FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

ROZHODOVACÍ STROMY

Autor textu:
Ing. Petr Honzík, Ph.D.

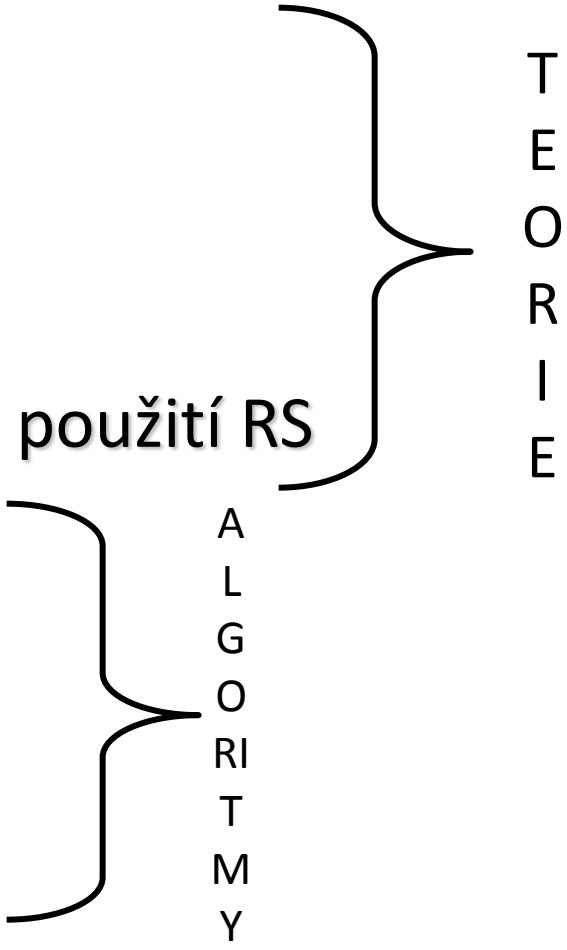
Květen 2014

Komplexní inovace studijních programů a zvyšování kvality výuky na FEKT VUT v Brně
OP VK CZ.1.07/2.2.00/28.0193

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Obsah přednášky

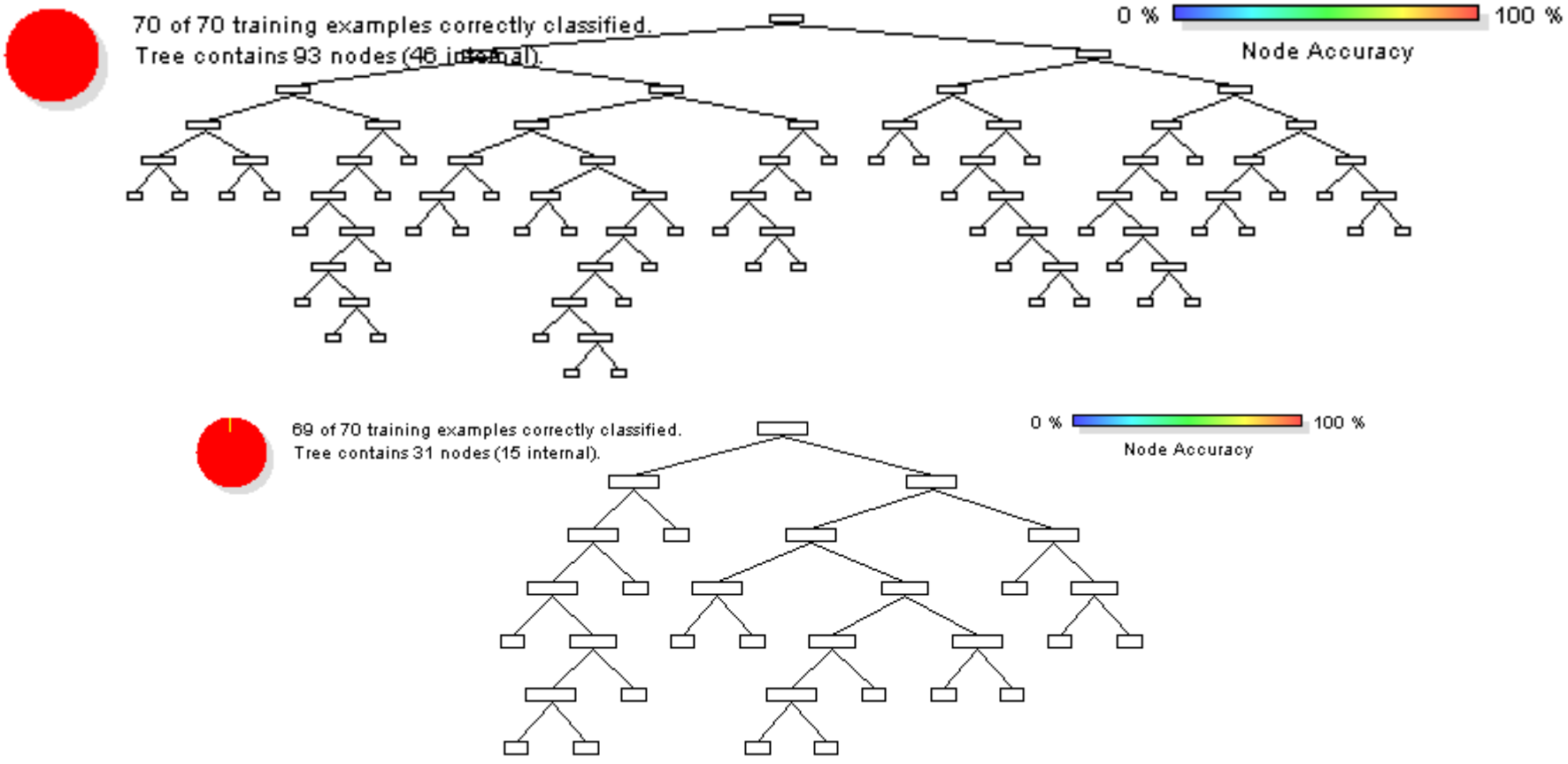
1. Úvod
2. Terminologie
3. Základní dělení
4. Princip tvorby, prořezávání a použití RS
5. Algoritmus ID3
6. C4.5
7. CART
8. Shrnutí



Stromové struktury a RS

- Obsah knihy
- Menu mobilního operátora
- Botanický klíč
- Hierarchie ve firmách, pyramidy, multilevel
- Zákony
- Strukturované znalosti – grafická reprezentace

Strom náhodný a generovaný



http://webdocs.cs.ualberta.ca/~aixplore/learning/DecisionTrees/Applet/DecisionTreeApplet.html

Důležité vlastnosti RS

Výhody

- rychlost predikce
- interpretovatelnost
- minimálně citlivé na irelevantní atributy
- nízké nároky na předzpracování dat

Nevýhody

- ortogonální dělení prostoru - nemusí nalézt ani jednoduchou funkční závislost (více u algoritmu CART)

? proč se nepoužívají v uzlech např. lineární kombinace veličin (např. $b_1x_1+b_2x_2$)

Terminologie I.

- Atribut
- Hodnota atributu
- Záznam

ATRIBUTY výčet **HODNOT ATRIBUTŮ**

VÁHA = {velká, střední, malá} => $X1 = \{X1_1, X1_2, X1_3\}$

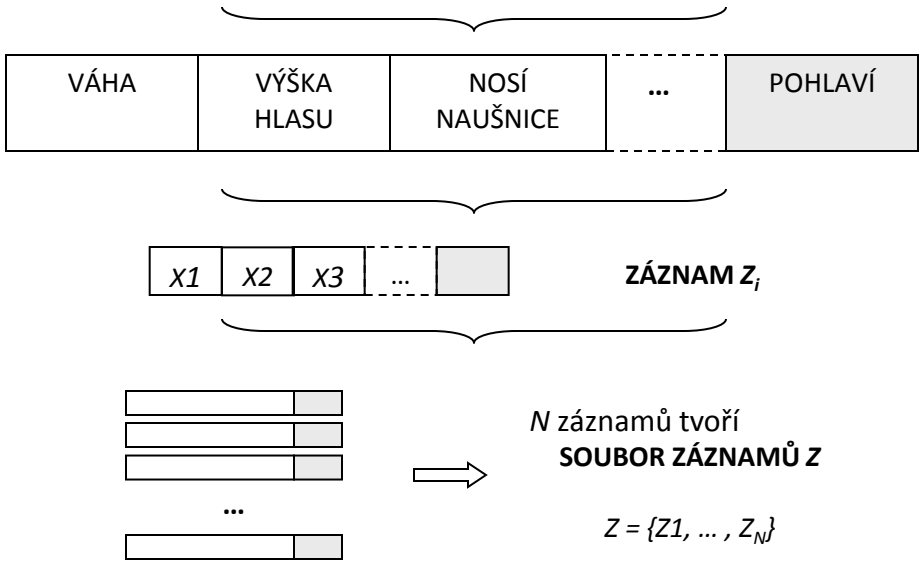
VÝŠKA HLASU = {hluboký, střední, vysoký}=> $X2 = \{X2_1, X2_2, X2_3\}$

NOSÍ NAUŠNICE= {ano, ne} => $X3 = \{X3_1, X3_2\}$

…

KLASIFIKAČNÍ ATRIBUT. Tvořen **TŘÍDAMI;** svou hodnotou určuje **TYP** záznamu.

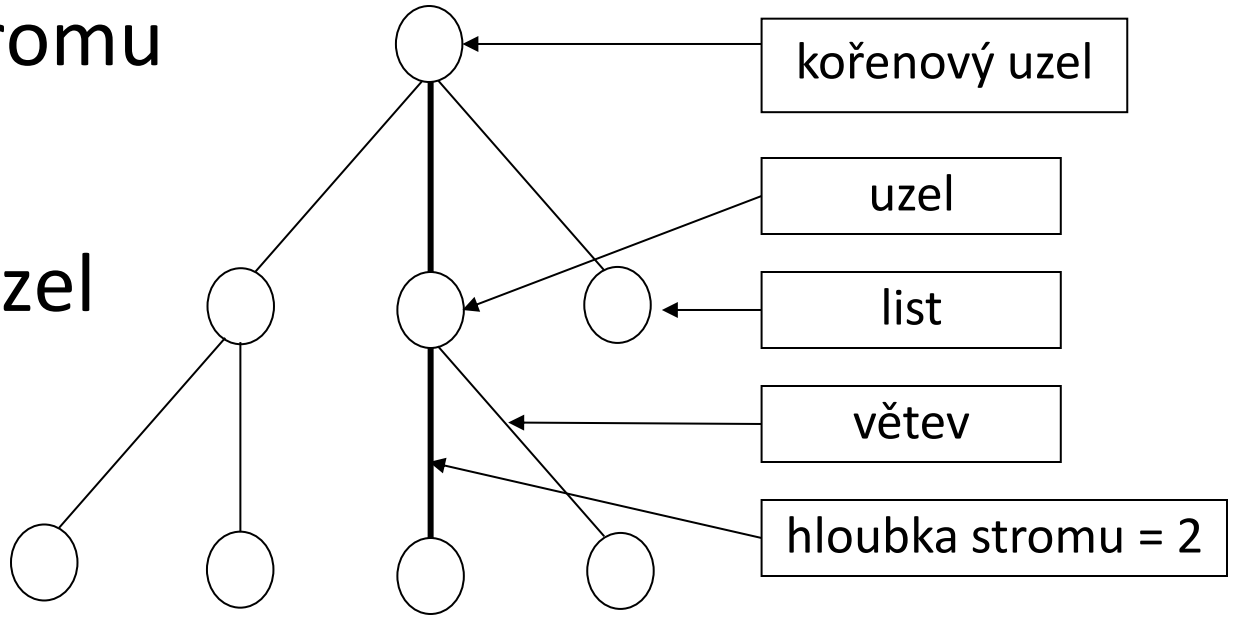
POHLAVÍ = {muž, žena}=> $G = \{G_1, G_2\}$



? vysvětli pojmy atribut, hodnota atributu, záznam, klasifikační atribut

Terminologie II.

- Rozhodovací strom – *hierarchický nelineární systém (model)*
- Hloubka stromu
- Uzel
- Kořenový uzel
- List
- Větev



***VIS: ?** doplň chybějící popisky*

Terminologie III.

- Entropie – *míra neuspořádanosti*
- Přeučený strom
- Prořezávání stromu
- Topologie stromu

? vysvětli význam jednotlivých pojmů

Hlavní problémy

- Jak zautomatizovat tvorbu rozhodovacího stromu na základě množiny dat?
- Jak rozdělovat uzly pro různé typy vstupních a výstupních veličin? Bude ukázáno:
  - vstup kvalitativní / výstup kvalitativní (ID3)
  - vstup kvant.+kval. / výstup kvalitativní (C4.5)
  - vstup kvant.+kval. / výstup kvant.+kval. (CART)
- Jak prořezat přeučení strom? Bude ukázáno:
  - rule-post pruning
  - reduced error pruning

Orientační dělení RS

- Topologie (binární, vícerozměrné)
- Vstupní proměnná (kvalitativní/kvantitativní)
- Výstupní proměnná (kvalitativní/kvantitativní)

Typ RS	CART	CLS	AID
Konkrétní algoritmy	CART tree(S)	CLS, ID3 (TDIDT), C4.5, C5	AID,THAID, CHAID – χ^2
Vstup	kvantitativní kvalitativní	kvantitativní (Cx.x kvalitativní)	všechny typy dat
Výstup	kvantitativní nominální	nominální	
Topologie	binární	dle atributu	všechny typy

Základní princip tvorby RS

- V cyklu opakuj
  1. Získej informace o uzlu
  2. Rozhodni o uzlu, zda bude dál dělen (krok 3.) nebo z něj udělej list a rozhodni o jeho výstupní hodnotě
  3. Vyber nejlepší atribut na větvení
  4. Rozděl data do nových uzlů
- Prořezej strom

- Existují i jiné typy stromů (ADTree – alternating decision tree, Decision tree forests, TreeBoost, …)

ID3 – základní údaje

- Klasifikace - **nominální** vstupy a výstupy
- Větvení podle **výčtu dělicího atributu**
- Kvalita dělení posouzena **entropií** – ML
- Předností je schopnost vybrat z velkého množství atributů ty vhodnější, **vždy** vytvořen **stejný strom** (deterministický)
- **Není garance** vygenerování optimálního stromu

Je algoritmus ID3 stochastický? Proč?

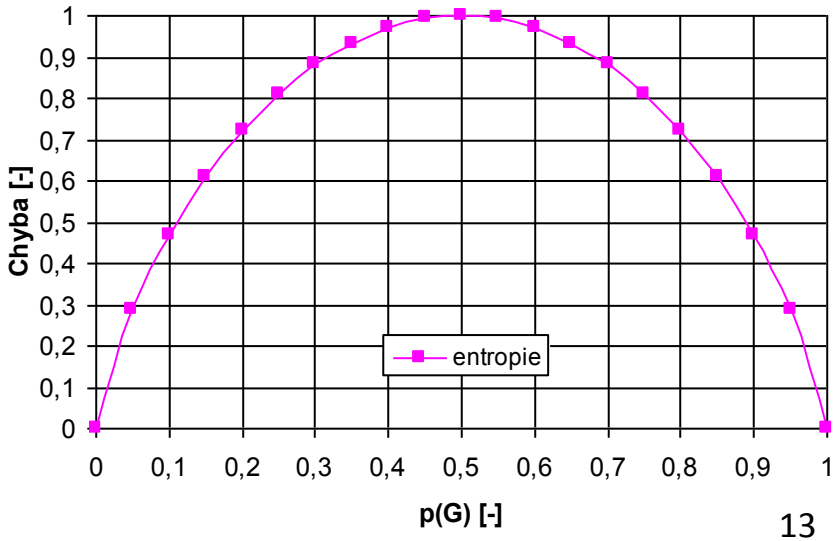
ID3 - entropie

- Míra neuspořádanosti – **entropie**
- Pravděpodobnost klasifikace do třídy G_i

$$\sum_{i=1}^{C} p(G_i) = 1$$

- Entropie

$$H = -\sum_{i=1}^{C} p(G_i) \log_2 p(G_i)$$



Uveď vztah pro výpočet entropie.

ID3 – binární klasifikace

- Binární klasifikace – výpočet entropie

$$\sum_{i=1}^{2} p_i = p_1 + p_2 = \frac{n_1}{n_1 + n_2} + \frac{n_2}{n_1 + n_2} = 1$$

$$H(S) = -\sum_{i=1}^{2} \frac{n_i}{n_1 + n_2} \log_2 \frac{n_i}{n_1 + n_2} =$$

$$-\frac{n_1}{n_1 + n_2} \log_2 \frac{n_1}{n_1 + n_2} - \frac{n_2}{n_1 + n_2} \log_2 \frac{n_2}{n_1 + n_2}$$

ID3 – průměrná neuspořádanost

- Occamovo ostří:

  „nejjednodušší RS konzistentní s trénovacími daty je nejpravděpodobněji ten nejvhodnější“

- Které dělení je nejlepší? **Podmíněná entropie** (též průměrná neuspořádanost), kde $|A|$=počet možných hodnot A, c – počet výstupních tříd G

$$H(S_i) = -\sum_{j=1}^{c} p_i(G_j)\log_2 p_i(G_j)$$

$$H(S \mid A) = \sum_{i=1}^{|A|}\left(\frac{N_i}{\sum_{j=1}^{|A|} N_j}\cdot H(S_i)\right) = \frac{1}{N}\sum_{i=1}^{|A|} N_i \cdot H(S_i) = -\frac{1}{N}\sum_{i=1}^{|A|} N_i \sum_{j=1}^{c}\left(p_i(G_j)\log_2 p_i(G_j)\right)$$

- **Informační zisk** $I(S,A) = H(S) - H(S|A)$

Uveď vztah pro výpočet průměrné neuspořádanosti a informačního zisku.

Tvorba stromu pomocí ID3

1. Informace o uzlu: H(S), počet prvků
2. Rozhodni, zda dělit
3. Vypočítej průměrnou neuspořádanost **dosud nepoužitých** atributů – $H(S|A_i)$
4. Vyber atribut s nejnižším $H(S|A_i)$

ID3 – chybějící atributy

- 2 základní principy
  - Nahraď chybějící hodnotu atributu hodnotou, která se vyskytuje v **daném uzlu nejčastěji**
  - Nahraď chybějící hodnotu atributu hodnotou, která je **nejčastější ve třídě**, kterou záznam popisuje

- V dalších verzích algoritmu typu ID3 (C4.5, C5.0) odlišný přístup

ID3 – zadání příkladu

Osoba č.	Barva vlasů	Výška	Váha	Opalovací krém	Spálila se
1	blond	průměrná	malá	ne	ano
2	blond	velká	průměrná	ano	ne
3	hnědé	malá	průměrná	ano	ne
4	blond	malá	průměrná	ne	ano
5	zrzavé	průměrná	velká	ne	ano
6	hnědé	velká	velká	ne	ne
7	hnědé	průměrná	velká	ne	ne
8	blond	malá	malá	ano	ne

ID3 – řešení příkladu

- Entropie jednotlivých atributů

$$H(S \mid Barva\ vals\mathring{u}) = \frac{4}{8}\left(-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right) + \frac{3}{8}\cdot 0 + \frac{1}{8}\cdot 0 = 0{,}5$$

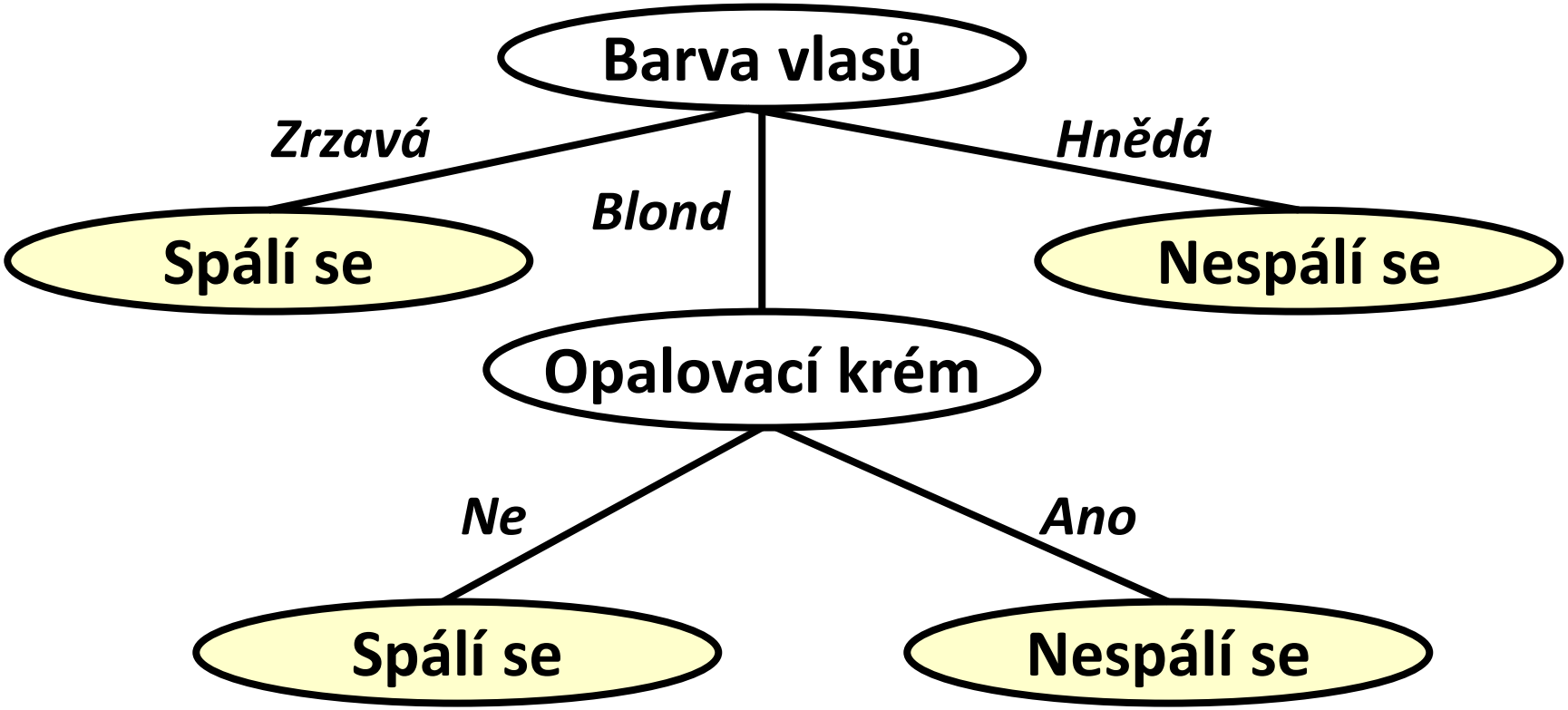
H(S|Výška) = 0,69 *H(S|Váha)* = 0,94 *H(S|Opalovací krém)* = 0,61

- Informační zisk

I(S,A) = H(S) – H(S|A) = 0,95 – 0,5 = 0,45

Umět spočítat …

ID3 - výsledný strom



Ukončovací podmínky

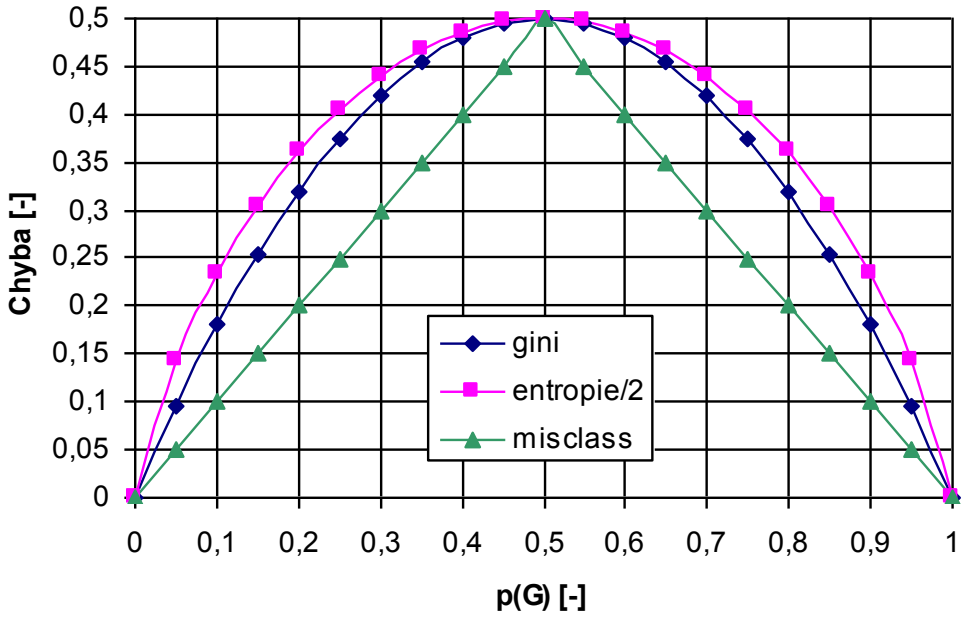
- Maximální hloubka
- Maximální počet uzlů
- Požadovaná přesnost
- Nedostatečný počet trénovacích dat

Uveď alespoň 3 různé ukončovací podmínky pro zastavení větvení RS.

Chybové funkce v RS

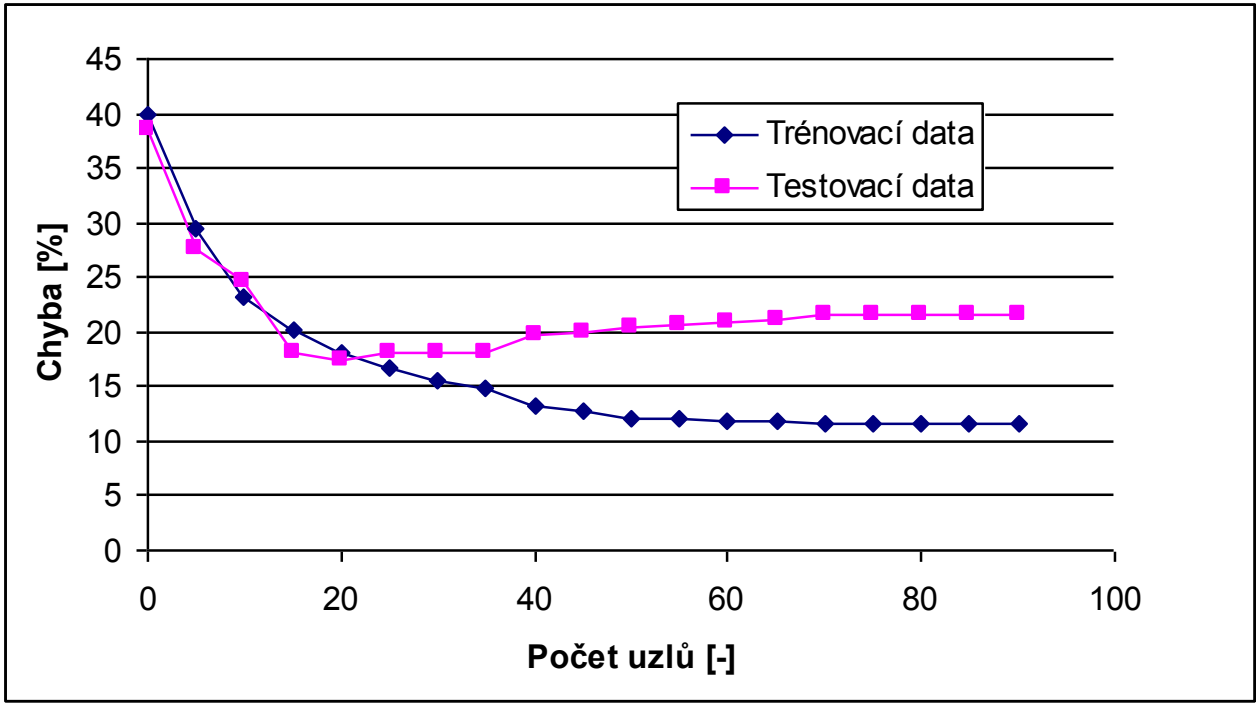
$$Entropie = -\sum_{i=1}^{C} p(G_i)\log_2 p(G_i)$$

$$Gini = 1 - \sum_{i=1}^{c} p(G_i)^2 \qquad Misclass = 1 - p(G_J)$$



Přeučení RS

- Zamezení ukončovací podmínkou
- Prořezávání



Prořezávání RS

- metoda **Reduced-Error Pruning**
- dělí data na trénovací (2/3) a validační (1/3)
- algoritmus:
  - nauč RS na všechna trénovací data (přeuč)
  - opakuj na validačních datech, dokud vzniká nový RS
    - zkus postupně všechny uzly v RS nahradit listem (dle validačních dat) a vyhodnoť všechny nové stromy
    - vyber z těchto nových stromů nejlepší a je-li přesnější než předtím platný RS, nahraď RS tímto novým stromem

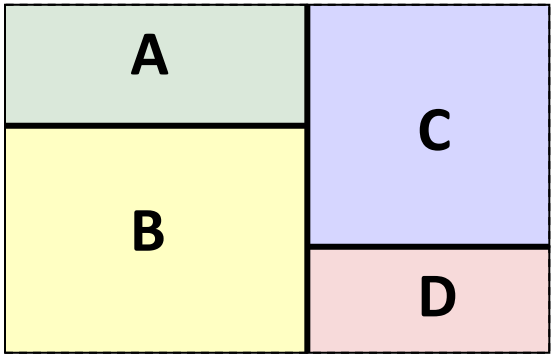
Prořezávání RS

- metoda **Rule Post-Pruning**
- nový model obecnější (zahrnuje i hypotézy pomocí RS nepokryté)
- algoritmus:
  - převeď přeučený RS na pravidla
  - v každém pravidle zkus odstranit všechny kombinace podmínek, urči přesnost (accuracy) a nejlepší podmínku (včetně původní) ulož
  - seřaď a používej pravidla v pořadí dle accuracy

Čím je specifický model po prořezávání metodou „Rule Post Pruning“?

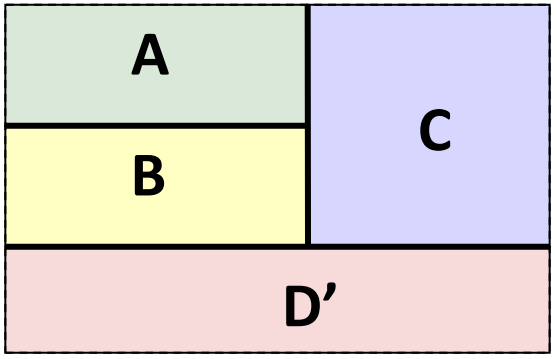
Prořezávání RS

- metoda **Rule Post-Pruning**



- rozdělení prostoru původním RS, platné i po přepisu do pravidel
- co se stane, pokud z pravidla D vypuštěna první podmínka (první dělení) a vznikne nové pravidlo D'?

- pokud bude pravidlo B před D', tak nic…
- pokud se ocitne pravidlo D' před B, dojde k novému rozdělení prostoru, které v původním RS neexistovalo



C4.5 – základní vlastnosti

- Algoritmus **rozšiřuje ID3** – vychází tedy z entropie a informačního zisku
- Hlavní zlepšení oproti ID3
  - **Vstupní kvantitativní** veličiny
  - Lepší ošetření **chybějících hodnot** atributů

  $\Rightarrow$prakticky použitelné
- Další zlepšení C5.0, chráněno licencí, autorská práva

Jaký je rozdíl mezi algoritmy C4.5 a ID3?

C4.5 – entropie a atributy

- U kvalitativních atributů jako u ID3
- U kvantitativních nastavení **prahu** θ, který atribut dělí na dvě poloviny (binární větvení)
- Kvantitativní atribut je převeden v uzlu na nominální atribut (práh – binární, intervaly - vícerozměrné)
- Kvantitativní atribut však může být použit opakovaně pro dělení (do nového uzlu se předává původní kvantitativní informace)

C4.5 – informace o uzlu

- Stav v uzlu S
  - $|S|$, počet prvků v uzlu S
  - *H(S),* entropie v uzlu S
  - *{A}*, množina atributů A_i, které lze použít k větvení
- Ohodnocení kvality dělení uzlu dle *A*
  - $|A|$, výčet atributu (počet možných hodnot *A*)
  - *I(S,A)*, informační zisk při větvení dle *A*
  - *P(S|A)*, **poměrový zisk** při větvení dle *A*
  - $I_P(S,A)$, **poměrový informační zisk** (gain ratio)

C4.5 – výpočty

- *I(S,A)*, informační zisk

$$I(S,A) = H(S) - H(S \mid A)$$

- *P(S,A)*, poměrový zisk (*kvůli ID, datumu, …*)

$$P(S \mid A) = -\sum_{i=1}^{|A|} \frac{|S_i|}{|S|} \log_2 \left(\frac{|S_i|}{|S|} \right)$$

- $I_P(S,A)$, poměrový informační zisk

$$I_P(S,A) = \frac{I(S,A)}{P(S,A)}$$

C4.5 – kritéria dělení

- Dvě kritéria
  - Dělí se atributem s **největším poměrným informačním ziskem** $I_P(S,A)$
  - Dělit lze pouze podle atributu, jehož informační zisk je **minimálně průměrný** (průměr spočten za použití všech použitelných atributů A_i)
- Vypočítám tedy pro všechny atributy *I(S,A),* určím jejich průměr a pouze pro hodnoty průměrné a vyšší zjistím poměrný informační zisk $I_P(S,A)$

C4.5 – chybějící atributy

- *I(S,A)*, informační zisk

$$I(S,A) = \frac{|S - S_0|}{S}\left(H(S) - H(S \mid A)\right)$$

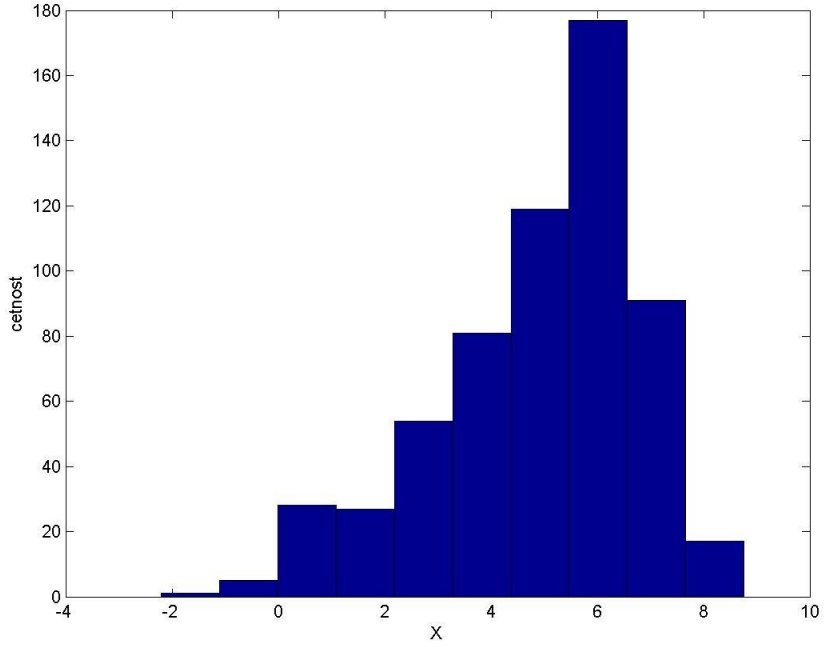
kde $|S_0|$ je počet prvků s chybějící hodnotou A

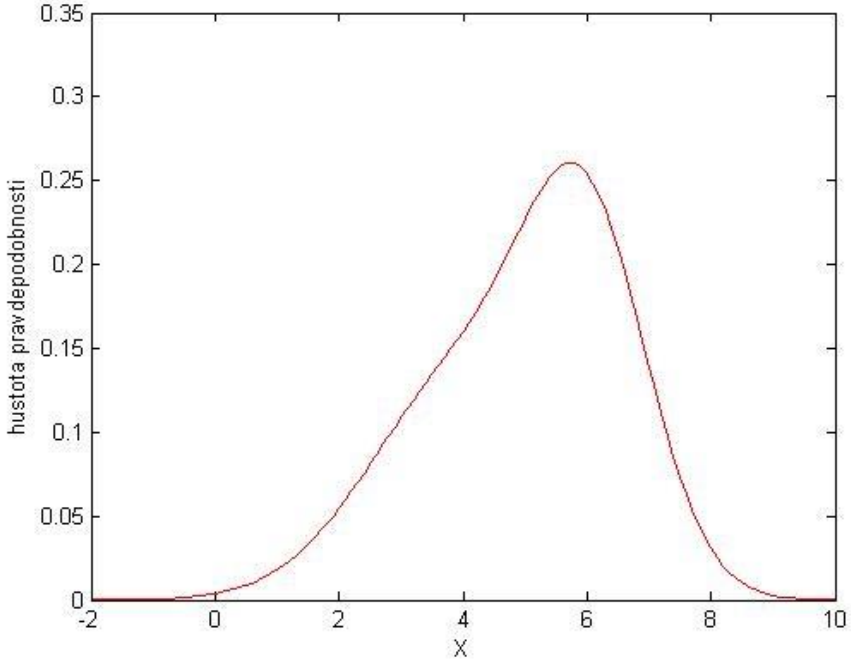
- *P(S,A)*, poměrový zisk

$$P(S,A) = -\frac{|S_0|}{|S|}\log_2\left(\frac{|S_0|}{|S|}\right) - \sum_{i=1}^{|A|}\frac{|S_i|}{|S|}\log_2\left(\frac{|S_i|}{|S|}\right)$$

C4.5 – kvantitativní atribut

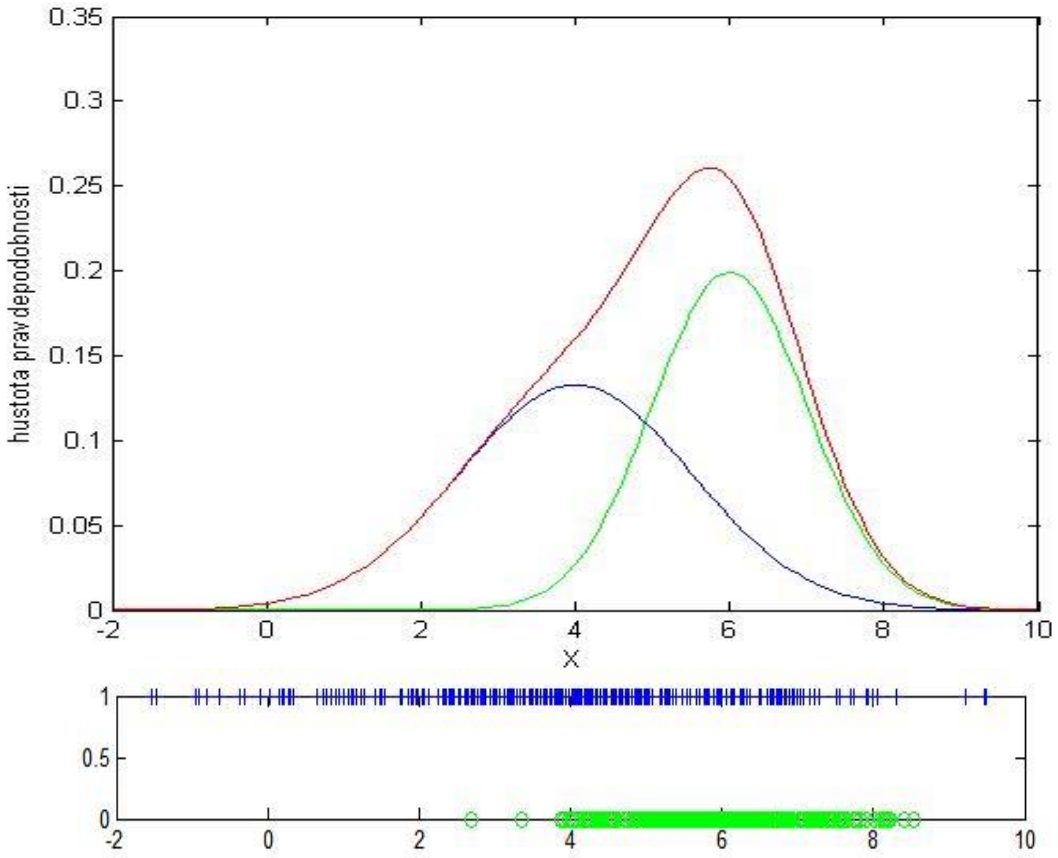
- Vypočti pro kvantitativní veličinu *X* optimální práh θ maximalizující informační zisk (minimalizující entropii).





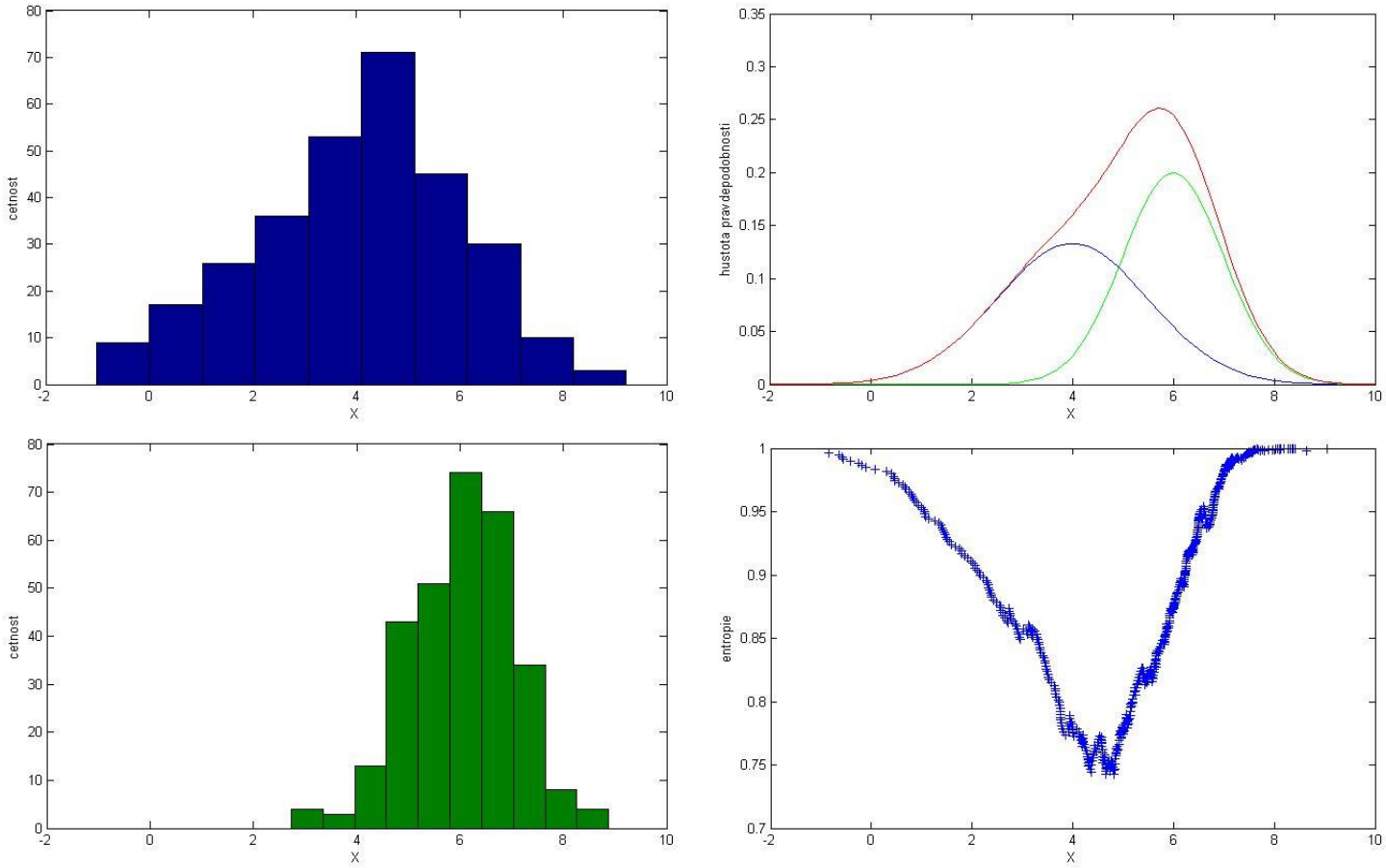
C4.5 – kvantitativní atribut

- na jehož základě klasifikujeme do 2 tříd



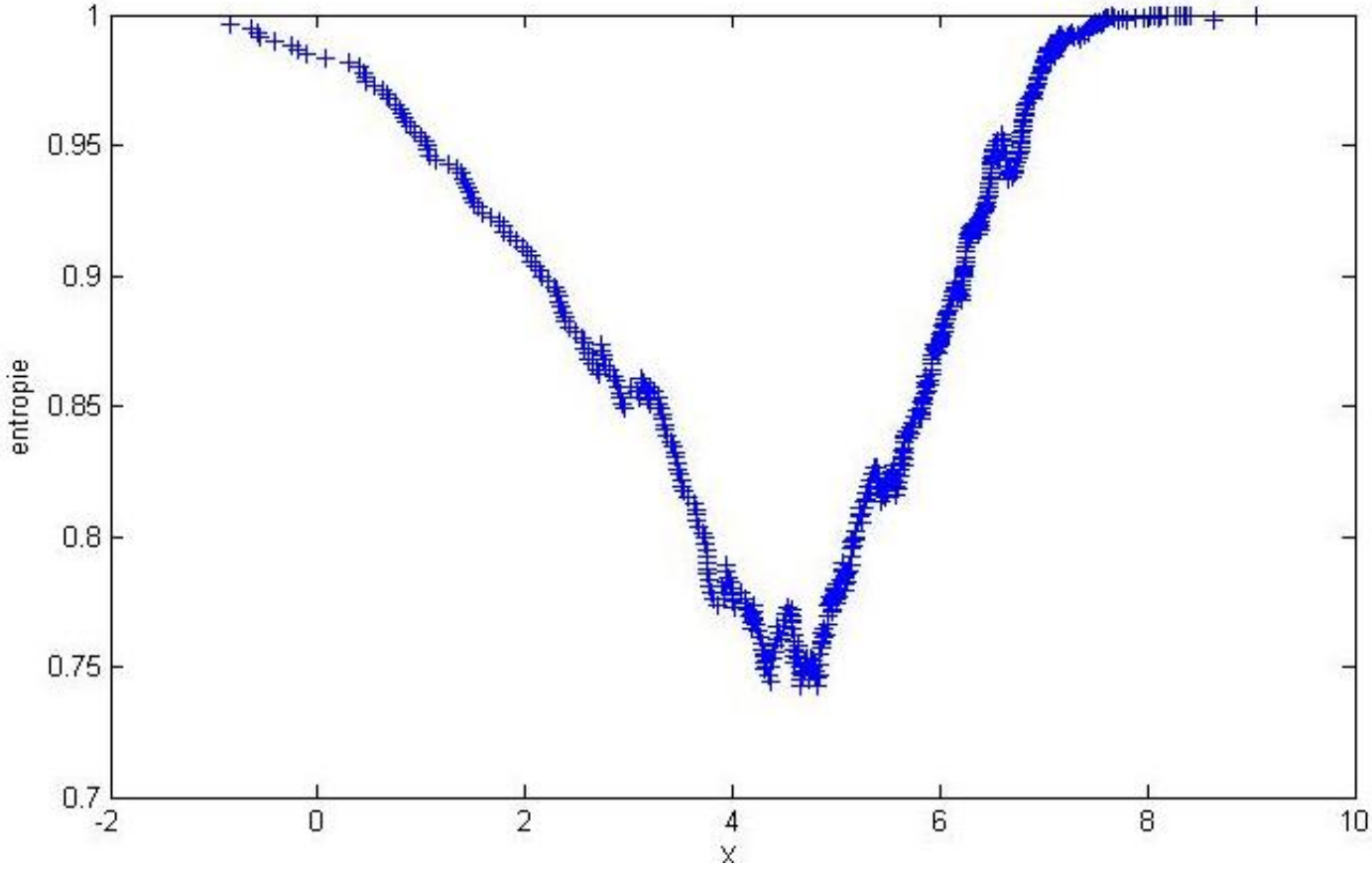
C4.5 – kvantitativní atribut

- Pro jednotlivé prahy θ hledáme optimum (maximální informační zisk = min. entropie)



C4.5 – kvantitativní atribut

- Z grafu stanoven optimální práh $\theta = 4{,}8$



C4.5 – příklad

den	obloha	teplota	Vlhkost	vítr	hrát tenis?
1.	slunečno	29,5	85	slabý	**NE**
2.	slunečno	26,5	90	silný	**NE**
3.	zataženo	28,5	78	slabý	**ANO**
4.	déšť	21	96	slabý	**ANO**
5.	déšť	20	80	slabý	**ANO**
6.	déšť	18,5	70	silný	**NE**
7.	zataženo	18	65	silný	**ANO**
8.	slunečno	22	95	slabý	**NE**
9.	slunečno	21	70	slabý	**ANO**
10.	déšť	24	80	slabý	**ANO**
11.	slunečno	24	70	silný	**ANO**
12.	zataženo	22	90	silný	**ANO**
13.	zataženo	27	75	slabý	**ANO**
14.	déšť	21,5	80	silný	**NE**

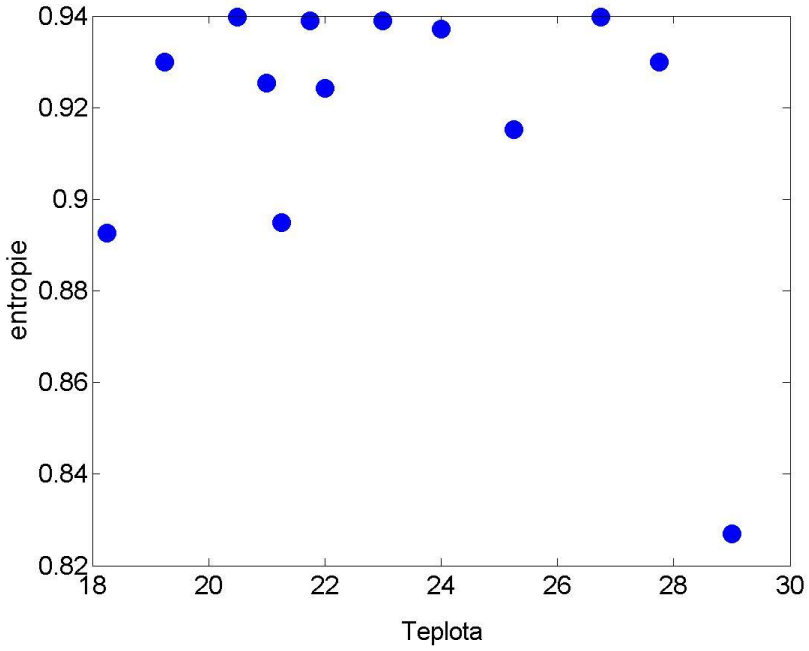
Bude se hrát 15. den? **<slunečno,20,90,silný>**

C4.5 – příklad, urči *I(S,A)*

$H(S|Obloha)=-(5+5)/14 \cdot (0,6\log_2 0,6+0,4\log_2 0,4)=0,694$

$\mathbf{I(S|Obloha)}=H(S)-H(S|Obloha)=0,94-0,694=\mathbf{0,247}$

$H(S|Teplota \leq 29)=0,827$



$\mathbf{I(S,Teplota \leq 29)}=H(S)-H(S|Teplota$

C4.5 – příklad, *vyber A*

I(S,Obloha)=H(S)-H(S|Obloha)=0,94-0,694=0,247

I(S,Teplota ≤ 29*)=H(S)-H(S|Teplota* ≤ 29*)=0,94-0,827=0,113*

I(S,Vlhkost ≤ 80*)=H(S)-H(S|Vlhkost* ≤ 80*)=0,94-0,838=0,102*

I(S,Vítr)=H(S)-H(S|Vítr)=0,94-0,892=0,048

Průměr $\bar{I}(S,A)$**=0,128**,

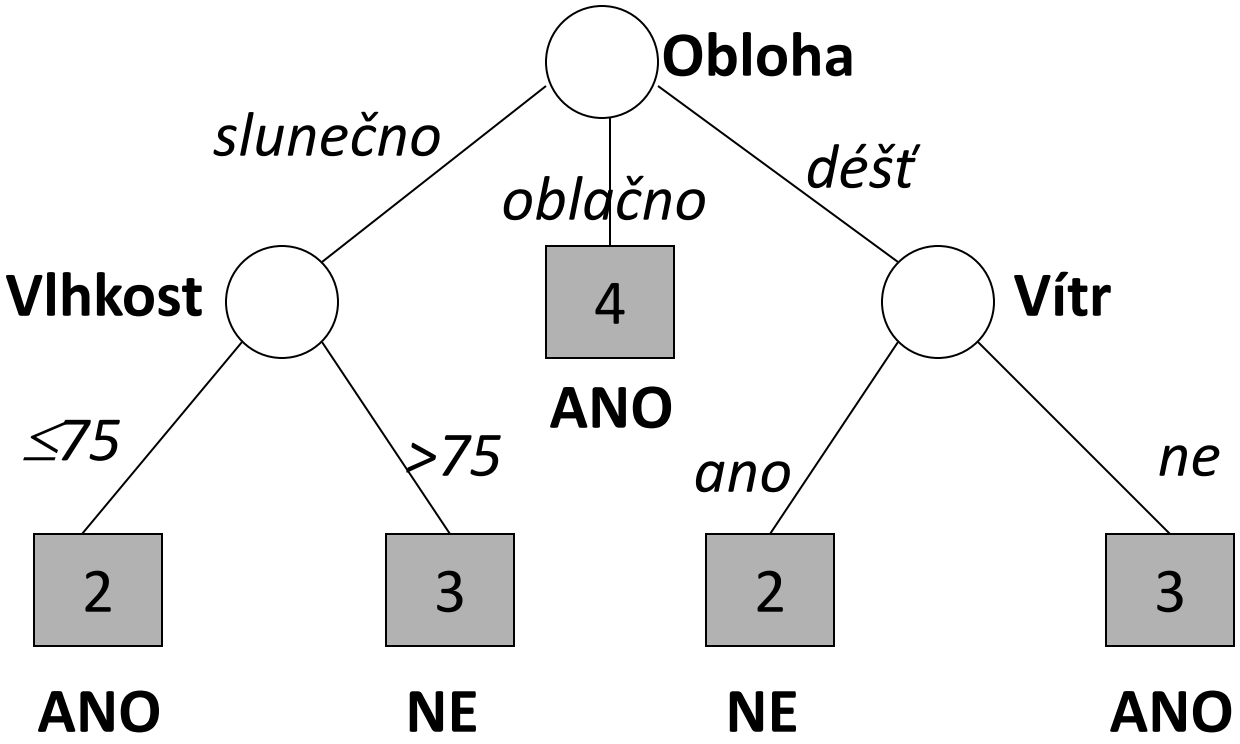
I(S,A) $\geq \bar{I}(S,A)$ splňuje pouze **obloha**

Následující výpočet pro tento případ tedy není nutný (jen jeden atribut splňuje podmínku), nicméně:

P(S,Obloha)= $2 \cdot \left(\frac{5}{14}\log_2\frac{5}{14}\right) + \left(\frac{4}{14}\log_2\frac{4}{14}\right) = 1{,}577$

I_P*(S,Obloha)=0,247 / 1,577 =* ***0,157***

C4.5 – příklad



Bude se hrát 15. den? **<slunečno,teplota=20,vlhkost=90,silný>**

CART - regresní

- CART – Classification and Regression tree
- binární topologie
- libovolné vstupní i výstupní proměnné
- dělení podle jednoho atributu – může způsobit, že není schopen aproximovat jednoduchou závislost

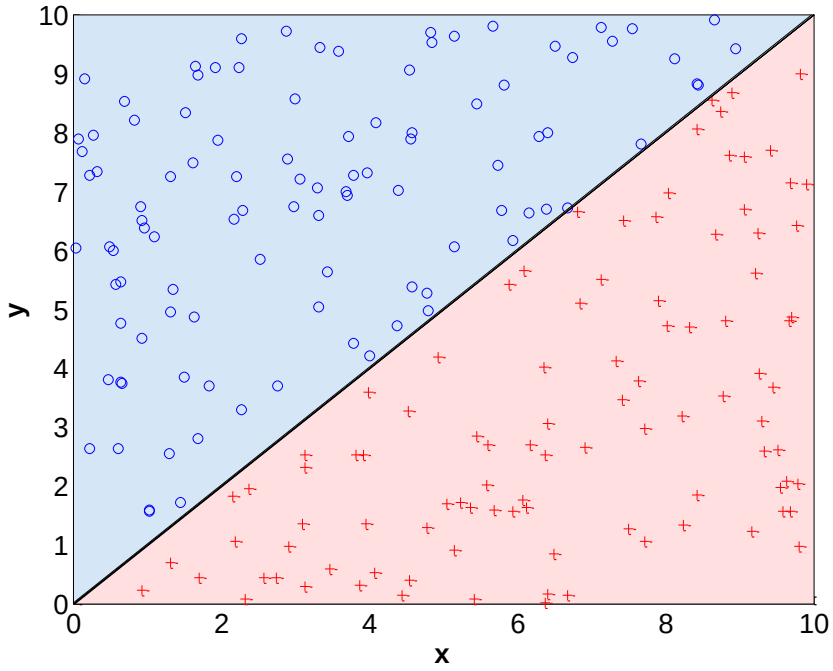
CART – příklad

- Modelovaná závislost:

$f(x,y)=x-y$

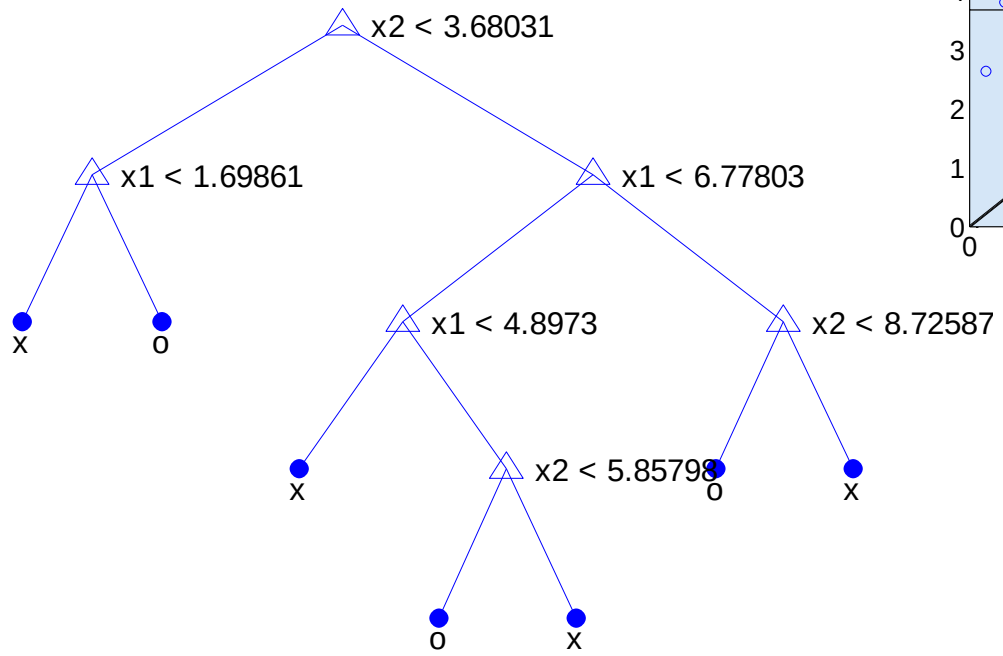
$f(x,y)>0 \Rightarrow$ o

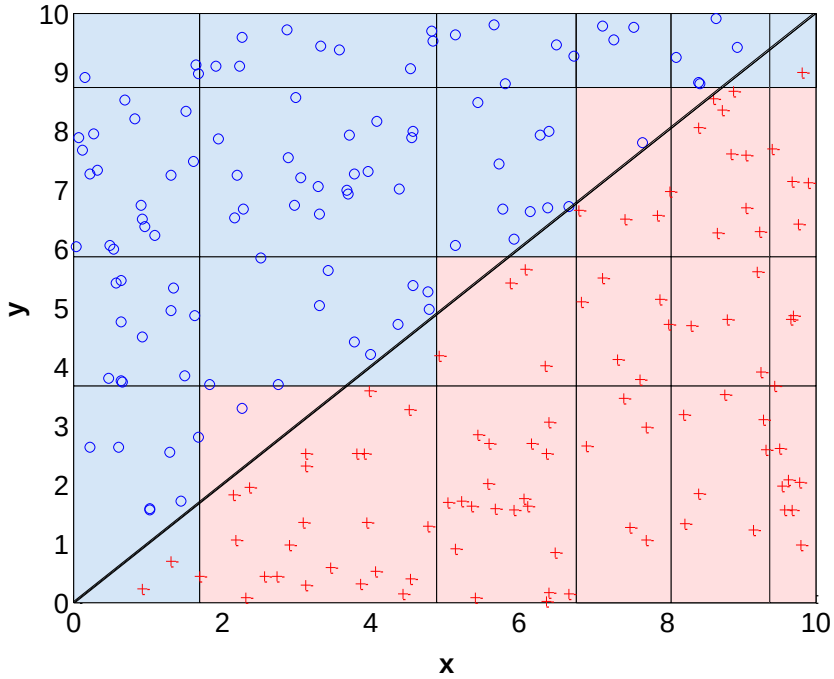
$f(x,y)\leq 0 \Rightarrow +$



CART – řešení

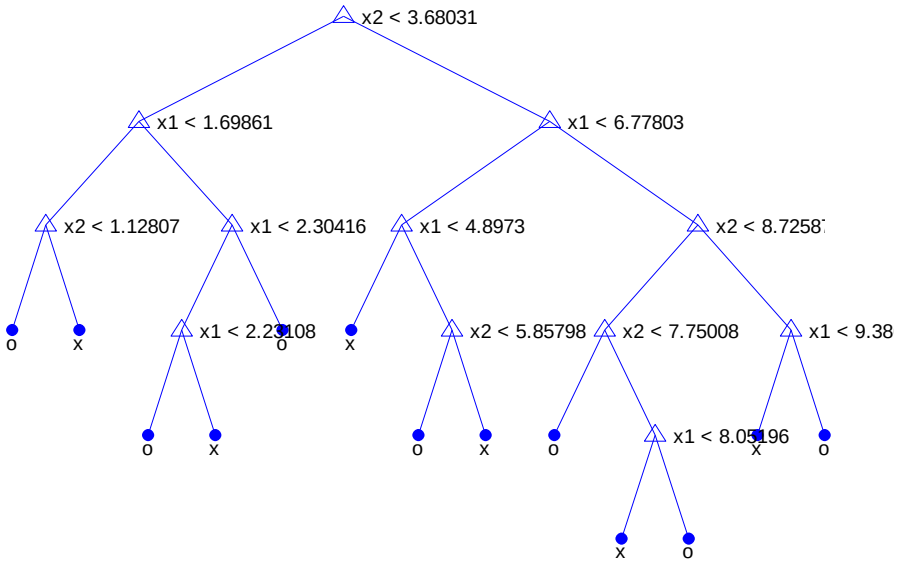
dělicí podmínka = 10 prvků
šum =0 %

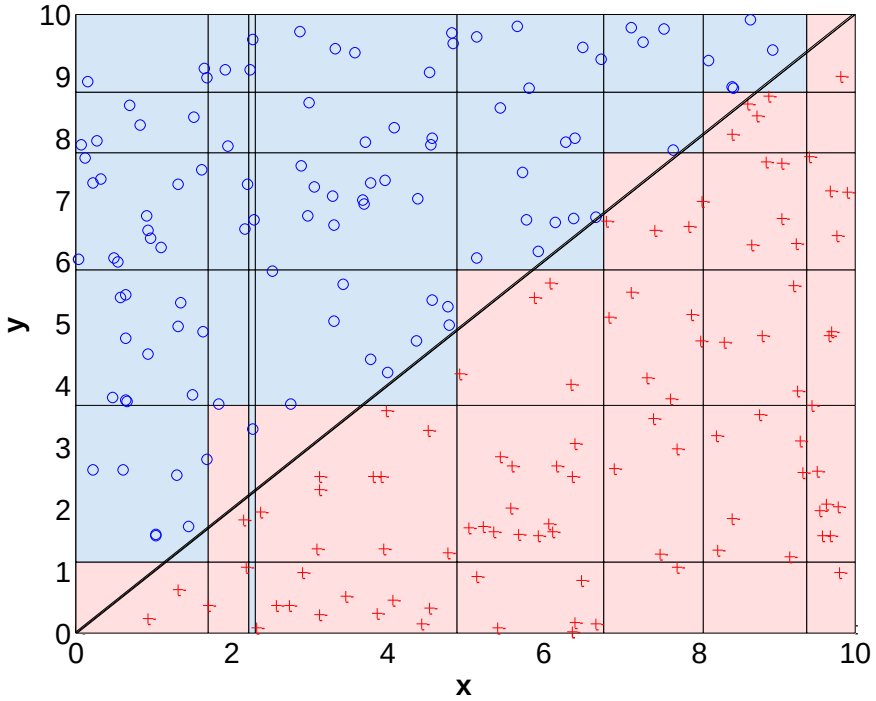




CART – řešení

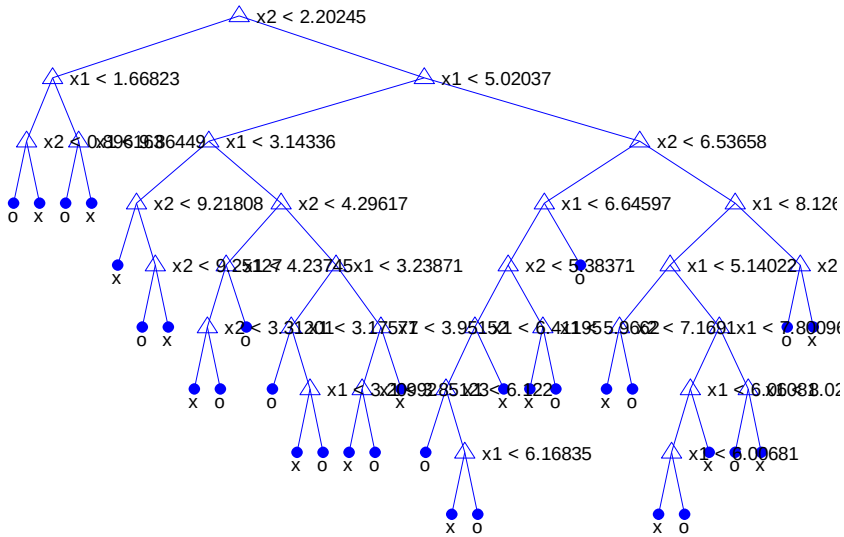
dělicí podmínka = 2 prvky
šum =0 %

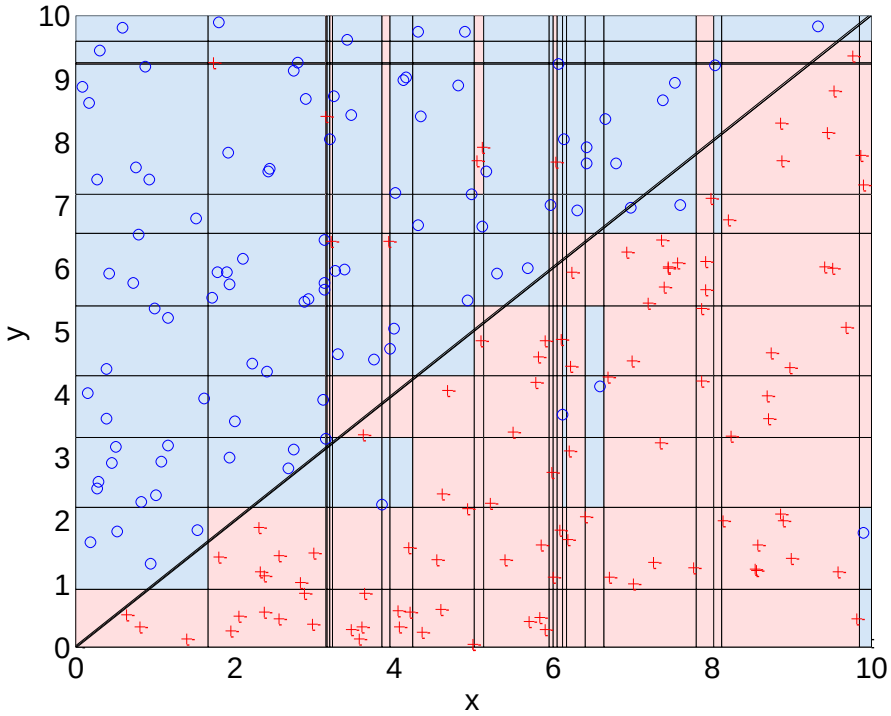




CART – řešení

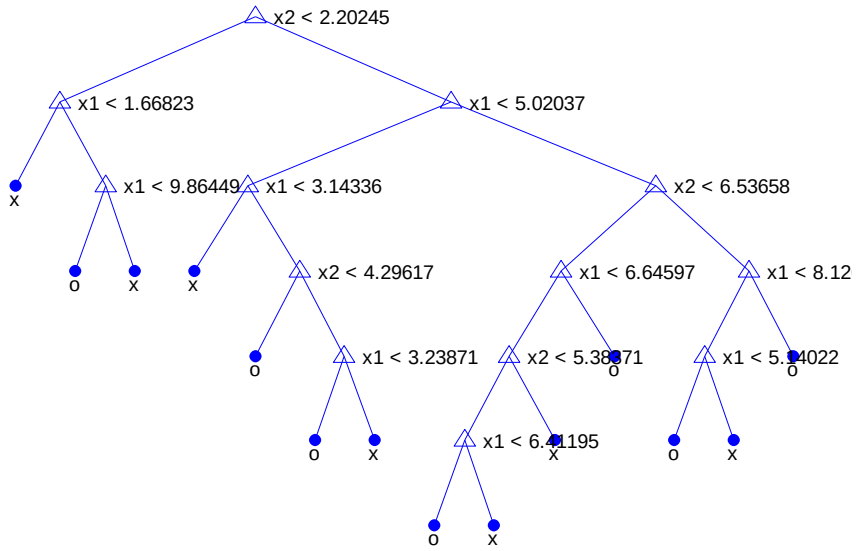
dělicí podmínka = 2 prvky
šum =5 %

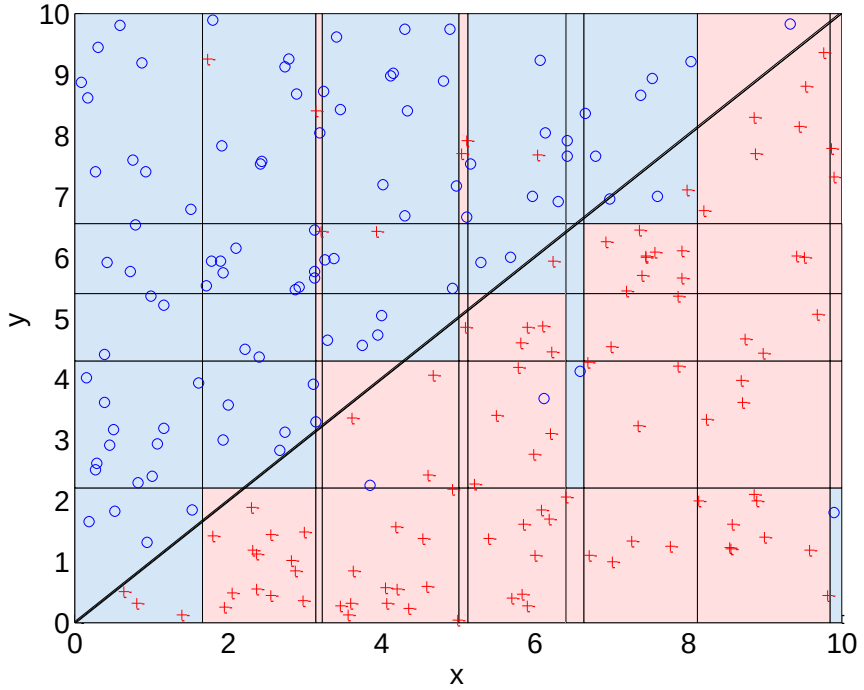




CART – řešení

dělicí podmínka = 10 prvky
šum =5 %





CART – použití

- neumí nalézt vícerozměrnou závislost
- skutečnost, že nelze vygenerovat kvalitní strom neznamená, že v datech není **jednoduchá** funkční závislost
- řešením při generování složitých stromů je zpřísnit dělící podmínky (generalizace vs. přeučenost)

CART – regresní

- V cyklu opakuj
  1. Získej informace o uzlu
  2. Rozhodni o uzlu, zda bude dál dělen
  3. Vyber nejlepší atribut na větvení
  4. Rozděl data do nových uzlů/listů
- Prořezej strom

1. Informace o uzlu U_k

1. Průměr výstupní veličiny	2. Pravděpodobnost uzlu
$\bar{y}_{U_k} = \frac{\sum_{i=1}^{N_k} y_i}{N_k}$	$p(U_k) = \frac{N_k}{N}$
3. Rozptyl v uzlu	**4. Riziko**
$s^2(U_k) = \frac{\sum_{i=1}^{N_k} (y_i - \bar{y}_{U_k})^2}{N_k}$	$risk(U_k) = p(U_k) \cdot s^2(U_k) = \frac{\sum_{i=1}^{N_k} (y_i - \bar{y}_{U_k})^2}{N}$

2. Rozdělit uzel U_k?

- Minimální požadovaný počet prvků N_k v uzlu U_k

$$N_k > N_{\min}$$

- Posouzení skutečného a požadovaného rozptylu (přesnosti) v uzlu U_k

$$s^2(U_k) > s^2_{\min} \qquad s^2(U_k) > 10^{-6} \cdot s^2(U_1)$$

3. Nejlepší dělení U_k

- Výpočet metodou nejmenších čtverců – minimalizace rozptylu (tedy chyby *Err*).
- Vždy binární dělení, U_L (U_P) rozptyl v levém (pravém) novém uzlu

$$Err(U) = \sum_{i=1}^{|U|} (y_i - \bar{y})^2 - \left[\sum_{i=1}^{|U_L|} (y_{Li} - \bar{y}_L)^2 + \sum_{i=1}^{|U_P|} (y_{Pi} - \bar{y}_P)^2 \right]$$

- Vážená metoda

$$Err(U) = \frac{\sum_{i=1}^{|U|} (y_i - \bar{y})^2}{|U|} - \left[\frac{p(U_L)}{p(U)} \cdot \frac{\sum_{i=1}^{|U_L|} (y_{Li} - \bar{y}_L)^2}{|U_L|} + \frac{p(U_P)}{p(U)} \cdot \frac{\sum_{i=1}^{|U_P|} (y_{Pi} - \bar{y}_P)^2}{|U_P|} \right]$$

4. Nejlepšího dělení, ukončovací podmínky

- Volí se atribut s **nejmenší chybou *Err***
- Ukončovací podmínky
  - Rozptyl menší než požadované minimum
  - Hloubka stromu
  - Počet prvků (v uzlu, v poduzlech)
  - Zpřesnění větvením menší než požadované minimum

CART – prořezávání

- Je-li normovaný součet chyb listů větší než chyba jejich nadřazeného uzlu, odstraň listy.
- **Prořezávání** umožňuje hlubší zásah do struktury stromu. Je-li normovaný součet chyb všech listů spadající pod **libovolný** uzel větší než chyba tohoto uzlu, je odstraněn celý podstrom

Shrnutí

- **Rozhodovací strom** je hierarchický nelineární systém umožňující nalezení a uložení znalostí a jejich využití k analýze nových dat. Rozhodovací stromy jsou primárně používány pro klasifikaci kvalitativních závislých proměnných na základě vstupních atributů.
- Hlavními přednostmi a charakteristickými rysy rozhodovacích stromů jsou: **hierarchická struktura**, **nelinearita**, **srozumitelnost a čitelnost**, **flexibilita** (co do typu analyzovaných dat i co do topologie) a **existence algoritmů**, které umožňují jejich automatické vytváření. Příkladem algoritmů jsou např. ID3, CART nebo CHAID.
- Rozhodovací strom se skládá z **kořenového uzlu** a dalších **uzlů** a **listů**. **Větve** spojují dva objekty (uzel-uzel nebo uzel-list) ze sousední hierarchické úrovně. Při průchodu uzlem jsou data rozdělena na základě podmínky do větví z uzlu vycházejících. Podmínka se může týkat hodnoty jednoho nebo kombinace více atributů. Dosažení listu při průchodu rozhodovacím stromem vede ke klasifikaci nebo predikci hodnoty výstupní veličiny analyzovaného záznamu.
- Dalšími důležitými pojmy jsou **přeučený strom** (extrahoval chybné znalosti ze šumu či chyb v trénovacích datech), **prořezávání** (metody ke zmenšování přeučených nebo příliš komplexních stromů) a **ukončovací podmínka** (určuje, kdy algoritmus přestává dále vytvářet rozhodovací strom; špatné nastavení vede k přeučení).