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

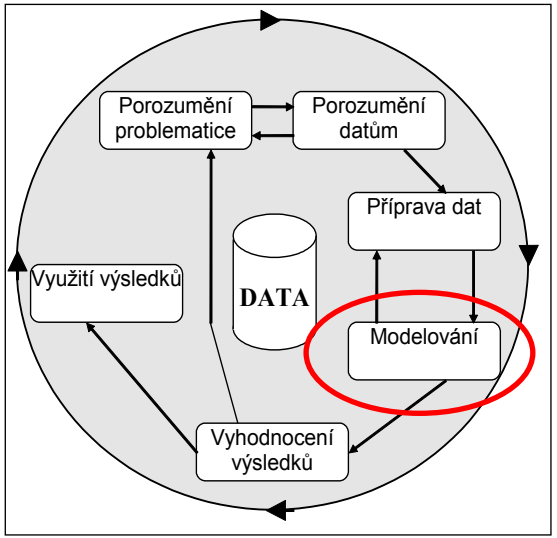
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Obsah přednášky

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Co jsou Rozhodovací stromy
- Obecný algoritmus a omezení
- Příklad na bankovních datech

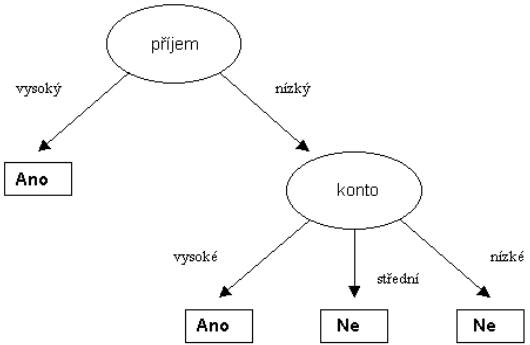
- Gini index
- Převod stromu na pravidla
- Prořezávání
- Práce s numerickými atributy



Rozhodovací stromy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Úloha klasifikace objektů do tříd (tedy učení s učitelem).
- 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.

Jak najít strom, který „pasuje“ na daná data?

Motivace:

Chyba (=1-správnost) pařezu vs chyba stromu s jedním uzlem

Příklad

Chyba(pařez) = 4/12

1. krok: Atribut jako kořen dílčího stromu

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Pro každý atribut (příjem, konto, pohlaví, nezaměstnaný) spočítáme chybu při jeho použití jako kořenového uzlu.

Atribut „příjem“:
- vysoký: 5x ano, 0x ne
- nízký: 3x ano, 4x ne

Atribut „konto“:
- vysoké: 2x ano, 0x ne
- nízké: 3x ano, 1x ne
- střední: 2x ano, 2x ne

Atribut „pohlaví“:
- žena: 4x ano, 1x ne
- muž: 4x ano, 3x ne

Atribut „nezaměstnaný“:
- ne: 6x ano, 2x ne
- ano: 2x ano, 2x ne

Pro výpočet chyby můžeme použít různé metriky, jako je Gini nebo entropie. Pro jednoduchost použijeme chybovou metriku, kde chyba je relativní počet chyb modelu na daných datech.

Příjem:
vysoký: 5x ano, 0x ne -> chyba = 0/5 = 0
nízký: 3x ano, 4x ne -> chyba = 3/7 = 0.43
Celková chyba (průměrná vážená): (5/12 * 0 + 7/12 * 3/7) = **0.25**

Konto:
vysoké: 2x ano, 0x ne -> chyba = 0/2 = 0
nízké: 3x ano, 1x ne -> chyba = 1/4 = 0.25
střední: 2x ano, 2x ne -> chyba = 2/4 = 0.5
Celková chyba (průměrná vážená): (2/12 * 0 + 4/12 * 0.25 + 4/12 * 0.5) = **0.25**

Pohlaví:
žena: 4x ano, 1x ne -> chyba = 1/5 = 0.2
muž: 4x ano, 3x ne -> chyba = 3/7 = 0.43
Celková chyba (průměrná vážená): (5/12 * 0.2 + 7/12 * 0.43) = 0.33

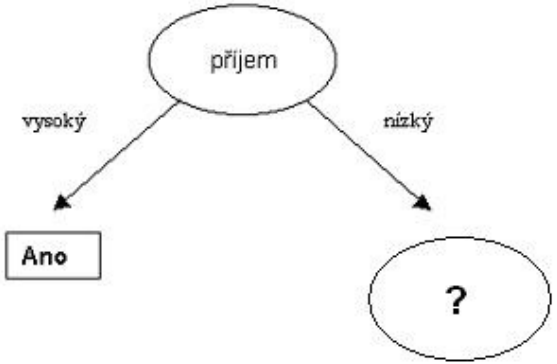
Nezaměstnaný:
ne: 6x ano, 2x ne -> chyba = 2/8 = 0.25
ano: 2x ano, 2x ne -> chyba = 2/4 = 0.5
Celková chyba (průměrná vážená): (8/12 * 0.25 + 4/12 * 0.5) = 0.33

2. krok: Rozdělíme data podle vybraného atributu a vytvoříme pro něj podstromy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Data rozdělená podle atributu „příjem“:

- vysoký příjem: 5x ano (listový uzel, žádná další rozdělení)
- nízký příjem: 3x ano, 4x ne (budeme dále rozdělovat)



3. krok: Opakování algoritmu pro podmnožinu s nízkým příjmem

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

příjem	konto	pohlavi	nezamestnany	uver
nizky	nizke	muz	ne	ne
nizky	nizke	zena	ano	ne
nizky	stredni	muz	ano	ne
nizky	stredni	zena	ano	ne
nizky	stredni	muz	ne	ano
nizky	vysoke	zena	ano	ano
nizky	vysoke	muz	ano	ano

Konto:

nízké: 1x ano, 2x ne -> chyba = 1/3 = 0.33
vysoké: 2x ano, 0x ne -> chyba = 0/2 = 0
střední: 1x ano, 2x ne -> chyba = 1/3 = 0.33
Celková chyba (průměrná vážená): (3/7 * 0.33 + 2/7 * 0 + 3/7 * 0.33) = **0.28**

Pohlaví:

žena: 2x ano, 2x ne -> chyba = 2/4 = 0.5
muž: 1x ano, 2x ne -> chyba = 1/3 = 0.33
Celková chyba (průměrná vážená): (4/7 * 0.5 + 3/7 * 0.33) = 0.43

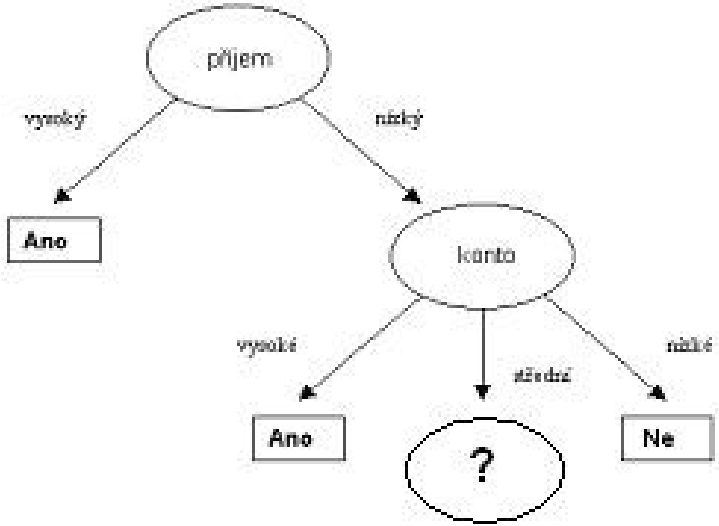
Nezaměstnaný:

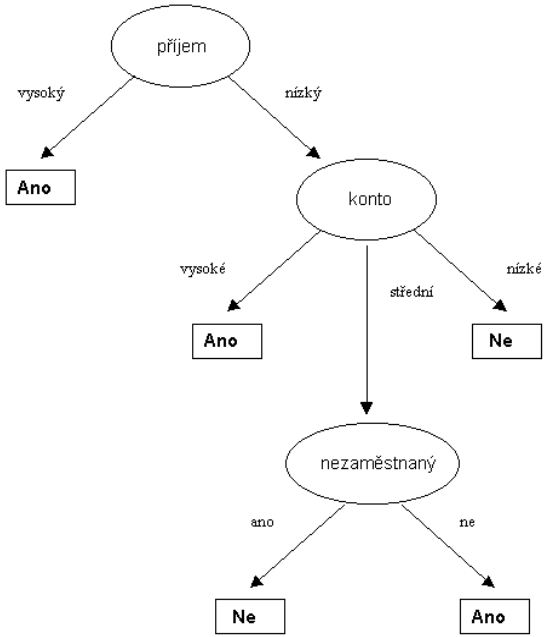
ne: 1x ano, 1x ne -> chyba = 1/2 = 0.5
ano: 2x ano, 3x ne -> chyba = 2/5 = 0.4
Celková chyba (průměrná vážená): (2/7 * 0.5 + 5/7 * 0.4) = 0.43

3. krok: Opakování algoritmu pro podmnožinu s nízkým příjmem

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

Opakujeme kroky 1 a 2 pro data s nízkým příjmem a zbývajícími atributy (konto, pohlaví, nezaměstnaný).



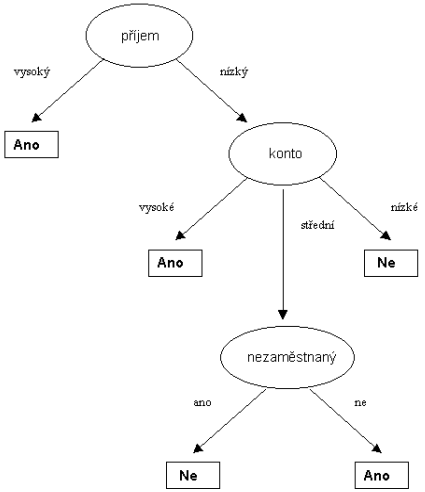


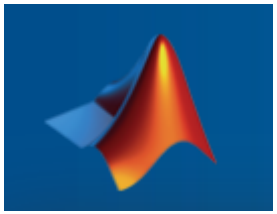
Shrnutí

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

Volba atributu (krok 1 algoritmu)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ

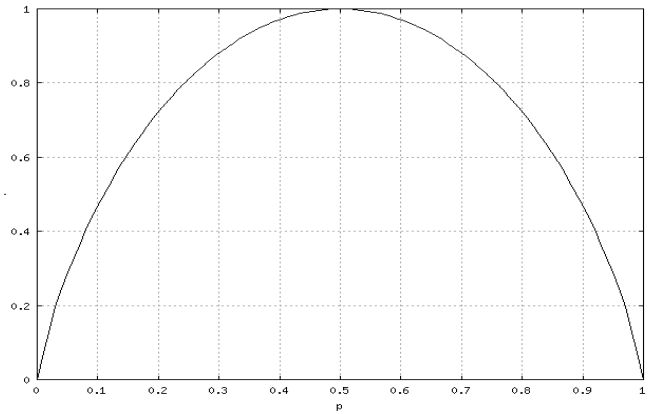
Entropie:

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

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

Pro *T*=2 je:
$p_1 = p$,
$p_2 = 1\text{-}p_1 = 1 - p$,
tedy
$H(\mathbf{p}) = H(p, 1\text{-}p)$

H(p, 1-p)



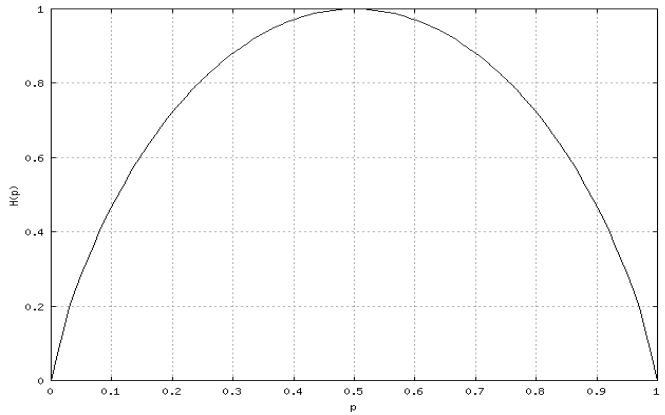
p

Před zkouškou:
$p_1 = p_2 = 0{,}5$
$H(p_1, p_2) = 1$

Po zkoušce:
$p_1 = 0$, $p_2 = 1$
$H(p_1, p_2) = 0$

Entropie spočtená z dat

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



$p_{ano} = 8/12 = 2/3$
$p_{ne} = 4/12 = 1/3$

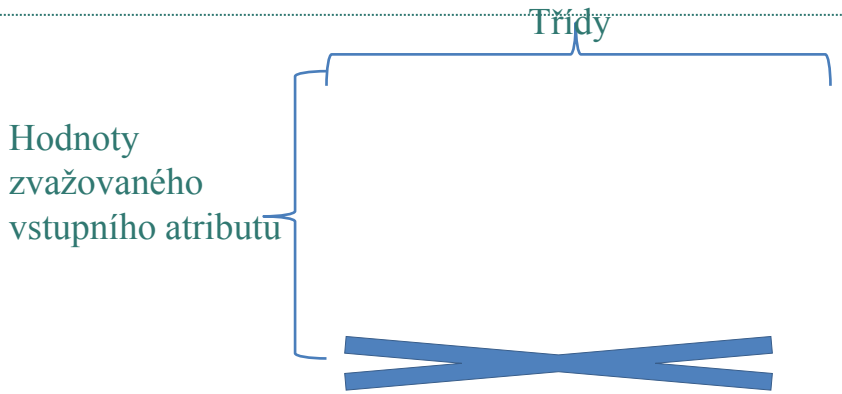
$\mathbf{p} = (2/3, 1/3)$

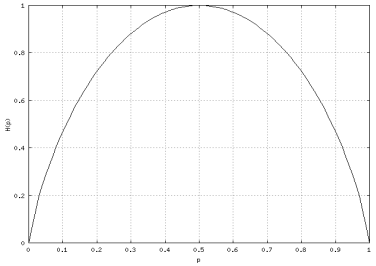
$$H(\mathbf{p}) = -\sum_{t=1}^{2} p_t log_2 p_t = -p_1 log_2 p_1 - p_2 log_2 p_2 = 0{,}92$$

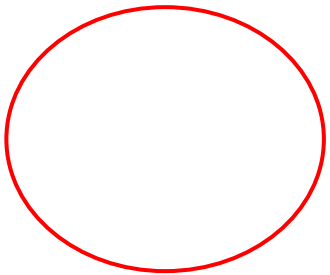
Volba atributu (krok 1 algoritmu)

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

X Y







	ano	ne	suma	entropie
vysoké	4	0	4	0
střední	2	2	4	1
nízké	2	2	4	1
suma			12	

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

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**

Příklad

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

Příklad

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

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

$$\mathbf{H(\text{příjem}) = 0.5747}$$

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

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

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

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

Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

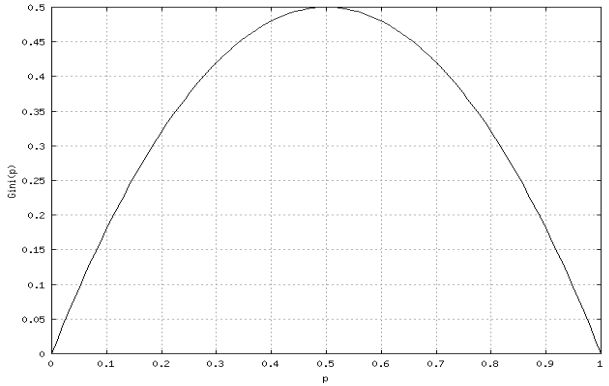
$$H(\text{pohlaví}) = 2/3 * H(\text{pohlaví(muž)}) + 1/3 * H(\text{pohlaví(žena)}) = 2/3 * 1 + 1/3 * 0 = 0.6667$$

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

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

Gini index

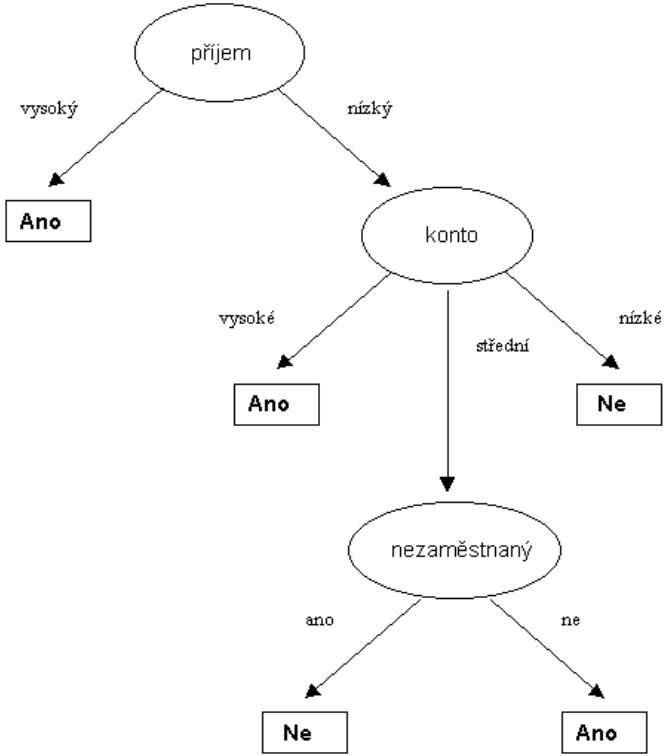
$$\text{Gini} = \sum_{t=1}^{T} p_t(1 - p_t) = 1 - \sum_{t=1}^{T} {p_t}^2$$



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.

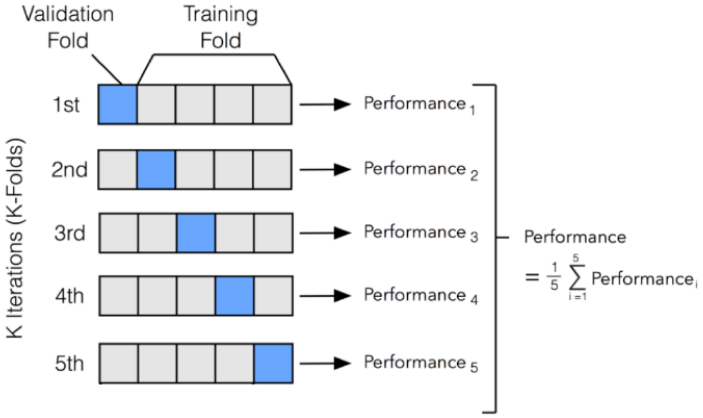
- **pre-pruning** – modifikuje se zastavovací kritérium (krok 3 algoritmu) = větvit se nebude pokud počet příkladů v uzlu klesne pod danou hodnotu nebo pokud relativní počet příkladů jedné třídy překročí danou hodnotu
- **post-pruning** – vytvoří se úplný strom, který se následně redukuje – ukazuje se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

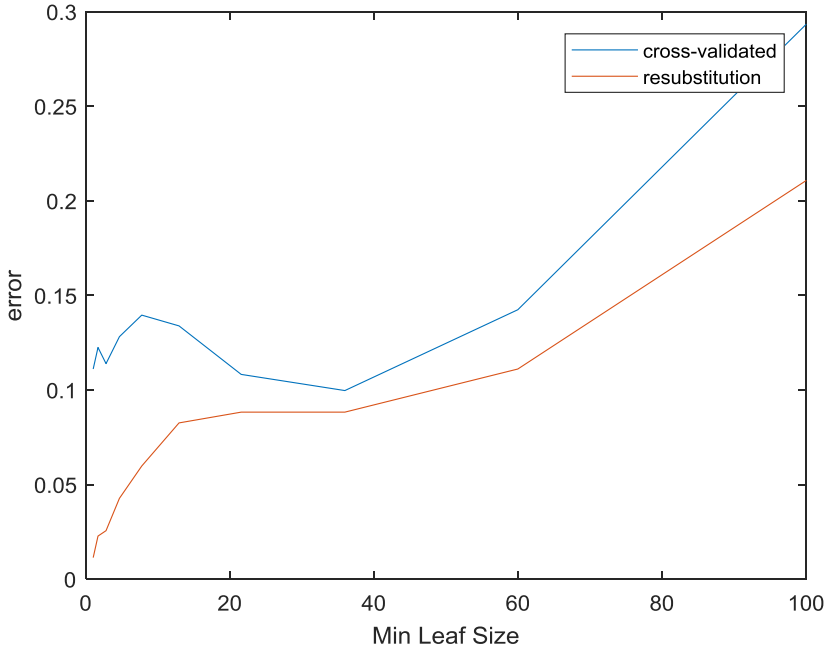
Lze kombinovat pre-pruning s post-pruningem.

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.

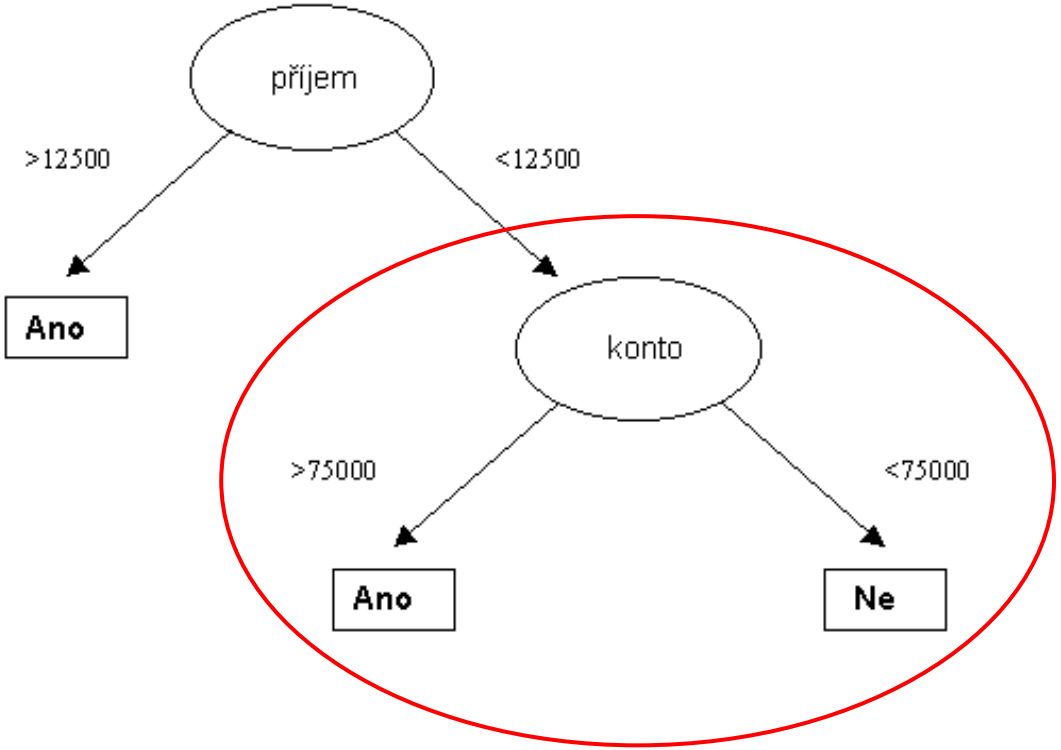
Příklad

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

$$H(konto_{22500}) = 3/12 * H(konto(<22500)) + 9/12 * H(konto(>22500)) =$$
$$1/4 * 0.9183 + 3/4 * 0.5640 = 0.6526$$
$$H(konto_{40000}) = 5/12 * H(konto(<40000)) + 7/12 * H(konto(>40000)) =$$
$$5/12 * 0.9706 + 7/12 * 0.5917 = 0.7497$$
$$H(konto_{55000}) = 6/12 * H(konto(<55000)) + 6/12 * H(konto(>55000)) =$$
$$1/2 * 1 + 1/2 * 0.6500 = 0.8250$$
$$\mathbf{H(konto_{75000})} = 7/12 * H(konto(<75000)) + 5/12 * H(konto(>75000)) =$$
$$7/12 * 0.9852 + 5/2 * 0 = \mathbf{0.5747}$$

Příklad

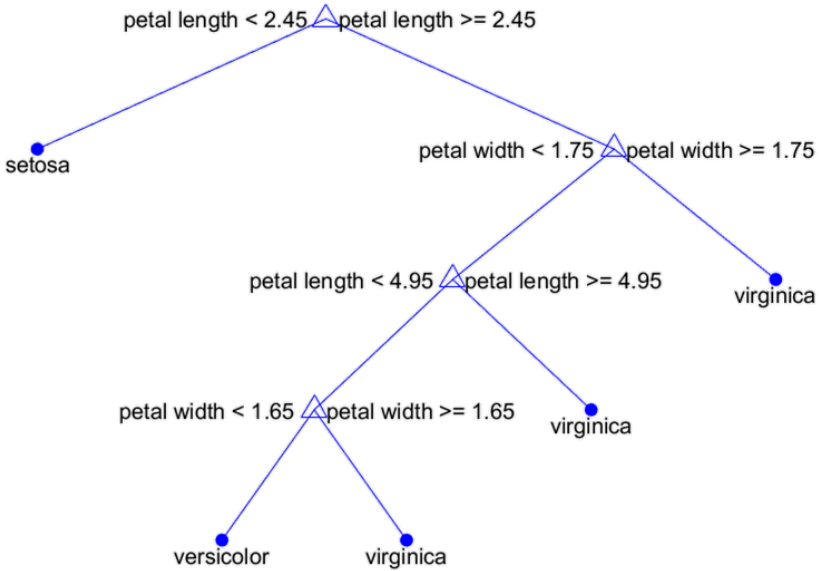
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ



Kategoriální vs numerické atributy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

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



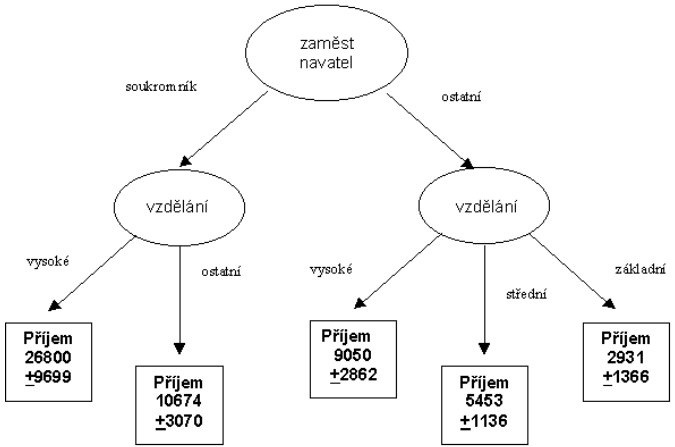
Regresní stromy

SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

- Úloha odhadu hodnoty nějakého numerického atributu.

Volba atributu (krok 1):

- kritérium **redukce směrodatné odchylky**



Použití rozhodovacích stromů

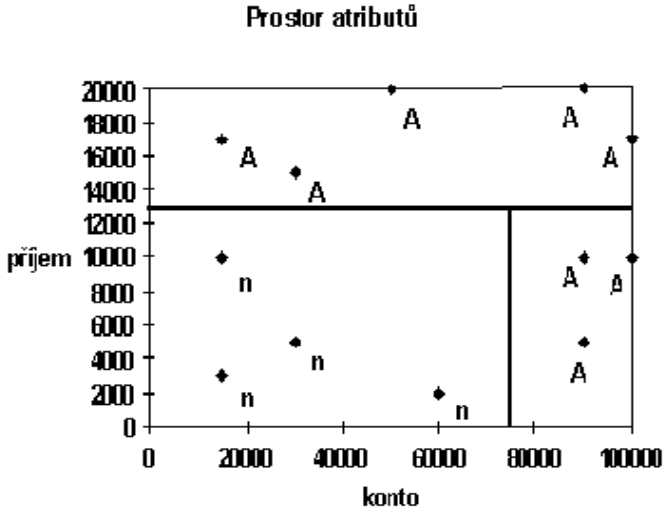
SLEZSKÁ UNIVERZITA
OBCHODNĚ PODNIKATELSKÁ FAKULTA V KARVINÉ

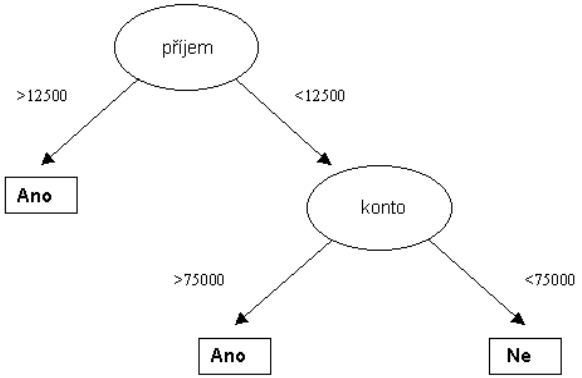
- příklady jsou reprezentovány hodnotami atributů,
- úkolem je klasifikovat příklady do konečného (malého) počtu tříd,
- trénovací data mohou být zatížena šumem,
- trénovací data mohou obsahovat chybějící hodnoty

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