Stromy typu CART - pokračování

	uzel	n_A	n_B	p_A	p_B	p_t	$Gini = 1 - p^2_A - p^2_B$	$p_t * Gini$	
D1	t_1	150	50	3/4	1/4	1/2	$1 - (3/4)^2 - (1/4)^2 = 3/8$	1/2 * 3/8	0,1875
	t_2	50	150	1/4	3/4	1/2	$1 - (1/4)^2 - (3/4)^2 = 3/8$	1/2 * 3/8	0,1875
								celkový	**0,375**
D2	t_3	100	200	1/3	2/3	3/4	$1 - (1/3)^2 - (2/3)^2 = 4/9$	3/4 * 4/9	0,3333
	t_4	100	0	1	0	1/4	1-1-0 = 0	1/4 * 0	0
								celkový	**0,3333**

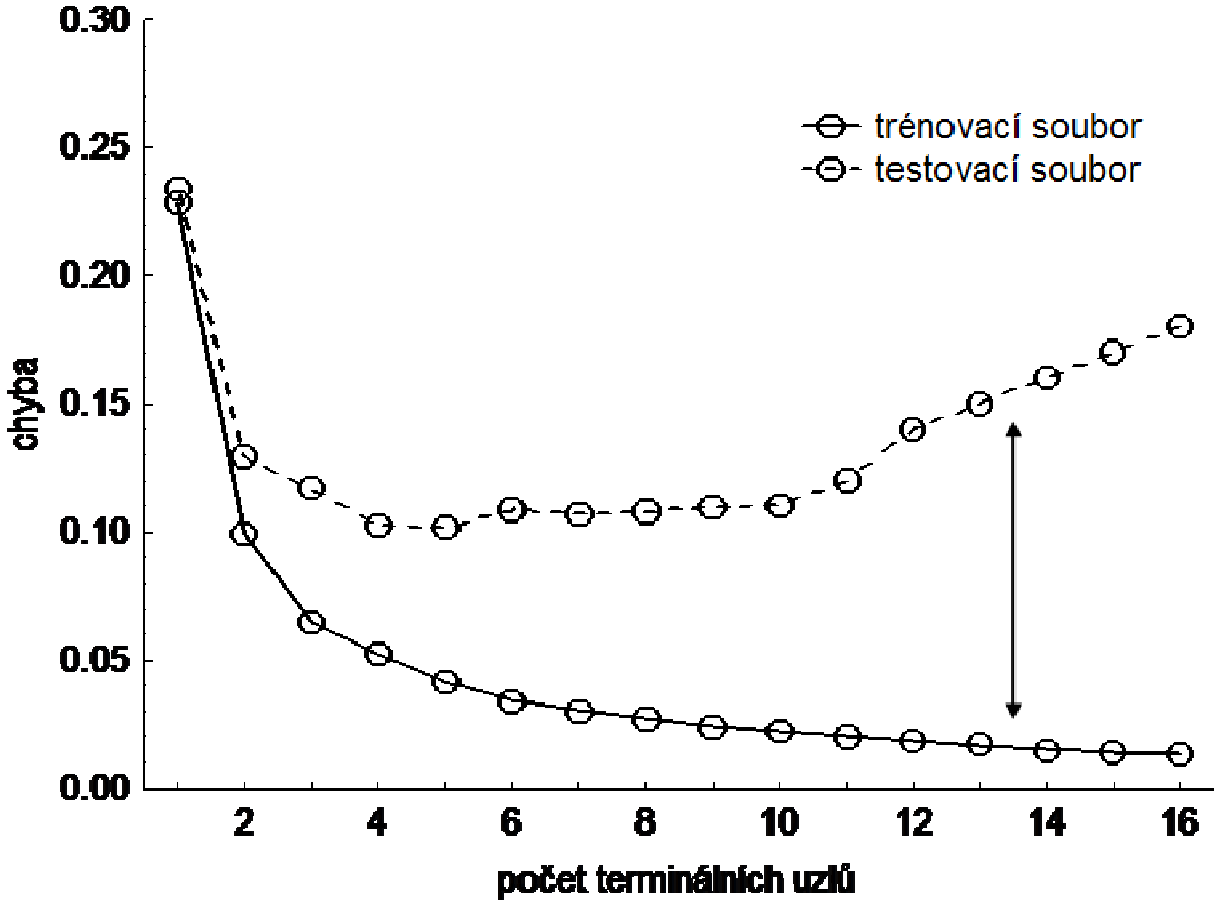
	uzel	n_A	n_B	p_A	p_B	p_t	$ME = 1 - \max(p_A, p_B)$	$p_t * ME$	
D1	t_1	150	50	3/4	1/4	1/2	1- 3/4 = 1/4	1/2 * 1/4	0,125
	t_2	50	150	1/4	3/4	1/2	1- 3/4 = 1/4	1/2 * 1/4	0,125
								celkový	**0,25**
D2	t_3	100	200	1/3	2/3	3/4	1- 2/3= 1/3	3/4 * 1/3	0,25
	t_4	100	0	1	0	1/4	1-1 = 0	1/4 * 0	0
								celkový	**0,25**

Výběr optimálního stromu

- strom bude mít velikost podle námi zvolených pravidel (nebo pravidel defaultně nastavených v softwaru), která mohou být subjektivní
- Jak tedy poznat, strom správné velikosti?
  - rozdělení souboru na trénovací a testovací
  - na trénovacím souboru se strom učí a roste
  - testovací soubor není při tvorbě stromu vůbec použit a slouží pouze k jeho otestování
- **nedoučený** (*underfitting*) strom → je příliš jednoduchý a chyba na testovacím i trénovací souboru bude velká
- **přetrénovaný** (*overfitting*) strom → je zbytečně složitý, trénovací chyba je většinou malá, ale testovací velká

!Je tedy třeba najít vhodný kompromis!

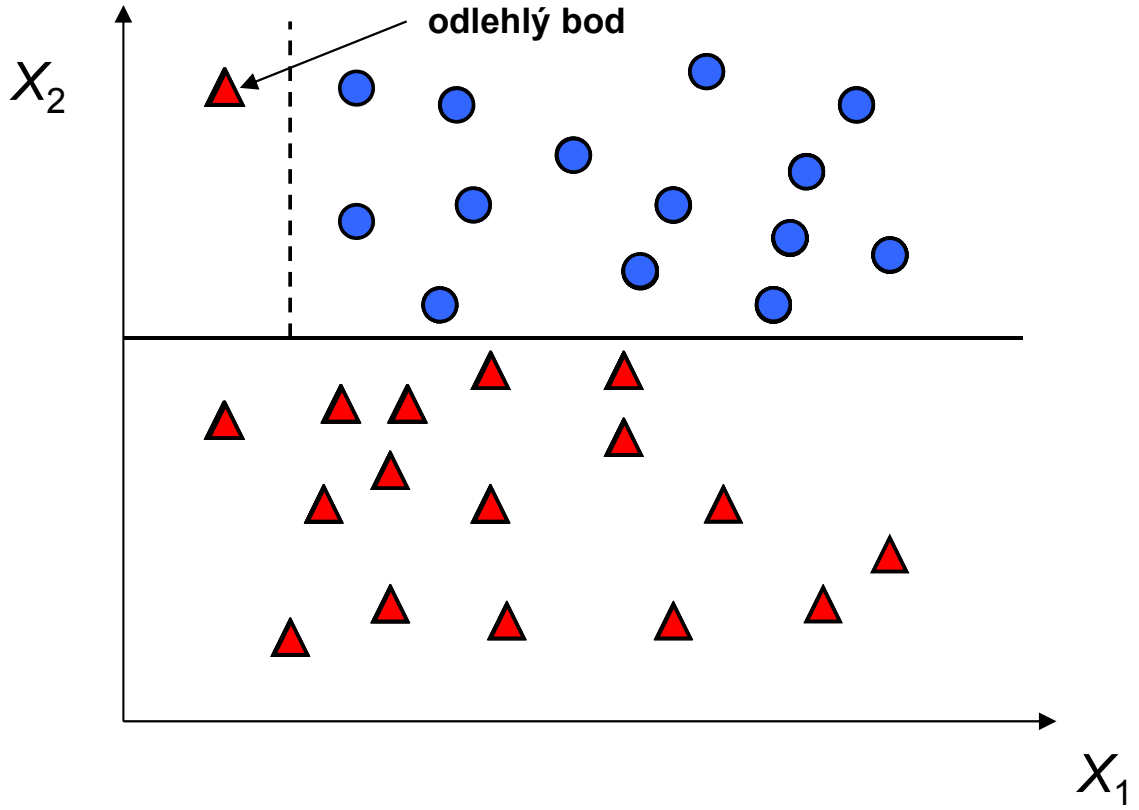
Rozdíl ve velikosti chyby mezi testovacím a trénovacím souborem při různé velikosti stromu, dané počtem terminálních uzlů



Nejprve byla spočítána chyba (procento chybně zaklasifikovaných pozorování) na testovacím a trénovacím souboru pro strom s 16 terminálními uzly. Postupně bylo vždy zpětně odstraněno poslední rozdělení uzlů, čímž se snížil počet terminálních uzlů o jedna. Pro takto zmenšený strom byla opět spočítána chyba pro oba soubory. Takto se postupně strom zmenšoval, až zbyl pouze jeden uzel –kořen stromu.

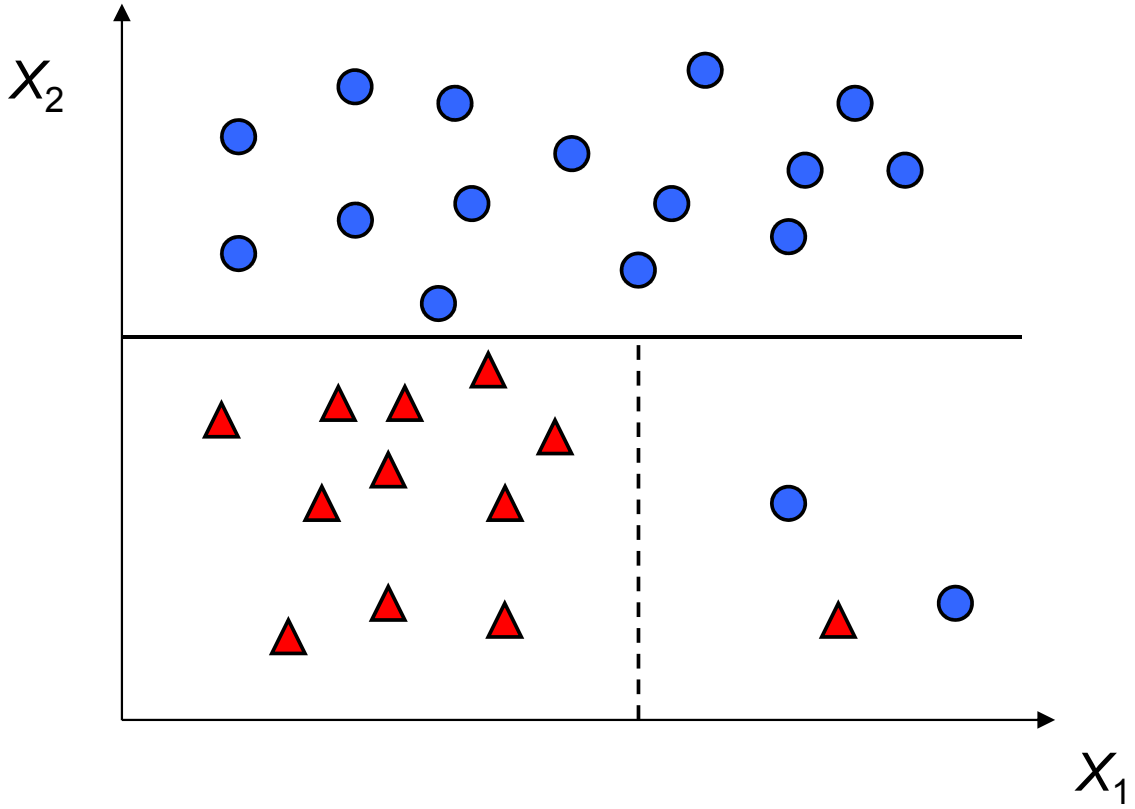
Příklad přetrénování stromu I

- kvůli odlehlé hodnotě



Příklad přetrénování stromu II

- z důvodu nedostatečného počtu trénovacích dat



Velikost stromu

Příliš velký strom

- Může být „přeučený“, tj. může být příliš specializovaný na datový soubor, který se použil pro jeho konstrukci.
- Pokud ho použijeme pro klasifikaci „neznámých“ případu, nemusí být příliš úspěšný.
- Neplatí tedy, že čím je strom větší, tím je lepší.
- Dobře naučený strom nepopisuje každý konkrétní případ, spíše by měl popisovat obecnější závislosti, které se v datech vyskytují.

Příliš malý strom

- Nemusí postihnout strukturu dat

Prořezávání stromu

- parametrem, který určuje složitost stromu, je jeho velikost.
- U CART začínáme s „přerostlým“, příliš detailně větveným stromem. Tento strom následně redukujeme pomocí některé z metod
  - **Prořezávání (*pruning*)**
  - **Zmenšování, scvrkávání se (*shrinking*)** - metoda pro regresní strom

K určení optimální velikosti stromu → kritérium složitosti stromu (*cost-complexity criterium*)

Kritérium složitosti stromu

Mějme strom T_0. Prořezáním určitého počtu koncových uzlů dostaneme strom T_1.

Cena jednoduššího stromu (*cost-complexity criterium):*

$$C_\alpha(T_1) = DT_1 + \alpha|T_1|,$$

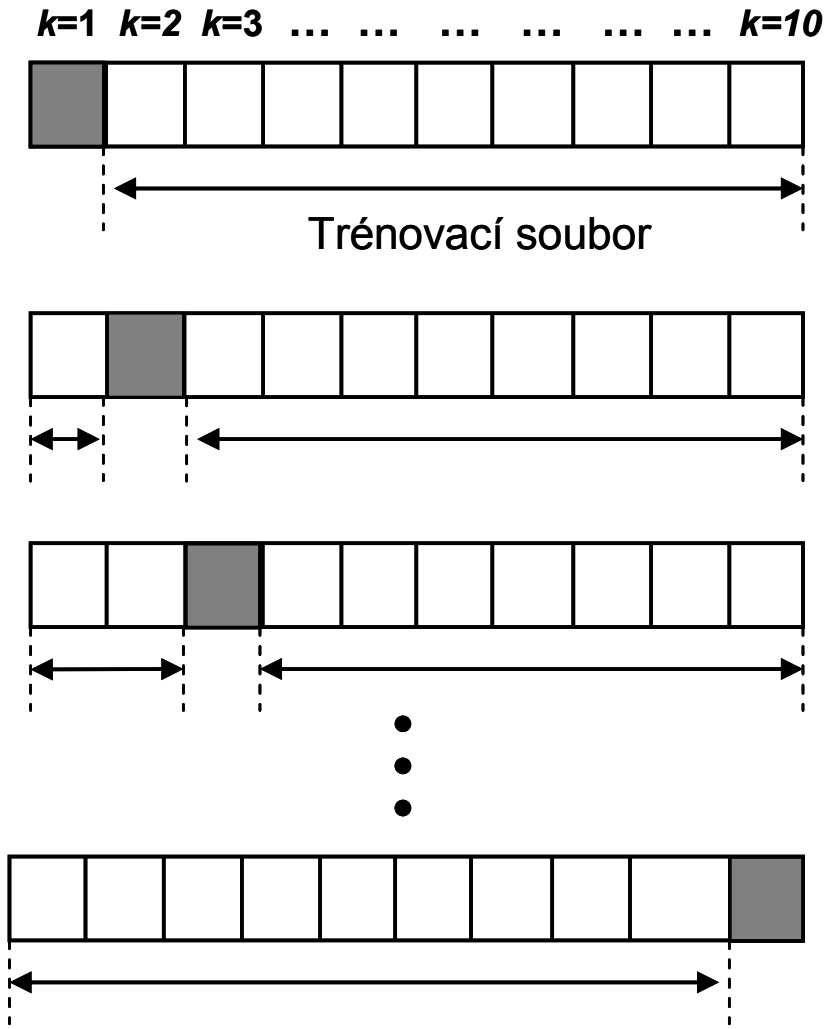
kde $|T_1|$ je počet terminálních uzlů stromu a DT_1 je deviance stromu. Parametr $\alpha \geq 0$ vyjadřuje kompromis mezi velikostí stromu a jeho vyčerpanou variabilitou. Hledáme tedy, pro každé α, takový strom , který minimalizuje $C\alpha(T)$.

K určení odhadu α se používá krosvalidace

Křížové ověřování (krosvalidace)

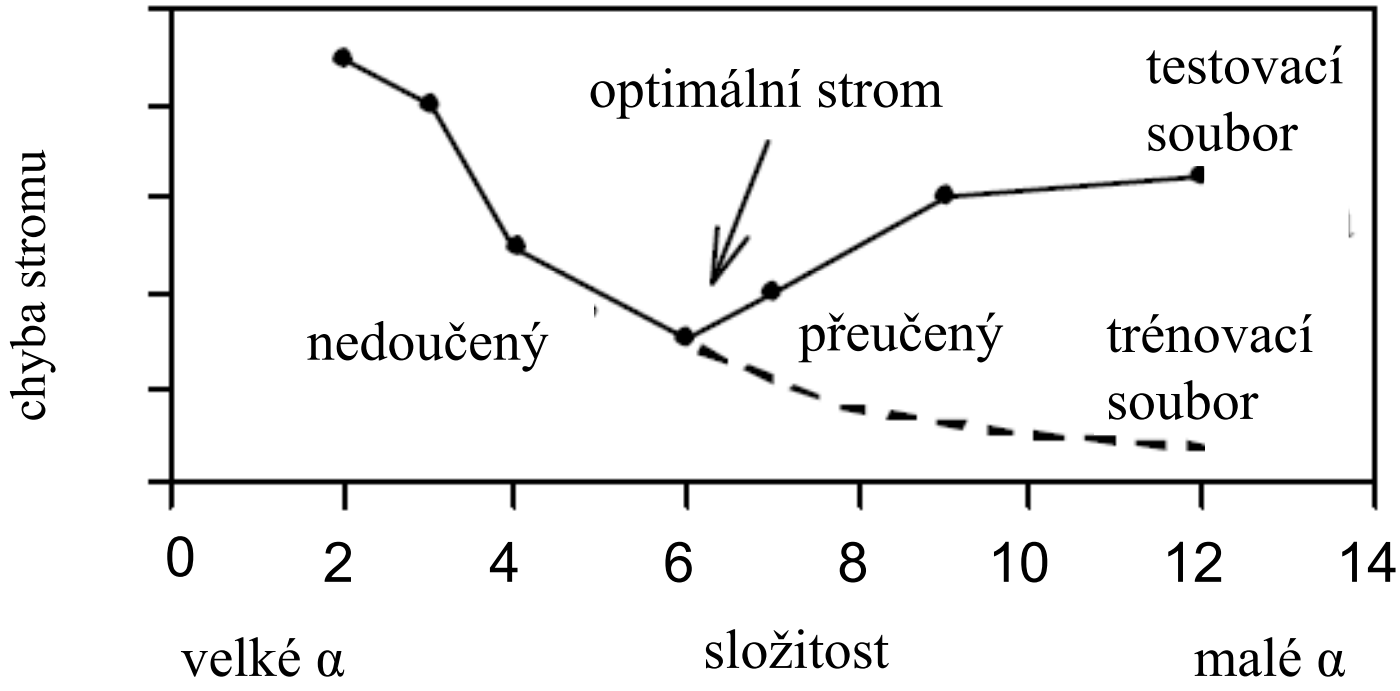
- Křížová validace patří mezi validační techniky
- Pozorování jsou rozdělena do *k* nezávislých podsouborů
- Jeden podsoubor se použije pro testování (pozorování nejsou použita při tvorbě modelu) všech ostatních *k*-1 skupin pro tvorbu modelu → je tedy vytvořeno *k* modelů otestovaných na *k* testovacích souborech
- Z výsledků testovacích souborů můžeme určit stabilitu metody (spočítat např. průměr a směrodatnou odchylku přesnosti na testovacím souboru) a její predikční schopnost.
- Stromy jsou obecně velmi nestabilní metody → i malá změna v datech může způsobit změny v rozhodovacích pravidlech a můžeme získat odlišný strom s jinou přesností.
  - Jak velká je tato variabilita, zjistíme z rozsahu hodnot přesnosti stromu pro jednotlivé testovací soubory.
- Výhoda křížové validace spočívá v použití nezávislého datového souboru pro testování - každé pozorování je pro testování použito právě jedenkrát

Křížové ověřování (krosvalidace)



Výběr optimálního stromu

- Pomocí křížové validace vybereme takové α, aby měl strom co největší přesnost, ale zároveň byl rozdíl v chybě mezi testovacím a trénovacím souborem co nejmenší



Měření přesnosti stromu

- Označme $e(t)$ chybu na trénovacím souboru (*re-substitution errors*) a $e'(t)$ chybu na testovacím souboru (*generalization errors*).
- Při použití pouze trénovacího souboru lze získat dva odhady celkové chyby stromu.
  - **optimistický odhad**, kdy předpokládáme, že chyba trénovacího souboru se rovná chybě na testovacím souboru $e'(t) = e(t)$
  - **pesimistický odhad**, kdy je pro každý terminální uzel $e'(t) = (e(t)+0,5)$
    - Celková chyba je tedy: $e'(T) = e(T) + N \times 0,5$,

      kde N je počet terminálních uzlů

Měření přesnosti stromu II

- Mějme soubor obsahující 100 měření. Pro strom s 20 terminálními uzly a 10 chybně zařazenými pozorováními z trénovacího souboru je:
  - optimistický odhad chyby = 10/100 = 10%
  - pesimistický odhad chyb = (10 + 20×0,5)/100 = 20%.
- Chyba na trénovacím souboru však není dobrým ukazatelem, jak dobře bude strom klasifikovat/predikovat nová data.
- Proto se k odhadu celkové (obecné) chyby stromu používá převážně testovací soubor.

Matice záměn (confusion matrix)

	pozorované				
	ano	**ne**	**celkem**		
predikované					
ano	**a** (TP)	**b** (FP, chyba I. druhu)	**a+b**	→	PPV
ne	**c** (FN, chyba II. druhu)	**d** (TN)	**c+d**	→	NPV
celkem	**a+c**	**b+d**	**n**		
	↓	↓			
	Sensitivita, FNR	Specificita, FPR			

a = pravdivě pozitivní (TP- true positive)
b = falešně pozitivní (FP - false positive)
c = falešně negativní (FN - false negative)
d = pravdivě negativní (TN - true negative)

Senzitivita-podíl (procento) správně zařazených pozitivních případů, ze všech případů, které byly předpovězeny jako pozitivní (např. Procento lokalit kde se taxon vyskytl, ze všech lokalit predikovaných jako výskyty)
specificita - podíl správně zařazených negativních případů, ze všech případů předpovězených jako negativní (např. procento lokalit kde se taxon nevyskytl, ze všech lokalit predikovaných jako nevýskyt).

Měření přesnosti klasifikačního stromu

- Celková správnost, (*Overall accuracy, Correct classification rate)*:

$$OA=(a+d)/n$$

- Klasifikační chyba:

$$MR = (b+c)/n$$

- Cohenovo kappa:

$$Kp= (OA-EA)/(1-EA),$$

kde $EA = ((a+c)(a+b)+(b+d)(c+d))/n^2$

- Na testovacím souboru, použití krosvalidačních technik pro zjištění obecnosti a stability stromu

Měření přesnosti klasifikačního stromu II

- Tato měření však nezohledňují různou velikost skupin ani rozdílnost oproti náhodnému výsledku, a proto může snadno dojít k nadhodnocení nebo naopak podhodnocení kvality modelu.

- Mějme příklad klasifikačního stromu pro závisle proměnnou se dvěma kategoriemi a počtem pozorování v jednotlivých kategoriích A = 100 a B = 10. Počet správně klasifikovaných pozorování v jednotlivých kategoriích je následující A = 100 a B = 0.

$$OA = 100/110=0,91$$

- Procento správně klasifikovaných pozorování by v tomto případě bylo zhruba 91% → takový strom nám není k užitku, protože nedokázal kategorie odlišit a všechna pozorování v kategorii *C* klasifikoval jako kategorii *A*.

Měření přesnosti klasifikačního stromu III

- Korekci na velikost kategorií lze však provést jednoduchou úpravou:

$$OA_{kateg} = \frac{1}{J}\sum_{c=1}^{J}\frac{n_{pc}}{n_c}$$

kde J je celkový počet kategorií, n_{pc} je počet správně klasifikovaných pozorování v kategorii c a nc je počet všech pozorování v kategorii c.

- Pro náš příklad se pak celková adjustovaná správnost stromu rovná:

$$\frac{1}{2}\left(\frac{100}{100}+\frac{0}{10}\right)=0{,}5$$

- Celková správnost se používá především pro srovnání s ostatními klasifikačními metodami nebo pro výběr vhodného stromu, v praxi nás však častěji zajímá procento správně klasifikovaných pozorování pro každou kategorii.

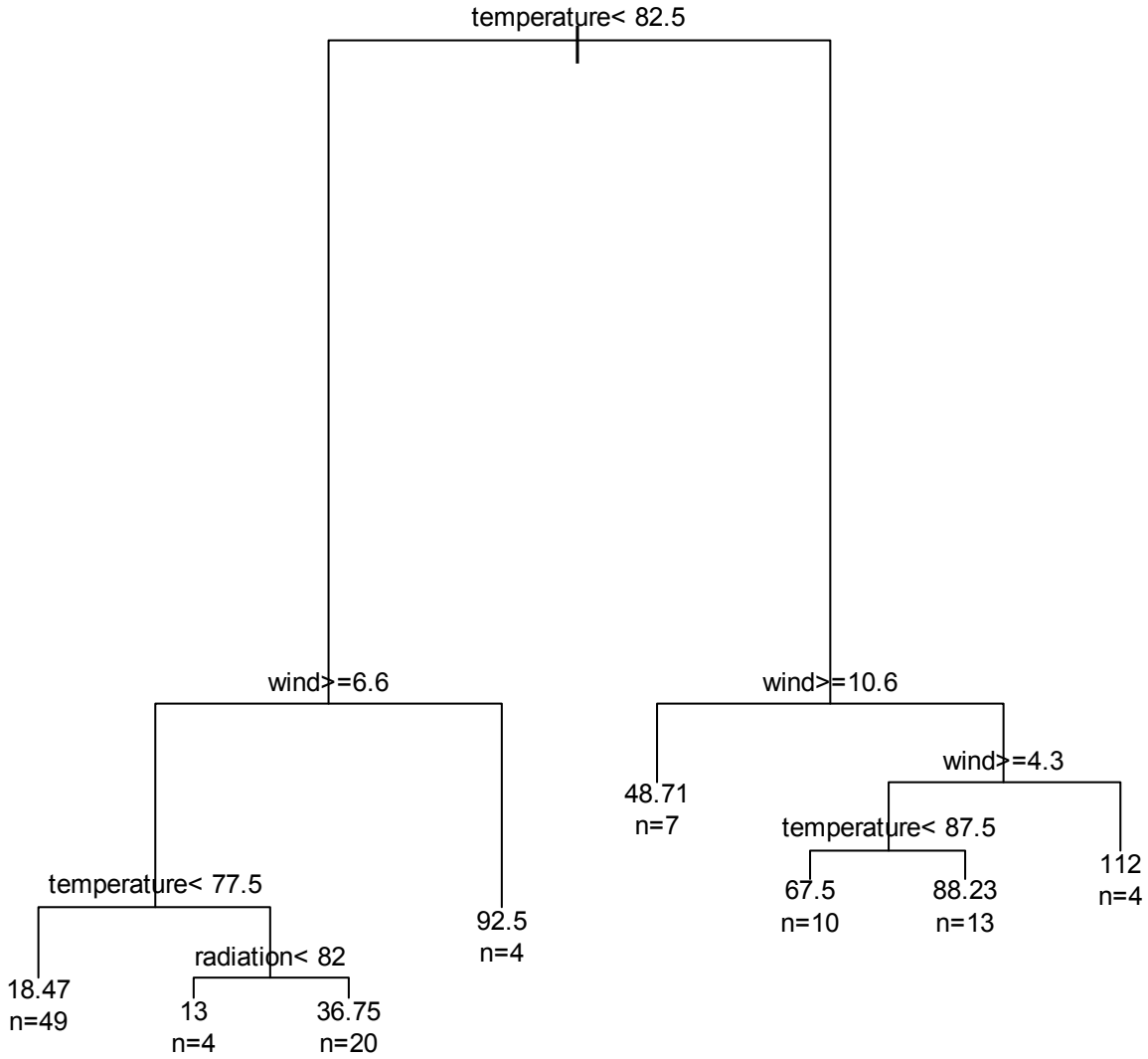
Určení přesnosti regresního stromu

- U regresního stromu je přesnost, určována stejně jako v lineární regresi, pomocí koeficientu determinace R^2.
- Koeficient determinace je obecně definován jako podíl variability závislé proměnné *Y*, vysvětlené modelem k celkové variabilitě proměnné *Y*.
- V našem případě jde o variabilitu vysvětlenou stromem

$$R^2 = \frac{variabilita\_vysvetlena\_modelem}{celkova\_variabilita\_Y} = 1 - \frac{residualni\_variabilita}{celkova\_variabilita\_Y} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2},$$

- kde $\hat{y}_i$ = je průměr v příslušných terminálních uzlech a odchylka od průměru uzlu *t* je spočítána vždy pro pozorování y_i zařazené do tohoto terminálního uzlu.
- Koeficient determinace nabývá hodnot od 0 do 1. Při hodnotě $R^2 = 1$ jsme vysvětlili veškerou variabilitu pomocí stromu a predikované hodnoty $\hat{y}_i$ se shodují s pozorovanými hodnotami y_i.
- Je opět možné spočítat chybu regresního stromu pro trénovací soubor $e'(t) = 1 - R^2_{tren}$ a testovací soubor $e(t) = 1 - R^2_{test}$

Příklad – ozón



Primární, zástupné a kompetitivní proměnné

- ***Primární proměnná*** dosahuje nejlepšího dělení daného uzlu a je použita jako pravidlo ve stromě
- Může se stát, že proměnná, která je téměř stejně vhodná (kriteriální statistika má podobnou hodnotu) jako vybraná primární proměnná, zůstane skrytá, i když může mít větší interpretační hodnotu → takovéto proměnné se nazývají zástupné (*surrogates*) a kompetitivní proměnné.
- ***Zástupné proměnné*** nesou podobnou informaci jako primární proměnná a většinou jsou s ní korelované. Pro každý uzel lze zjistit, nakolik rozdělují pozorování v dceřiných uzlech stejně jako primární proměnná.
- ***Kompetitivní proměnná*** rozděluje daný uzel odlišně než primární

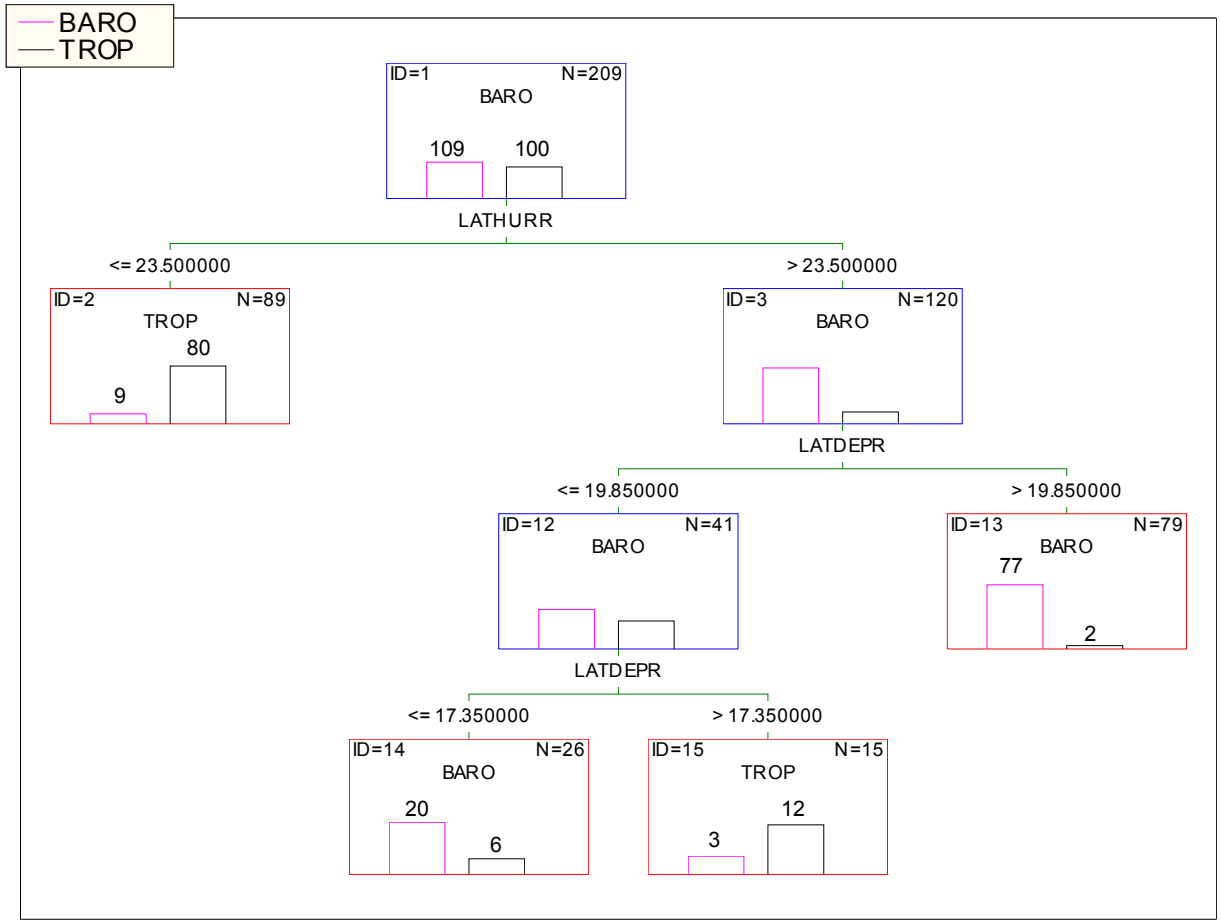
- Na základě hodnot kriteriální statistiky se tak v případě absence primární proměnné rozdělí uzel podle kompetitivní nebo zástupné proměnné.
- Je tedy vybrán jiný prediktor s další nejlepší hodnotou kriteriální statistiky.
- Velký význam pro interpretaci

Určení primární, kompetitivní a zástupné proměnné při rozdělení pozorování kategorií A, B, C do dvou terminálních uzlů

A = 100, B = 100, C = 100

proměnná *X*	kategorie	uzel 1	uzel 2
primární	A	**90**	10
	B	**90**	10
	C	20	**80**
zástupná	A	**80**	20
	B	**85**	15
	C	25	**75**
kompetitivní	A	80	**20**
	B	**20**	80
	C	**10**	90

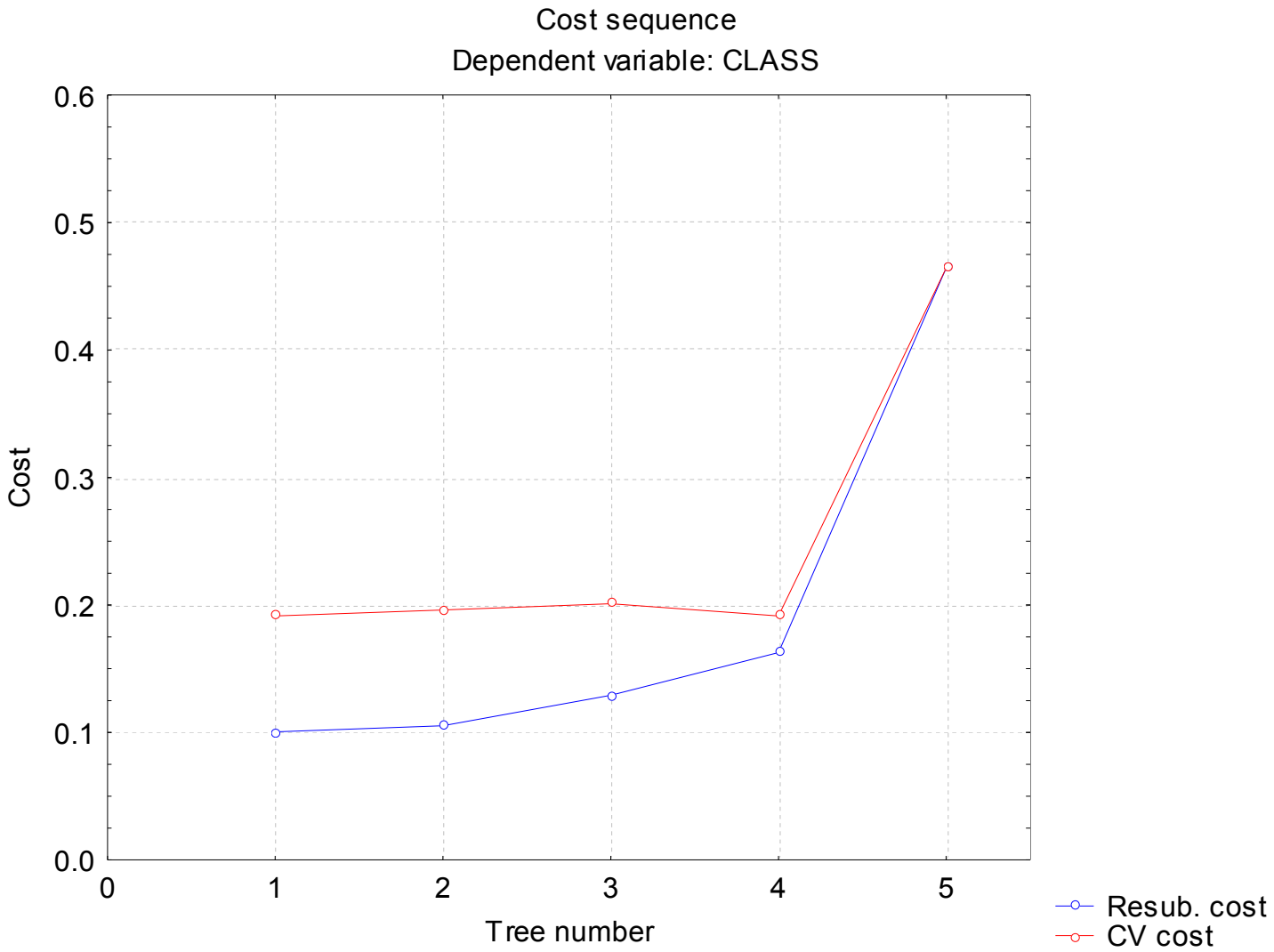
Příklad hurikány



Co vše můžeme zjistit ze stromu……

Jak interpretovat strom ?
Jaká je celková přesnost stromu ?
Která ze dvou skupin je lépe klasifikována?
Které parametry jsou významné ?

Příklad hurikány



Má strom správnou velikost?

Výhody stromů

- ☺ Snadné grafické znázornění – jednoduchá interpretace
- ☺ Neklade žádné podmínky na typ rozdělení
- ☺ Algoritmy tvorby stromu jsou odolné vůči odlehlým hodnotám
- ☺ Možno použít korelované proměnné
- ☺ Prediktory mohou být všech typů
- ☺ Výsledky přesnosti stromu lze snadno porovnat s výsledky jiných modelů
- ☺ rychlá metoda při klasifikaci nových případů
- ☺ Metoda je vhodná pro klasifikaci i regresi (pro regresi s jistými omezeními)

Klasifikační (rozhodovací) strom
Nevýhody

☹ Nestabilita - tvar stromu velmi závisí na datech, malá změna v datech způsobí změny v rozhodovacích pravidlech uvnitř uzlů

+ změna výsledných klasifikací/predikcí

- Vzhledem k nestabilitě je nutná opatrnost při interpretaci.
- Řešení: např. Bagging – kombinace většího počtu stromů, aby se minimalizovala jejich variabilita (bude vysvětleno později viz. klasifikační lesy)

☹ měření přesnosti stromu (*accuracy*) je výrazně závislé na krosvalidačním mechanizmu, selekčních kritériích a výběru mechanizmu pro minimalizaci chyby stromu

☹ nevhodné pro malý počet vzorků a velký počet tříd

☹ vytváření stromů vyžaduje zkušenosti

Algoritmy učení

Je celá řada algoritmů pro růst stromu
obecně nelze říci, který z algoritmů je lepší, záleží na řešeném problému
výsledkem je strom, který se však liší obsahem uzlů i jejich počtem

- ID3 (Quinlan 1979)
- CHAID - Chi-squared Automatic Interaction Detector (Kass,1980)
- CART (Breiman et al. 1984)
- Assistant (Cestnik et al. 1987)
- MARS - Multivariate Adaptive Regression Splines (Friedman,1991)
- RETIS (Karalič 1992) – pro regresní stromy
- C4.5 (Quinlan 1993)
- QUEST - Quick, Unbiased and Efficient Statistical Tree (Loh & Shih, 1997)
- C5 (Quinlan 1997)
- PRIM - Patient Rule Induction Method (Friedman & Fisher, 1999)
- Stromy ve Wece (Frank 2000)
- Stromy v Orange (Demšar, Zupan 2000)