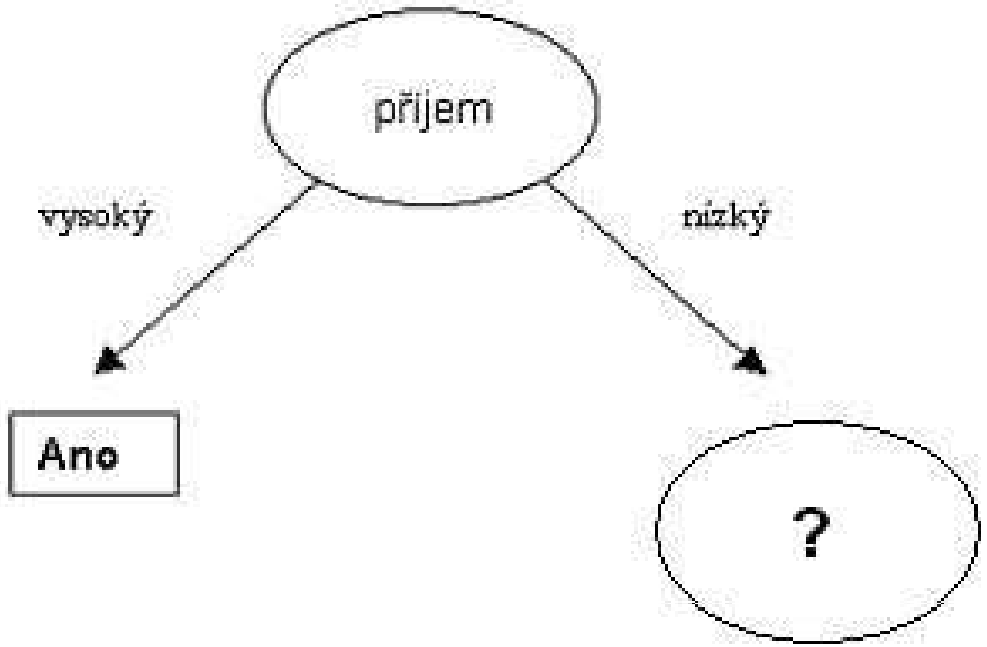
H(konto) = 4/12*H(konto(vysoké)) + 4/12*H(konto(střední)) + 4/12*H(konto(nízké)) = 1/3 * 0 + 1/3 * 1 + 1/3 * 1 = **0.6667**

H(pohlaví) = 6/12*H(pohlaví(muž)) + 6/12*H(pohlaví(žena)) = 1/2 * 0.9183 + 1/2 * 0.9183 = **0.9183**

H(nezaměstnaný) = 6/12*H(nezaměstnaný(ano)) + 6/12*H(nezaměstnaný(ne)) = 1/2 * 1 + 1/2 * 0.6500 = **0.8250**

Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

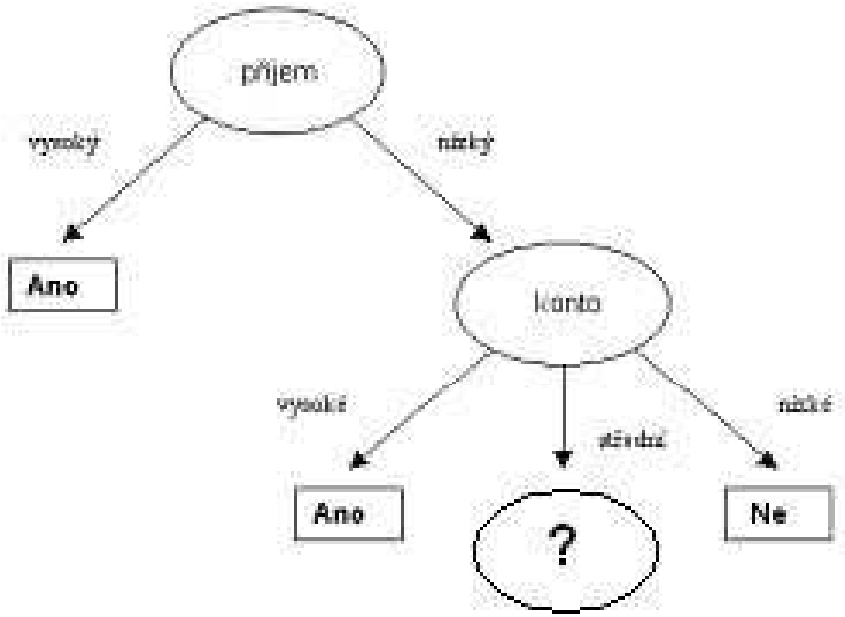
$$H(konto) = 2/7*H(konto(vysoké)) + 3/7*H(konto(střední)) + 2/7*H(konto(nízké)) = 2/7 * 0 + 3/7 * 0.9183 + 2/7 * 0 = 0.3935$$

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

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

Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

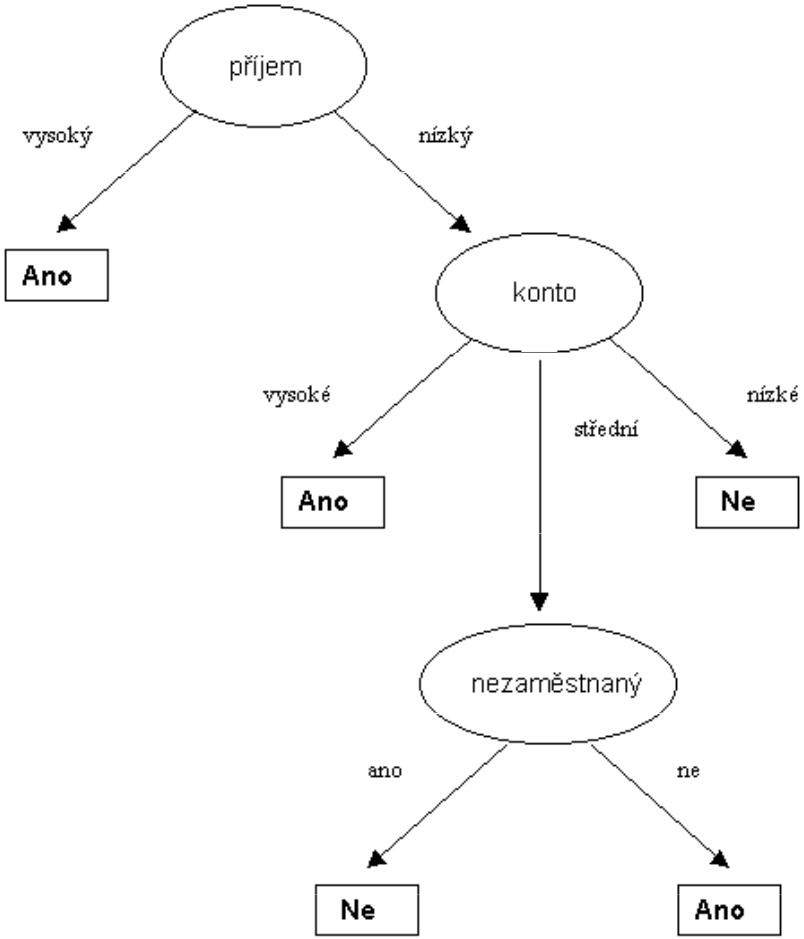
$$H(\text{pohlaví}) = 2/3 * H(\text{pohlaví(muž)}) + 1/3 * H(\text{pohlaví(žena)}) = 2/3 * 1 + 1/3 * 0 = 0.6667$$

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

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

Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

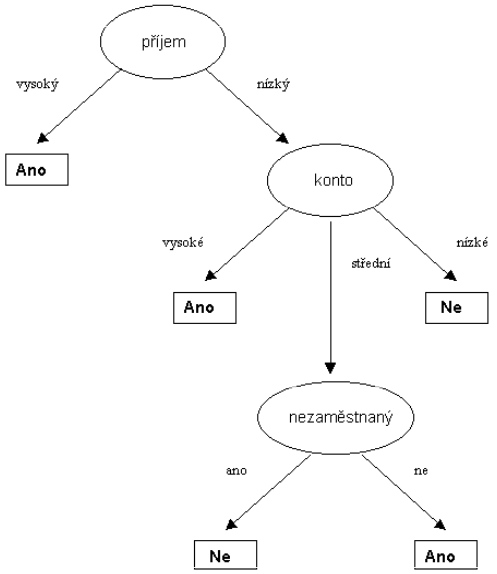


Shrnutí

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

- Shora dolů
- Heuristické



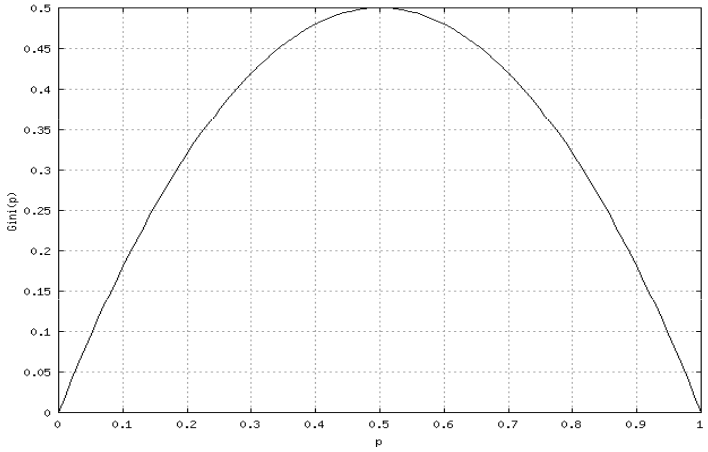
Dále:

- 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 KARVINÉ

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

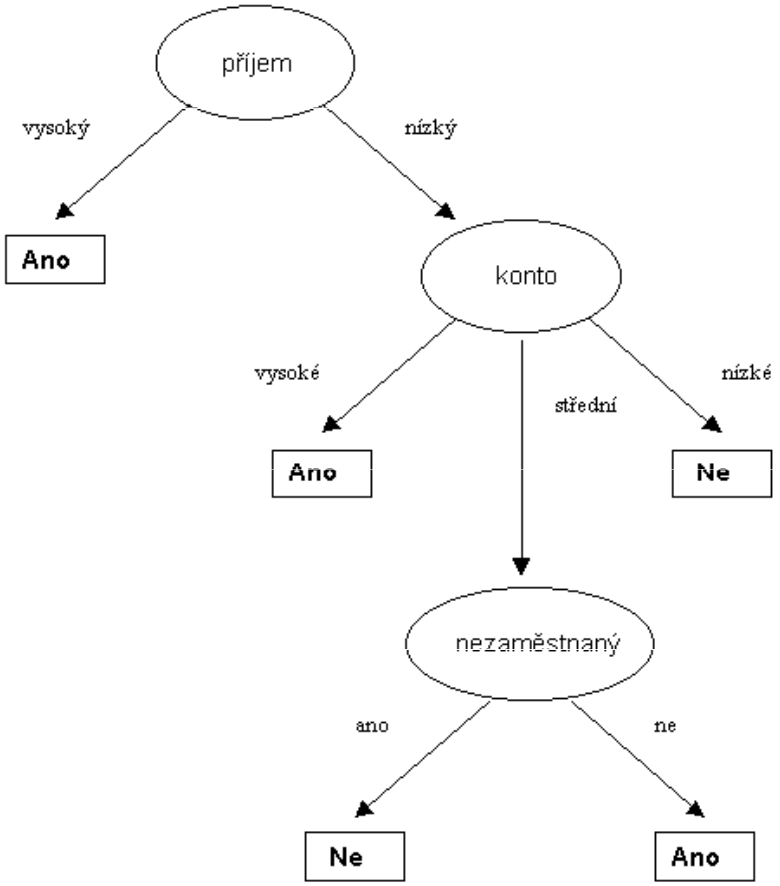


$$Gini(A) = \sum_{i=1}^{R} \frac{r_i}{n} \left(1 - \sum_{j=1}^{S} \left(\frac{a_{ij}}{r_i} \right)^2 \right)$$

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

Převod stromů na pravidla

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



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

Prořezávání stromů

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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.

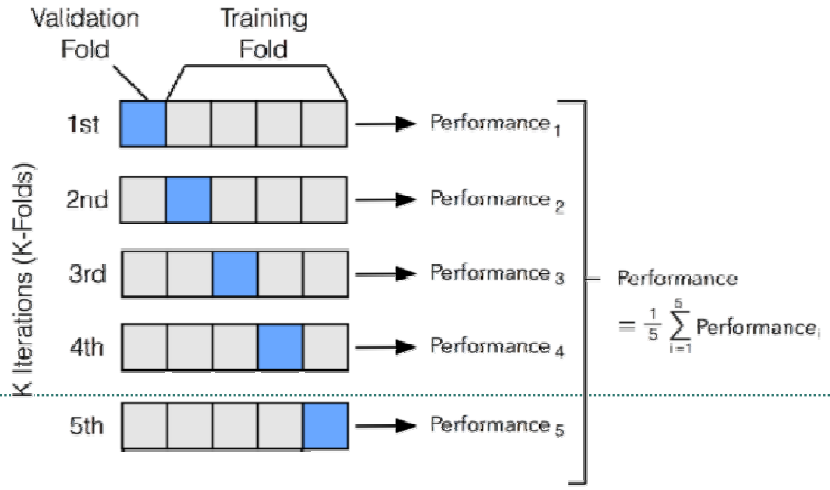
- **pre-pruning** – modifikuje se zastavovací kritérium (krok 3 algoritmu) = větvit se nebude pokud počet příkladů v uzlu klesne pod danou hodnotu nebo pokud relativní počet příkladů jedné třídy překročí danou hodnotu
- **post-pruning** – vytvoří se úplný strom, který se následně redukuje – ukazuje se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

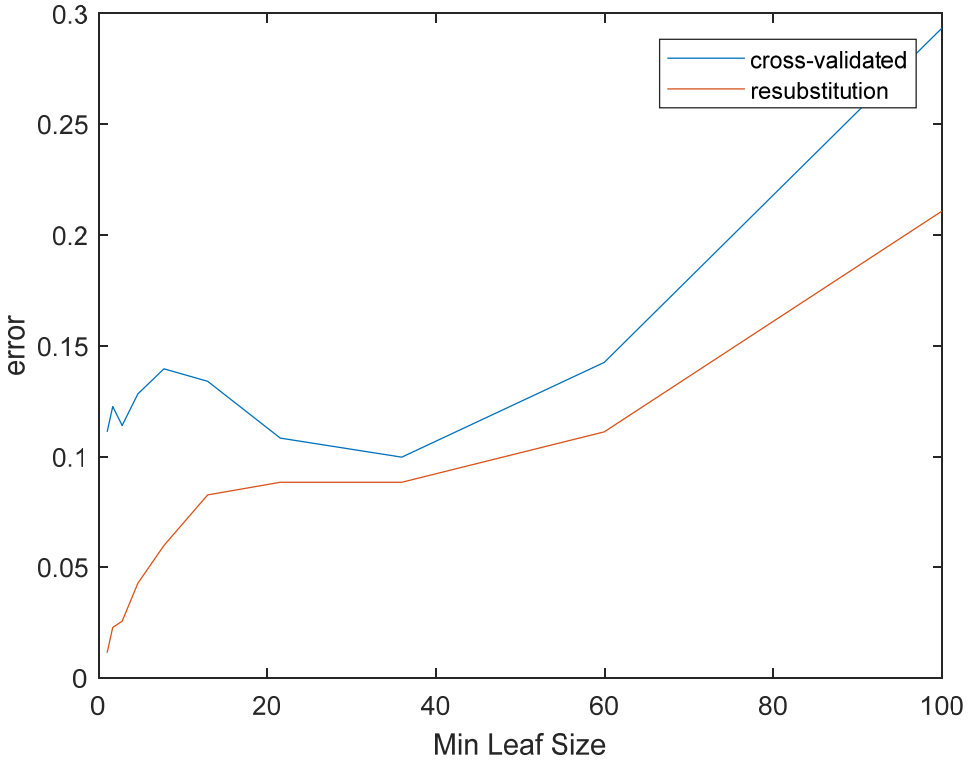
Lze kombinovat pre-pruning s post-pruningem.

Pre-pruning

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- 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

Práce s numerickými atributy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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 KARVINÉ

1. Seřaď vzestupně hodnoty diskretizovaného atributu A,
2. Pro každou možnou hodnotu dělícího bodu θ spočítej entropii $H(A_\theta)$
3. Vyber dělící 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 KARVINÉ

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

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

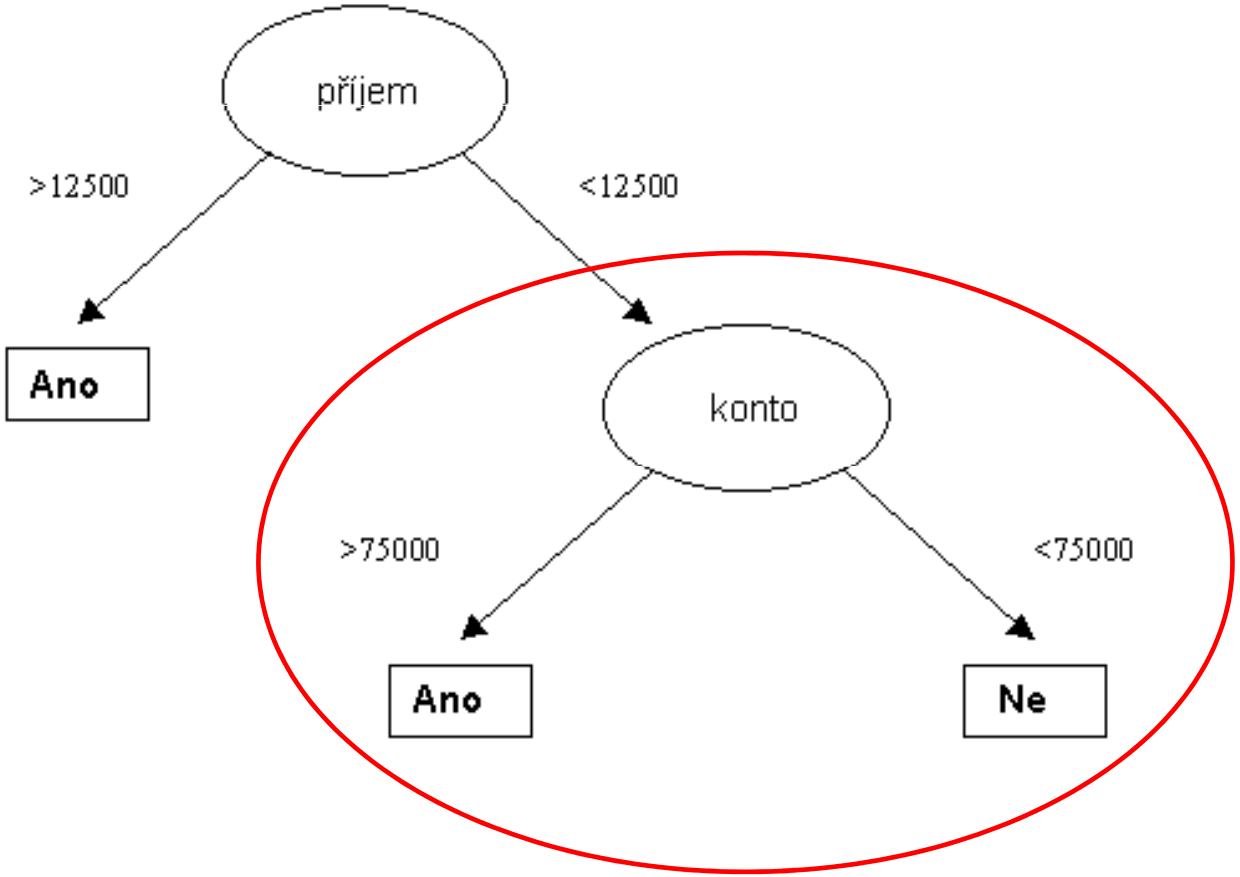
$$H(konto_{40000}) = 5/12 * H(konto(<40000)) + 7/12 * H(konto(>40000)) =$$
$$5/12 * 0.9706 + 7/12 * 0.5917 = 0.7497$$

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

$$\mathbf{H(konto_{75000})} = 7/12 * H(konto(<75000)) + 5/12 * H(konto(>75000)) =$$
$$7/12 * 0.9852 + 5/2 * 0 = \mathbf{0.5747}$$

Příklad

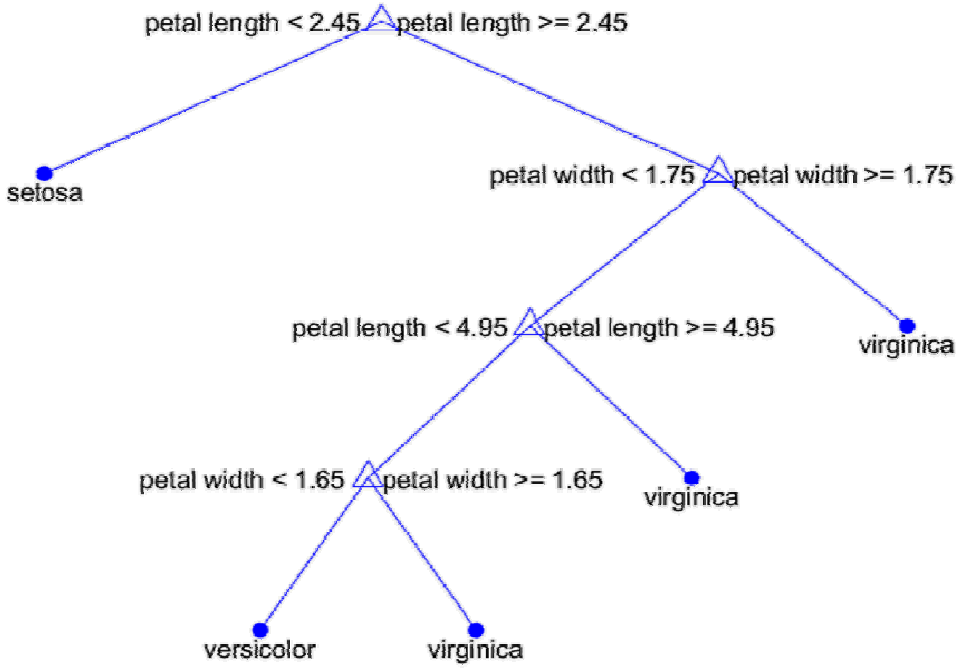
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



Kategoriální vs numerické atributy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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



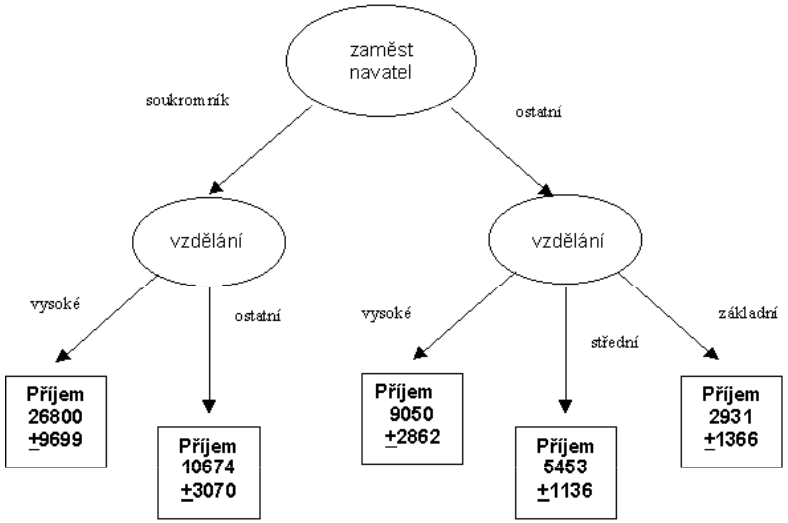
Regresní stromy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Úloha odhadu hodnoty nějakého numerického atributu.

Volba atributu (krok 1):

- kritérium **redukce směrodatné odchylky**



Použití rozhodovacích stromů

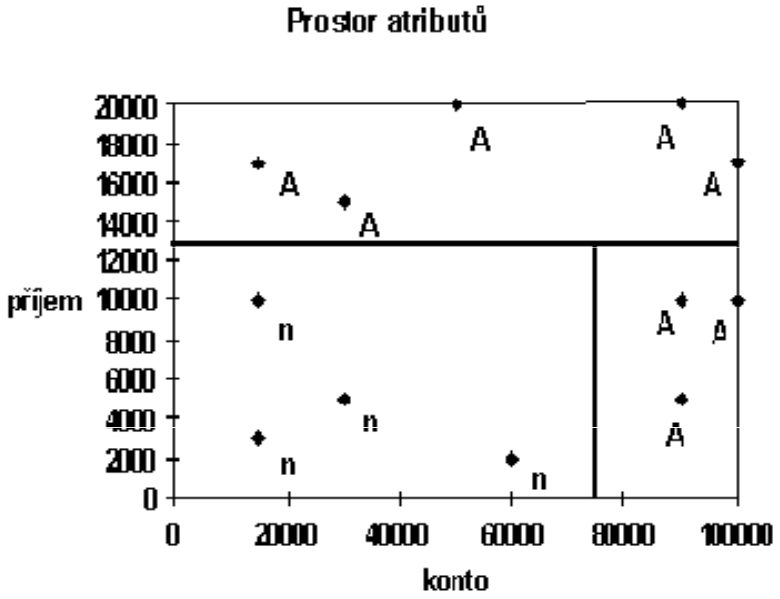
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

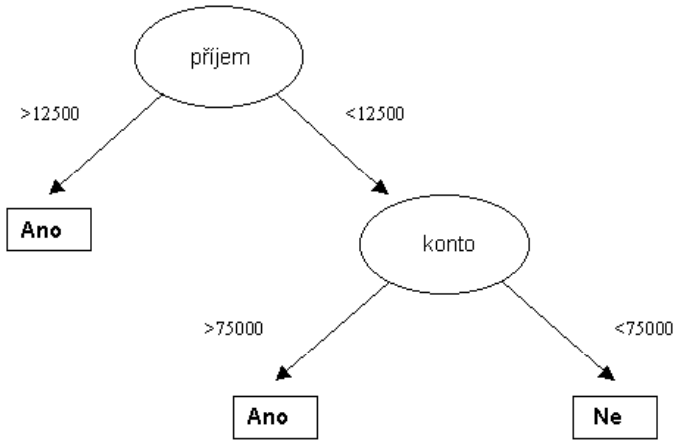
- příklady jsou reprezentovány hodnotami atributů,
- úkolem je klasifikovat příklady do konečného (malého) počtu tříd,
- trénovací data mohou být zatížena šumem,
- trénovací data mohou obsahovat chybějící hodnoty

Vyjadřovací síla rozhodovacích stromů

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- 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