Pokročilé neparametrické metody
Rozhodovací stromy

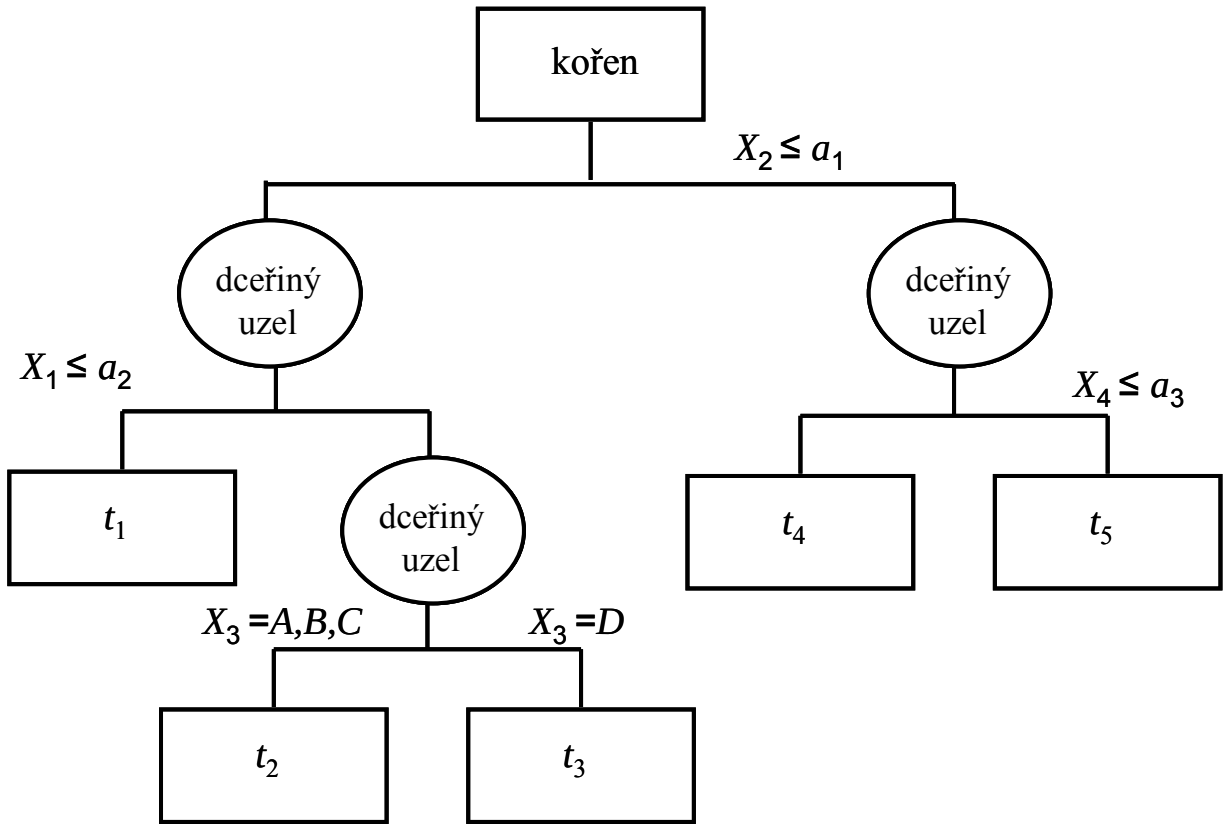
Klára Komprdová

INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Stromy typu CART

Strom typu CART

- Breiman et al. 1984
- vhodné pro kategoriální i regresní úlohy
- rostou na základě rekurzivního binárního dělení



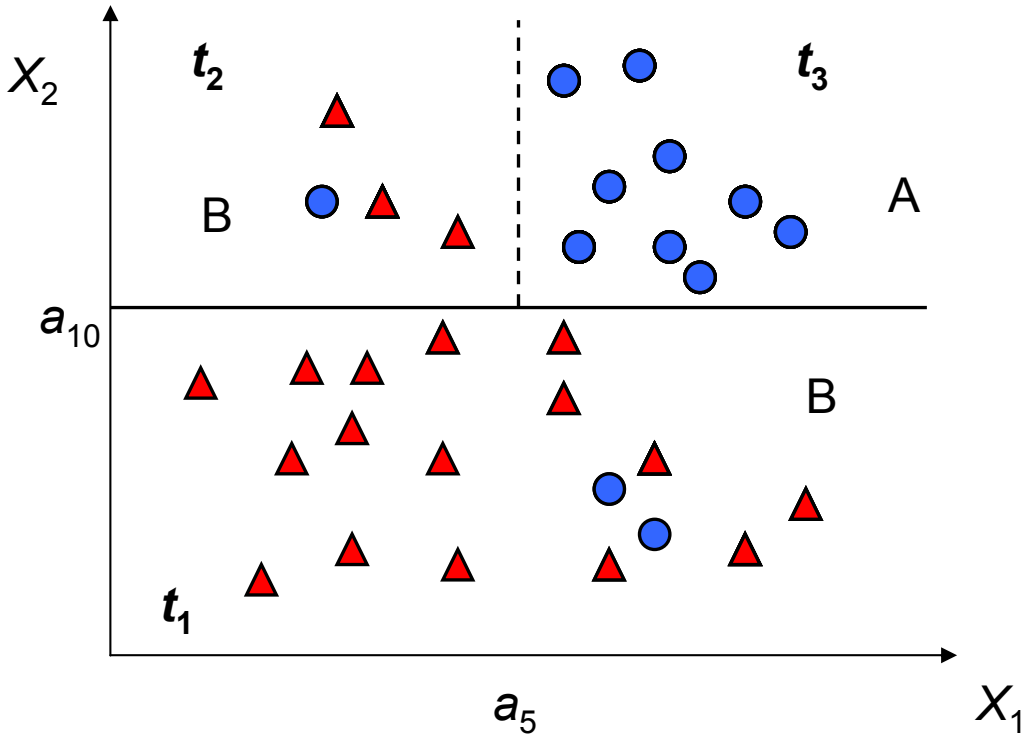
Jak roste strom CART?

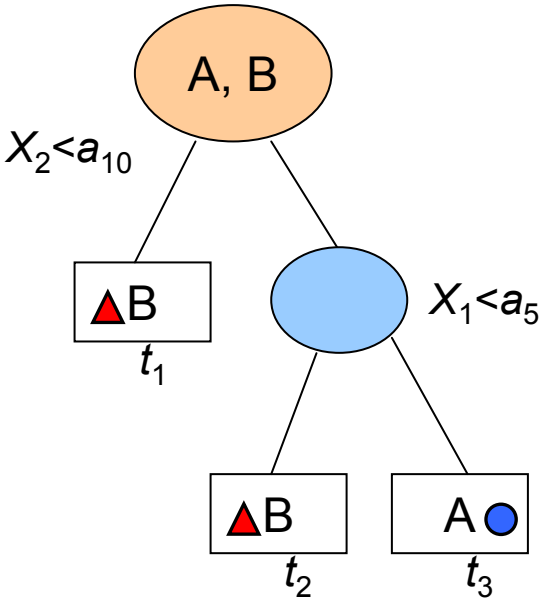
- pozorování rozdělena do dvou dceřiných uzlů, na základě hodnoty *a* prediktoru *X*, které jsou dále děleny opět binárně na další uzly

- hodnoty vysvětlujících proměnných, použité při větvení, rozdělují daný prostor na sadu pravoúhelníků a pak pro každý z nich fitují jednoduchý model

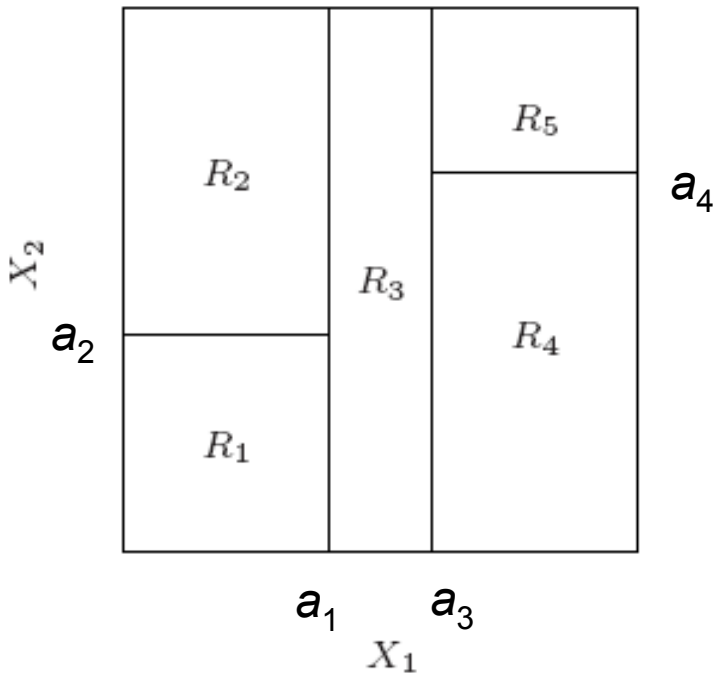
Grafické znázornění stromu CART

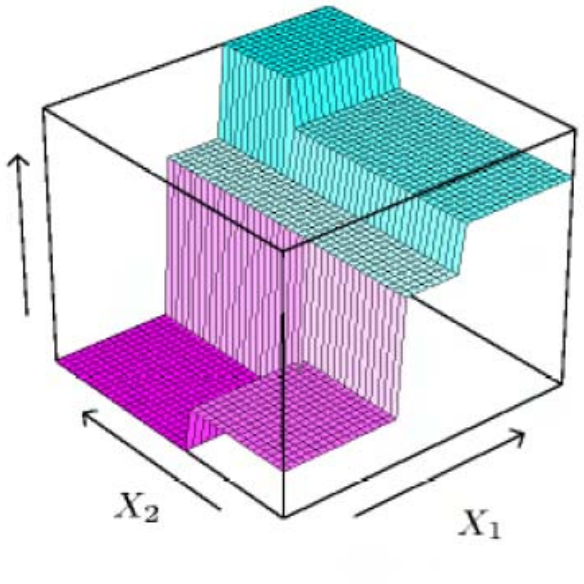
- rozdělení pozorování do kategorií A a B závisle proměnné Y s použitím dvou spojitých prediktorů X_1, X_2





Jak na to?





(Tibshirani et. al, 2001).

Jak najít správné rozdělení?

- existuje mnoho **algoritmů**, jak vybírat proměnné a hranice podle kterých bude probíhat dělení datového souboru
- hlavní princip: snažíme se najít takové rozdělení závisle proměnné *Y* prediktorem *X*, aby hodnoty proměnné *Y* byly uvnitř uzlu co nejhomogennější a zároveň mezi uzly co nejrozdílnější
- který prediktor (a jeho hodnota) nám zajistí nejlepší rozdělení zjistíme pomocí tzv. **kriteriální statistiky** (*spliting criterium*), která určuje homogenitu uzlu
- existuje několik měření kriteriálních statistik, které se navíc liší podle toho, zda se jedná o klasifikační nebo regresní strom
- nejčastěji používanými měřeními pro stromy typu CART: Kritérium minima kvadratické chyby , Gini index, Entropie a klasifikační chyba

Kriteriální statistika pro regresní stromy

- Předpokládejme, že máme strom rozdělený do určitého počtu terminálních uzlů a odpověď závisle proměnné modelujeme jako konstantu pro každý terminální uzel.
- Pokud použijeme kritérium, které minimalizuje střední kvadratickou chybu, nejlepším odhadem bude průměr.
- Kritérium Kritérium minima kvadratické chyby (*Least Square Deviation LSD*):

$$\bar{y}_t = \frac{1}{N_t} \sum y_{i(t)}$$

$$Q_t(T) = \frac{1}{N_t} \sum_{i=1}^{N_t} \left(y_{i(t)} - \bar{y}_t \right)^2$$

kde N_t je počet pozorování v uzlu t a $y_{i(t)}$ jsou hodnoty závisle proměnné v uzlu t

Kriteriální statistika pro klasifikační stromy

- Gini index: $$GI = \sum_{c=1}^{J} p_{tc}\left(1 - p_{tc}\right) = 1 - \sum_{c=1}^{J} p_{tc}^2$$

- Entropie: $$H = -\sum_{c=1}^{J} p_{tc} \log_2 p_{tc}$$

- Klasifikační chyba: $$ME = 1 - \max\left\{p_{tc}\right\}$$

kde p_{tc} je podíl pozorování y_i s kategorií c v uzlu t z celkového počtu všech pozorování y_i v tomto uzlu neboli pravděpodobnost kategorie c v uzlu t.

- Gini index – nejčastěji používané měření pro klasifikační stromy - hodnota Giny indexu se rovná nule, pokud je v konečném uzlu pouze jediná třída a dosahuje maxima, pokud je v konečném uzlu v každé třídě stejný počet pozorování.

- *Impurity measurement*

Celkové hodnoty indexů pro rozdělení

- Ve chvíli, kdy dojde k rozdělení uzlu na dva dceřiné uzly, je *GI* spočítán pro každý dceřiný uzel.
- Hodnota *GI* indexů jednotlivých dceřiných uzlů je vážena velikostí dceřiného uzlu.
- GI_{celk} = součet *GI (i)* dceřiných uzlů, které jsou vynásobeny příslušným podílem pozorování v daném dceřiném uzlu z celkového počtu pozorování v původním mateřském uzlu.

$$GI_{celk} = \sum_{i=1}^{K} \frac{N_i}{N_t} GI(i)$$

kde K je počet dceřiných uzlů (v případě binárního stromu se $K = 2$), N_t je počet pozorování v mateřském uzlu t a N_i jsou počty v dceřiných uzlech.

Stejně pro další indexy…Entropie

- Celková entropie: $$H_{celk} = \sum_{i=1}^{k} \frac{N_i}{N_t} H(i)$$

- Entropie dosahuje maxima, pokud jsou jednotlivé kategorie proměnné *Y* rovnoměrně zastoupeny v uzlech a minima pokud pozorování v uzlu náležejí pouze do jediné kategorie.
- Entropie je často používána v algoritmu C4.5.

- *GAIN* (*information gain*, informační zisk) a měří pokles v entropii.

$$GAIN_{celk} = H - \left(\sum_{i=1}^{k} \frac{N_i}{N_t} H(i) \right)$$

Klasifikační chyba

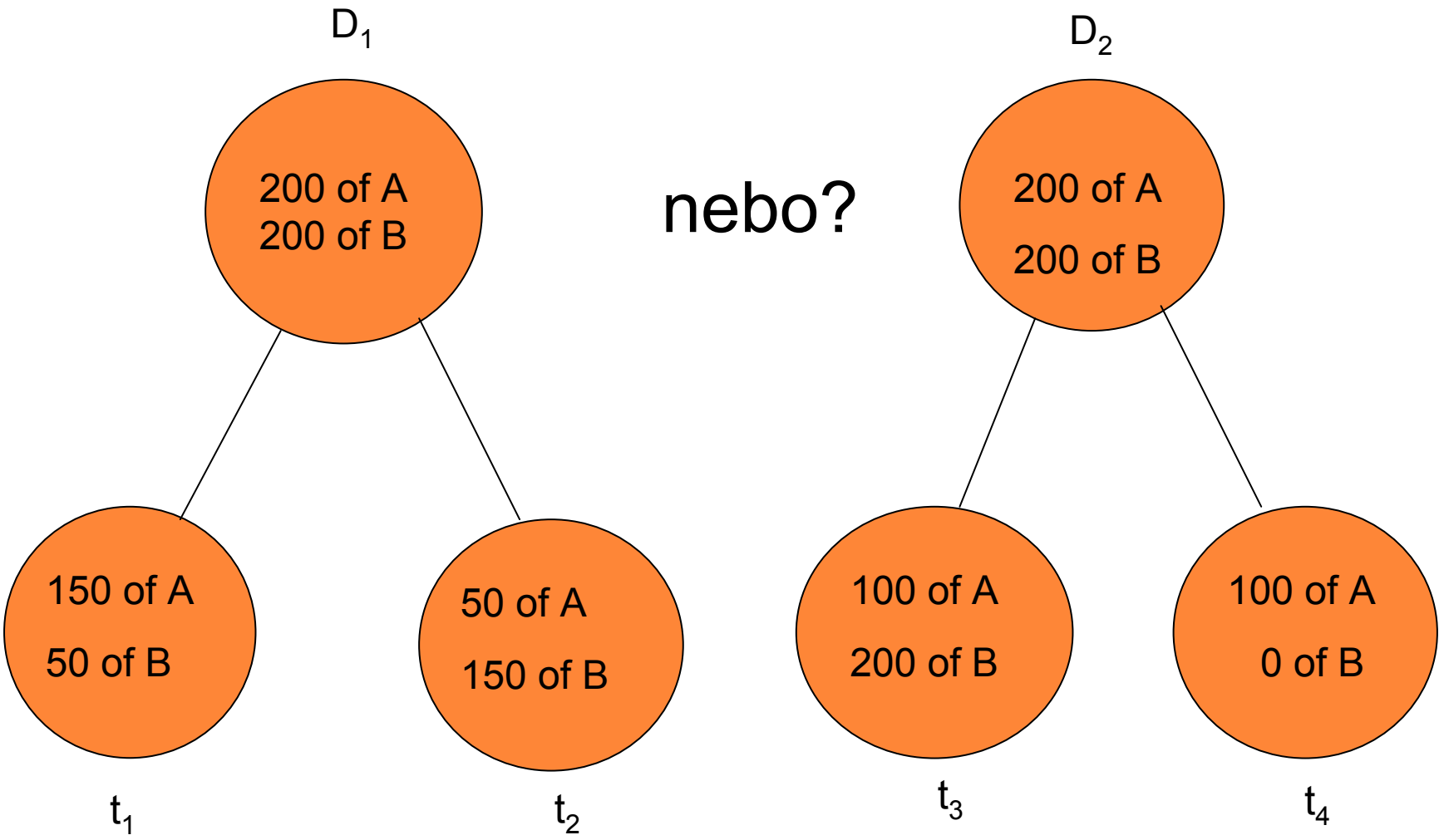
$$ME_{celk} = \sum_{i=1}^{k} \frac{N_i}{N_t} \, ME(i)$$

- Celková klasifikační chyba pro dané dělení = vážený součet *ME* v dceřiných uzlech.
- *ME* je podíl chybně klasifikovaných pozorování
- 1 - *ME* je celková přesnost stromu = podíl správně klasifikovaných pozorování

- Klasifikační chyba je obvykle používána k finálnímu měření přesnosti klasifikačního stromu, proto je logické její použití jako kriteriální statistiky

- preferovány jiné indexy → Entropie a Gini index jsou mnohem více citlivé na změny v pravděpodobnostech uzlů než ME

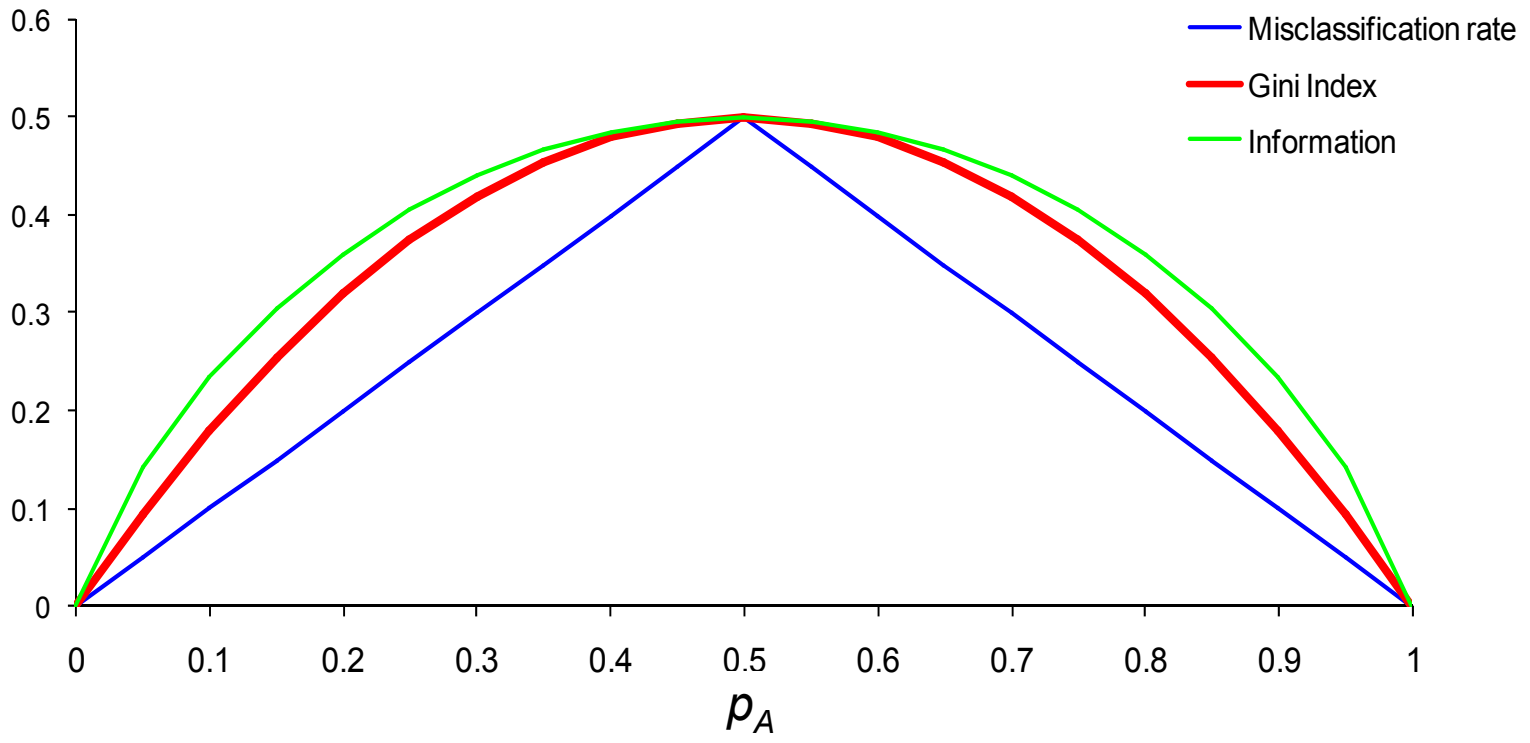
Příklad


D1
200 of A
200 of B
150 of A
50 of B
50 of A
150 of B
t1
t2
nebo?
D2
200 of A
200 of B
100 of A
200 of B
100 of A
0 of B
t3
t4

	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							
	t_2							
								celkový
D2	t_3							
	t_4							
								celkový

	uzel	n_A	n_B	p_A	p_B	p_t	$ME = 1 - \max(p_A, p_B)$	$p_t * ME$
D1	t_1							
	t_2							
								celkový
D2	t_3							
	t_4							
								celkový

Obecný průběh kriteriálních statistik pro rozdělení do dvou kategorií *A* a *B* závisle proměnné *Y* jako funkce podílu první kategorie p_A.



Všechny kriteriální statistiky dosahují svého maxima, pokud je kategorie rovnoměrně rozmístěna mezi uzly (p_A = 0,5) a minima, pokud je zastoupena pouze jedna kategorie (p_A = 1 nebo p_A = 0 p_B = 1).

(Tibshirani et. al, 2001).

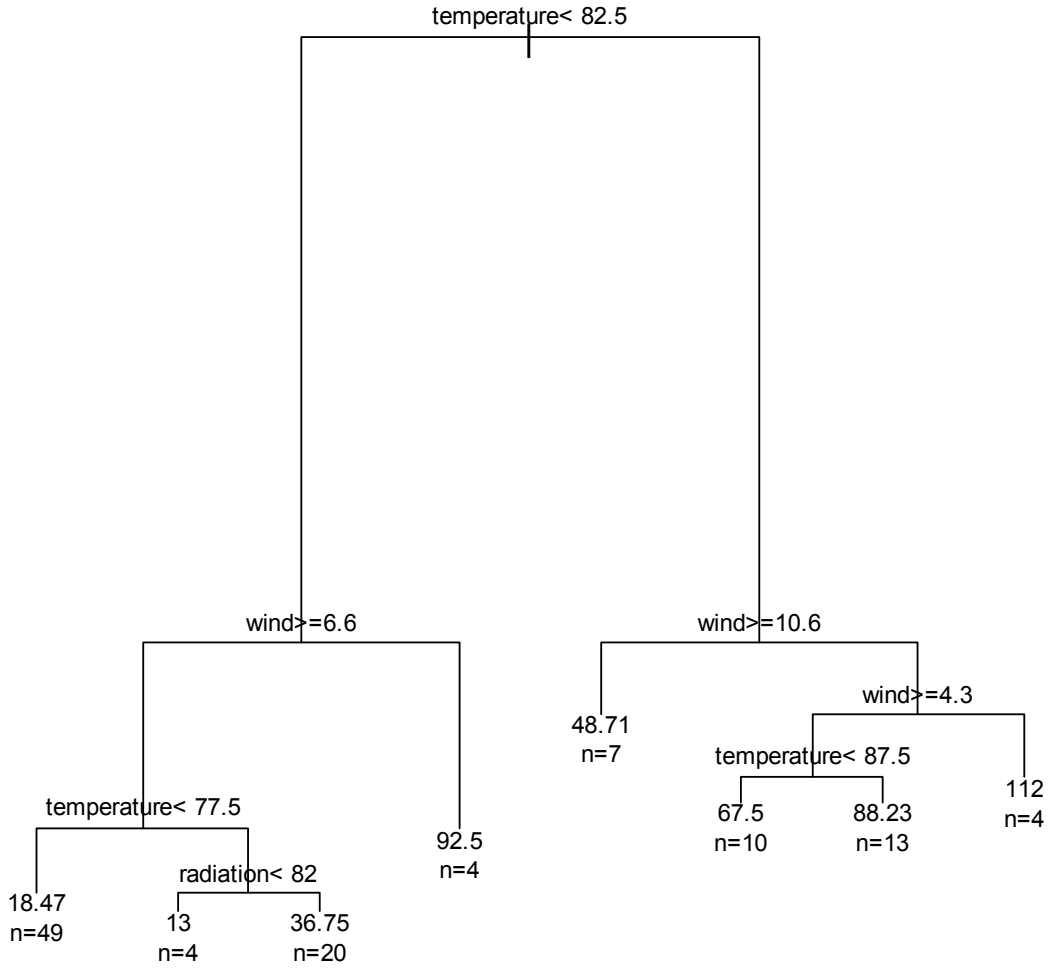
Výsledná hodnota predikce – regresní strom

- Každému objektu z koncových listů je přiřazena hodnota, kterou vypočteme jako aritmetický průměr hodnot všech objektů v příslušném listu.

- Výsledný odhad hodnot závisle proměnné tak bude nabývat pouze t_n hodnot, kde t_n je počet terminálních uzlů

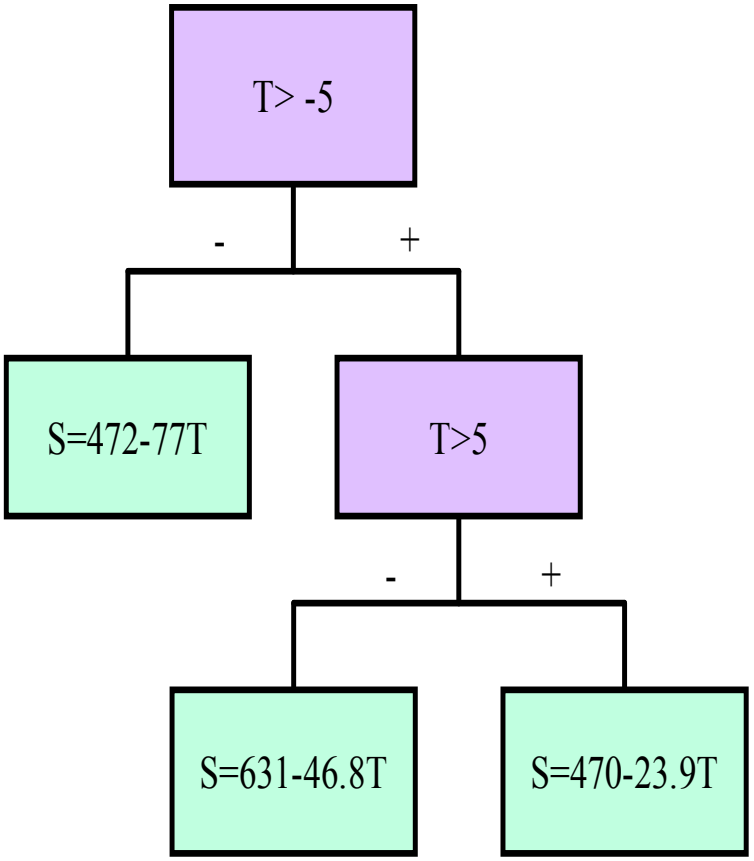
- Další možností je vytvořit pro jednotlivé listy regresní modely
  - Nemusí však být dostatečný počet dat v koncové uzlu
  - Výsledný vztah nelze popsat regresí (není zde závislost, vzorky v terminálním uzlu nesplňují předpoklady regrese)
  - Metoda začne nabývat na složitosti

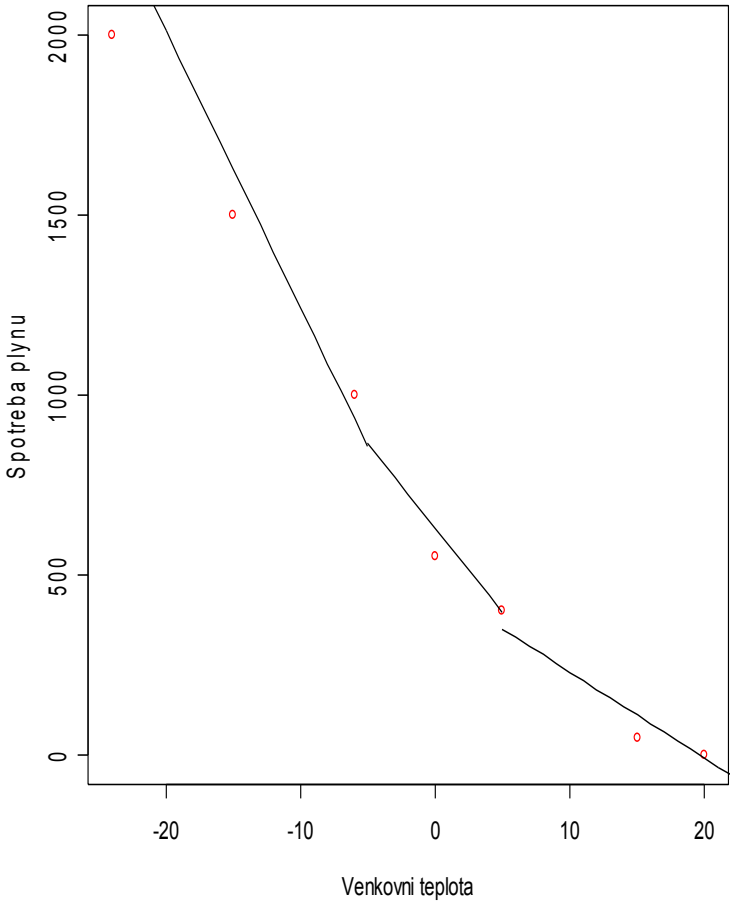
Příklad – ozón



Příklad: Ukázka regresního stromu

Závislost spotřeby plynu na venkovní teplotě





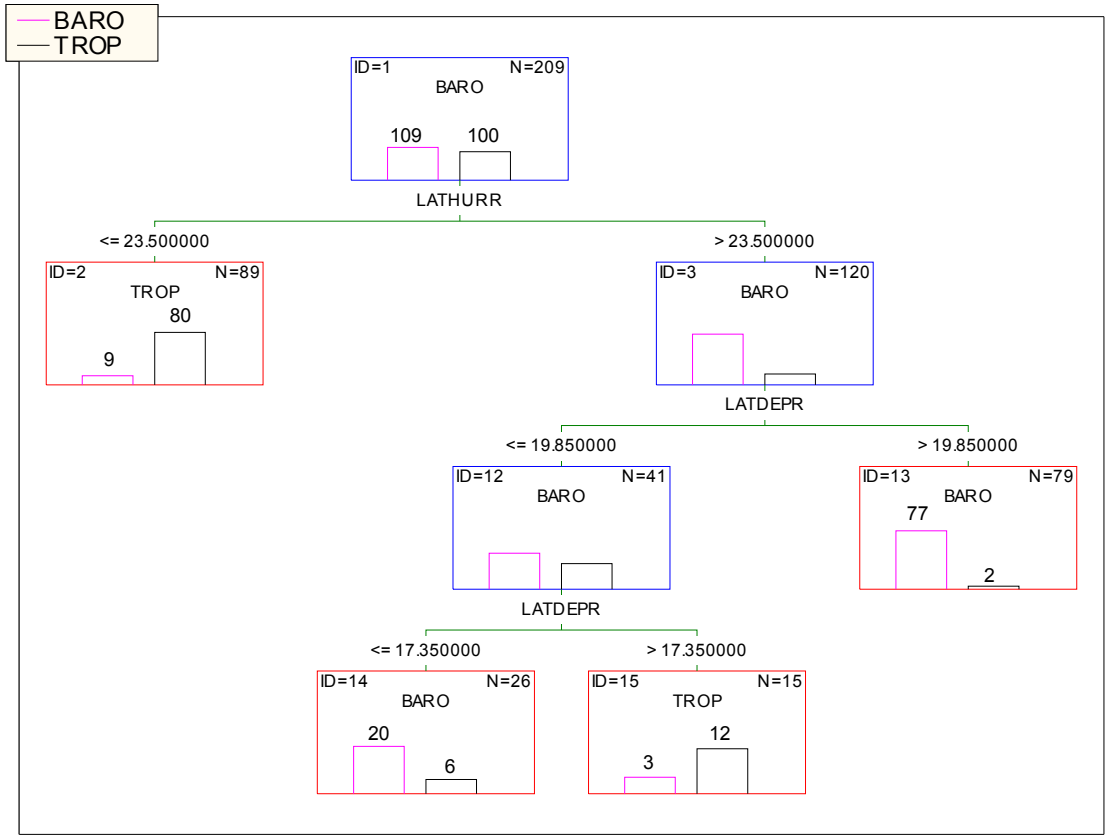
Přiřazení hodnoty terminálnímu uzlu

- klasifikační strom - každému uzlu, včetně kořenového, je přiřazena výsledná kategorie závisle proměnné
- výsledná kategorie - má v daném uzlu největší zastoupení
- nové pozorování je klasifikováno podle kategorie uzlu, do kterého je stromem zařazeno
- může se stát, že po rozdělení do dvou terminálních uzlů bude oběma uzlům přiřazena stejná kategorie, zejména je-li podíl kategorií proměnné *Y* nevyrovnaný → výhodu mají kategorie, které jsou u proměnné *Y* více zastoupeny
  - možnost použít vážení jednotlivých kategorií

Příklad hurikány

- Atlantické hurikány jsou klasifikovány podle ovlivnění tropickými (Trop) nebo baroklinickými (Baro) jevy.
- Tropická cyklóna při vývoji prochází třemi stádii: *tropická deprese → tropická bouře → hurikán*.
- K dispozici je šest prediktorů, na základě kterých by mělo být možné tyto dvě třídy hurikánů odlišit.
- Jedná se o datum, zeměpisnou šířku a délku tropické deprese (LATDEPR, LONDEPR) (První stádium při vzniku hurikánu) a datum, zeměpisnou šířku a délku, kdy bouře dosáhla statutu hurikánu (LATHUR, LONHUR).

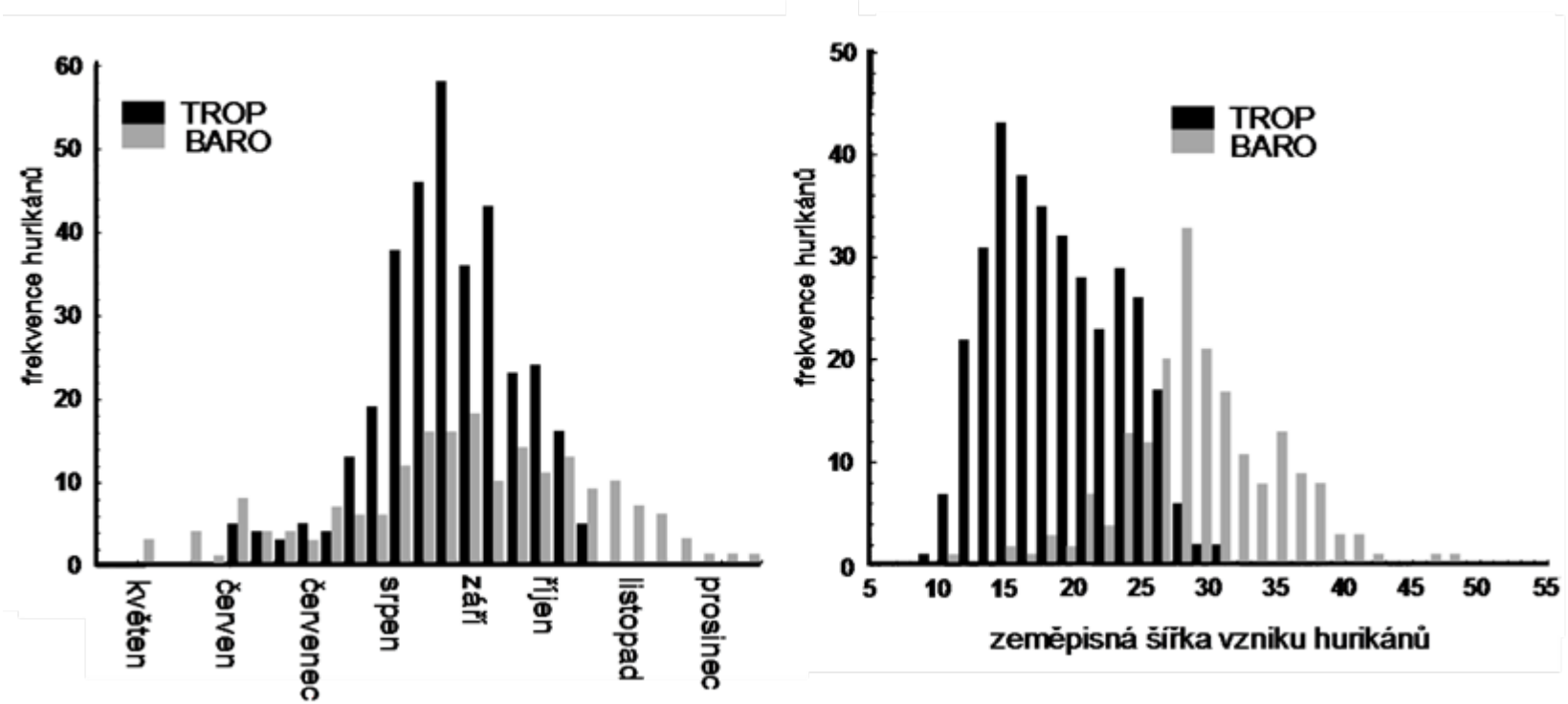
Příklad hurikány



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 ?

Rozložení hurikánů BARO a TROP v období jejich výskytů a podél zeměpisné šířky vzniku hurikánů.



Algoritmus růstu stromu CART

- Rozděl soubor na trénovací a testovací → poměr se určuje na základě počtu pozorování a účelu studie
- Najdi nejlepší rozdělení každého z prediktorů:
  - Pro spojité proměnné
    - seřaď hodnoty každého prediktoru od nejmenší po největší.
    - Projdi všechny hodnoty prediktoru *X* a spočítej kriteriální statistiku všech možných rozdělení proměnné *Y* na dva potenciální dceřiné uzly.
    - Pokud je dělicí hodnota *a* prediktoru *X* větší nebo rovna hodnotě x_i, pozorování y_i náleží do levého uzlu, jinak do pravého (popřípadě naopak).
    - Hodnota *a*, pro kterou je kriteriální statistika minimální, je vybrána jako nejlepší možné dělení závisle proměnné *Y* pomocí daného prediktoru.
    - Pro každý prediktor tak získáme jednu hodnotu (nejlepší potenciální rozdělení) kriteriální statistiky → Následně je vybrán prediktor s nejnižší hodnotou kriteriální statistiky a hodnota *a* je použita k rozdělení souboru (hodnot y_i) do dvou dceřiných uzlů.
  - Pro kategoriální prediktor
    - projdi všechny možné kombinace, tvořené jednotlivými kategoriemi prediktoru a hodnot nebo kategorií závisle proměnné →použij dělení s nejnižší hodnotou kriteriální statistiky.
- Rozděl soubor na dva dceřiné uzly t_1 a t_2 podle hodnoty prediktoru vybrané v kroku 2.
- Opakuj krok 2 a 3, dokud se dělení nezastaví na předem definované hodnotě (dokud není dosaženo některého z pravidel pro zastavení růstu stromu). Protože vybíráme vždy z celé množiny prediktorů, může být stejný prediktor použit ve stromě vícekrát.
- Použij testovací soubor k ověření vhodné velikosti stromu, a pokud je strom příliš velký, prořež strom.

Pravidla pro zastavení růstu stromu (*stopping rules*)

- Strom nemůže růst donekonečna → maximální velikost je dána velikostí souboru

- Strom se zastaví sám v těchto případech:
  - terminální uzel obsahuje pouze jedno pozorování;
  - všechna pozorování v uzlu mají stejnou hodnotu všech prediktorů;
  - všechna pozorování v uzlu mají stejnou hodnotu závisle proměnné.

- Strom můžeme v růstu omezit nastavením některých parametrů a k dalšímu rozdělení nedochází, pokud je dosaženo zadaných hodnot:
  - maximální počet větvení daného stromu;
  - maximální počet pozorování v koncovém uzlu;
  - frakce pozorování v uzlu, která již nemůže být oddělena;
  - velikosti chyby v potenciálních dceřiných uzlech - například uzel se nerozdělí, pokud střední kvadratická chyba (MSE) nebo procento nesprávně klasifikovaných vzorků v důsledku rozdělení překročí určitou hranici.

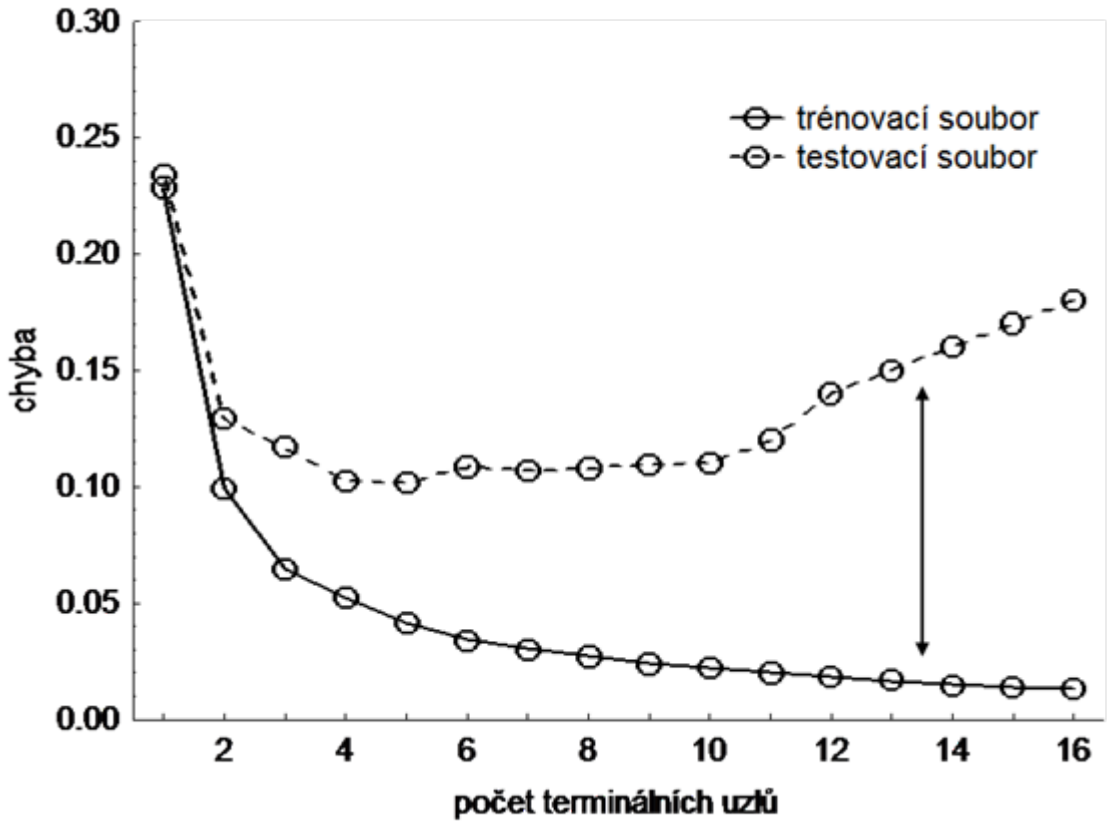
Stromy typu CART - pokračování

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!

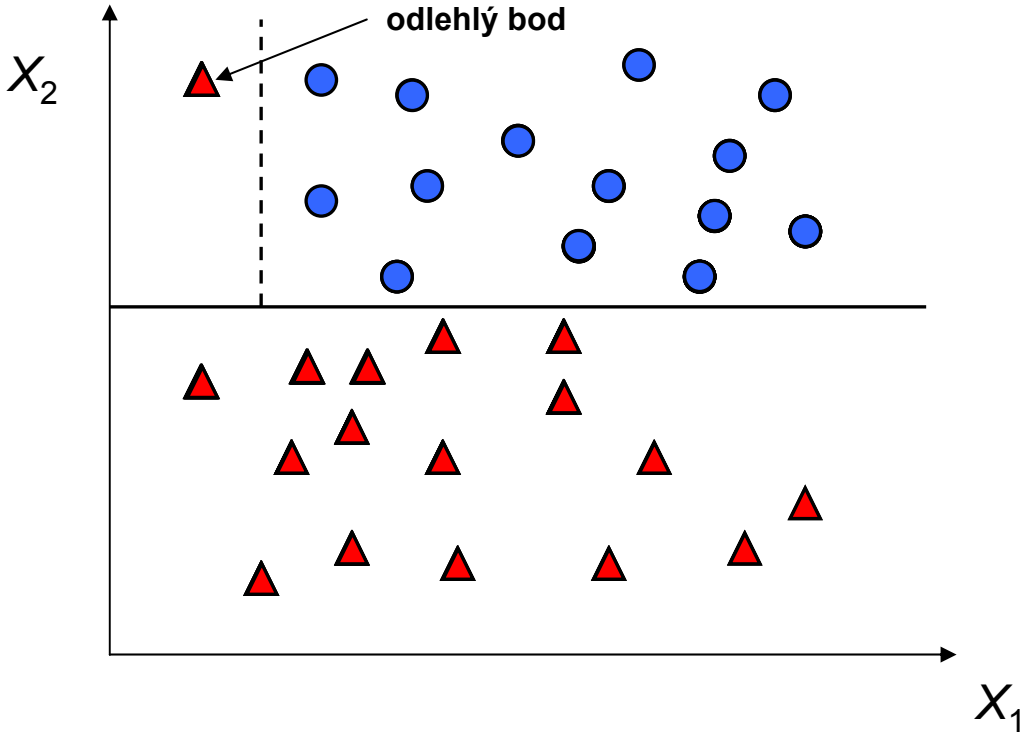
Rozdíl ve velikosti chyby mezi testovacím a trénovacím souborem při různé velikosti stromu, dané počtem terminálních uzlů



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.

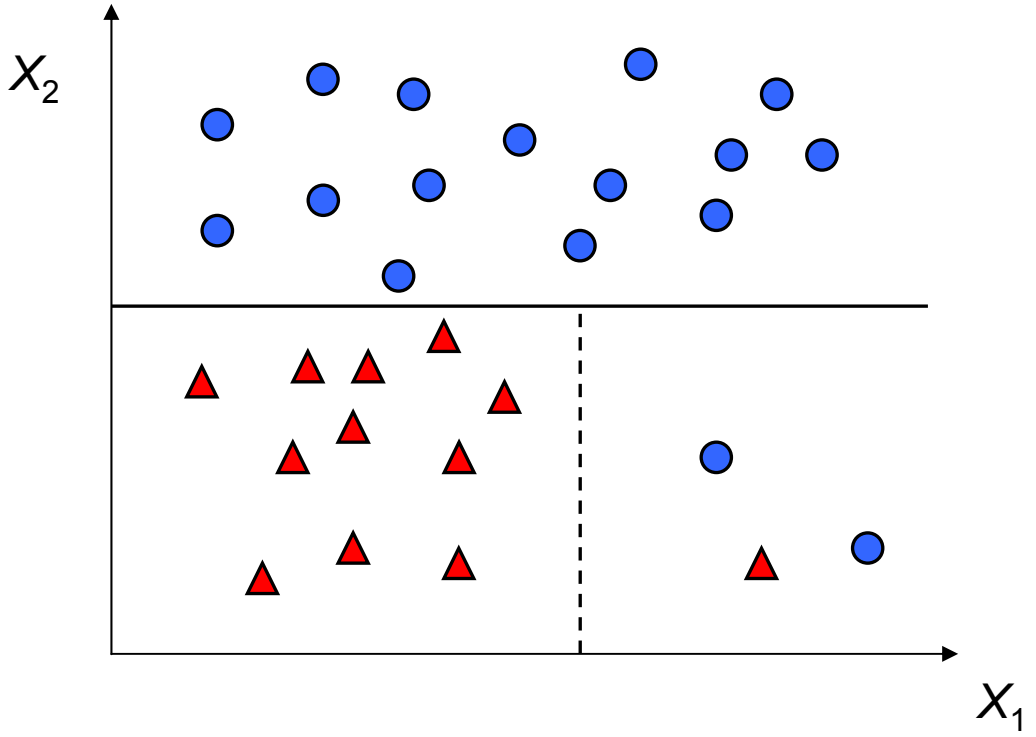
Příklad přetrénování stromu I

- 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):*

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

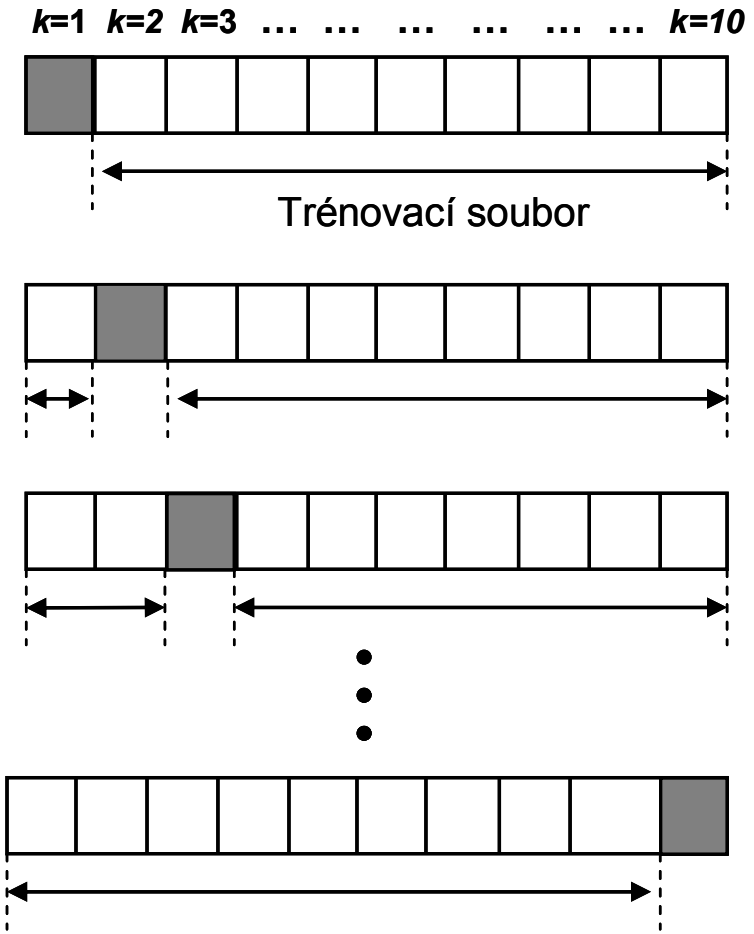
kde $|T_1|$ je počet terminálních uzlů stromu a DT_1 je deviance stromu. Parametr $\alpha \geq 0$ vyjadřuje kompromis mezi velikostí stromu a jeho vyčerpanou variabilitou. Hledáme tedy, pro každé α, takový strom , který minimalizuje $C\alpha(T)$.

K určení odhadu α se používá krosvalidace

Křížové ověřování (krosvalidace)

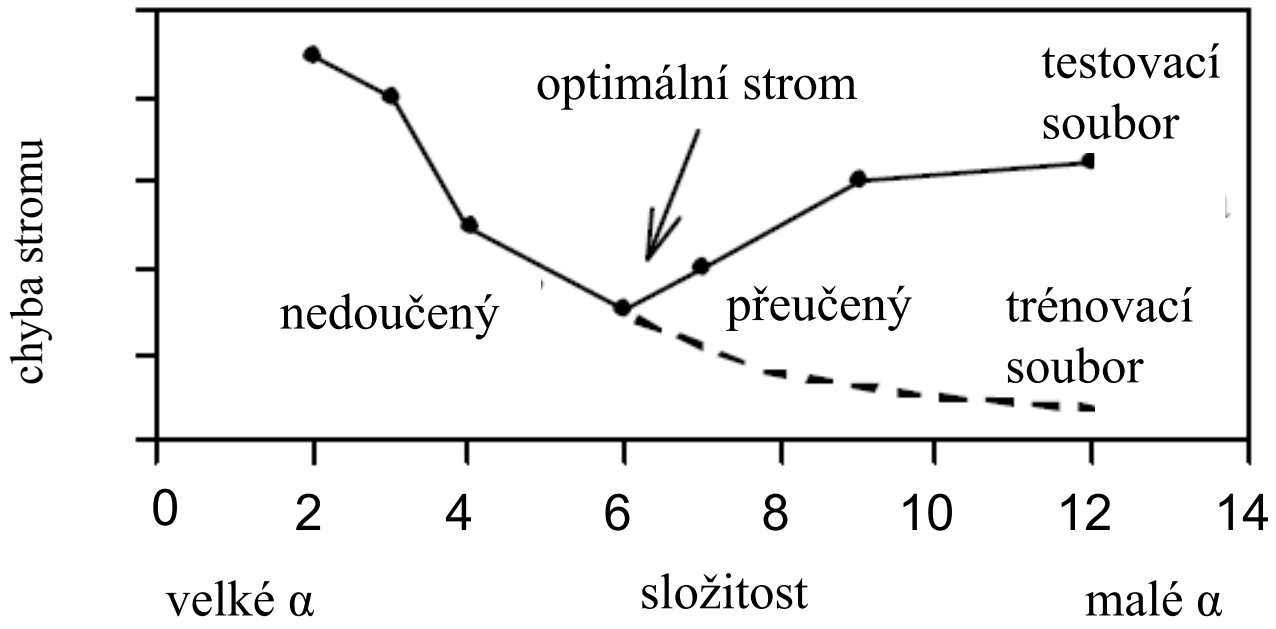
- Křížová validace patří mezi validační techniky
- Pozorování jsou rozdělena do *k* nezávislých podsouborů
- Jeden podsoubor se použije pro testování (pozorování nejsou použita při tvorbě modelu) všech ostatních *k*-1 skupin pro tvorbu modelu → je tedy vytvořeno *k* modelů otestovaných na *k* testovacích 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)

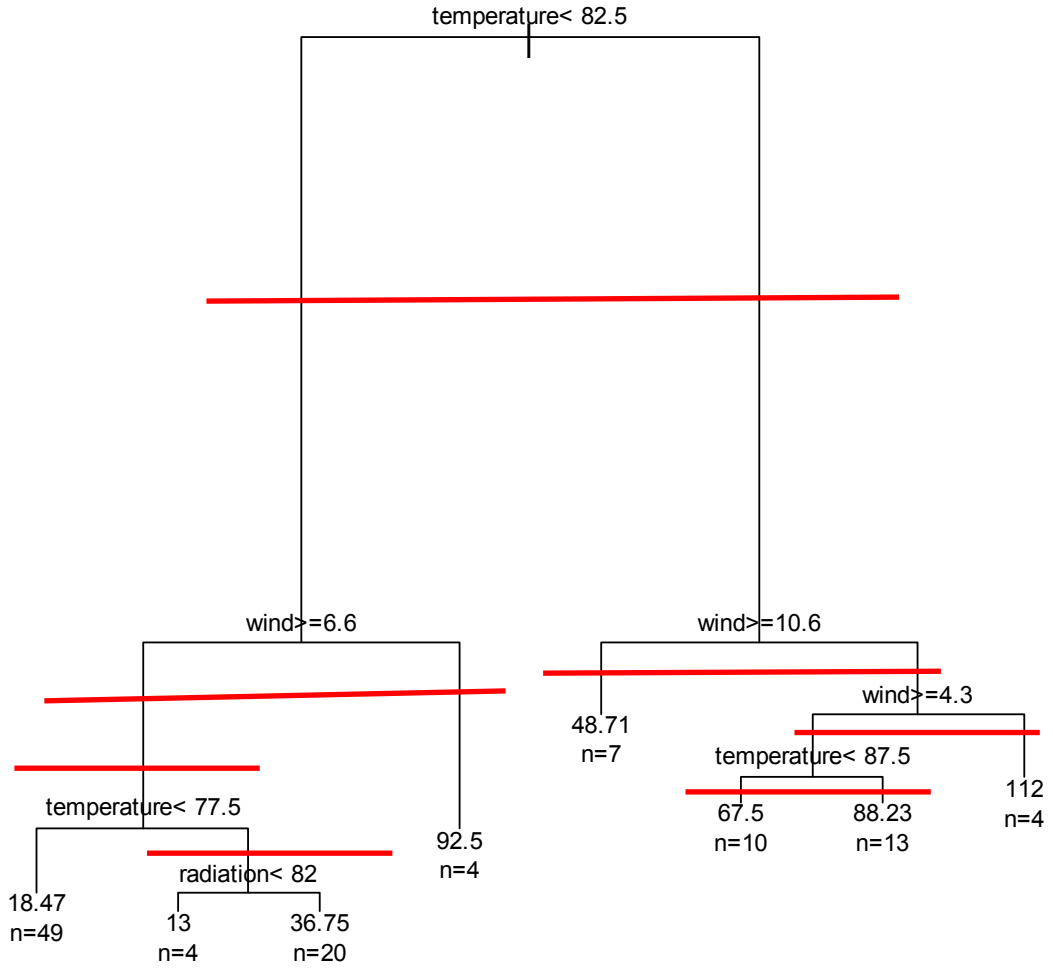


Výběr optimálního stromu

- 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



Příklad – ozón



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.

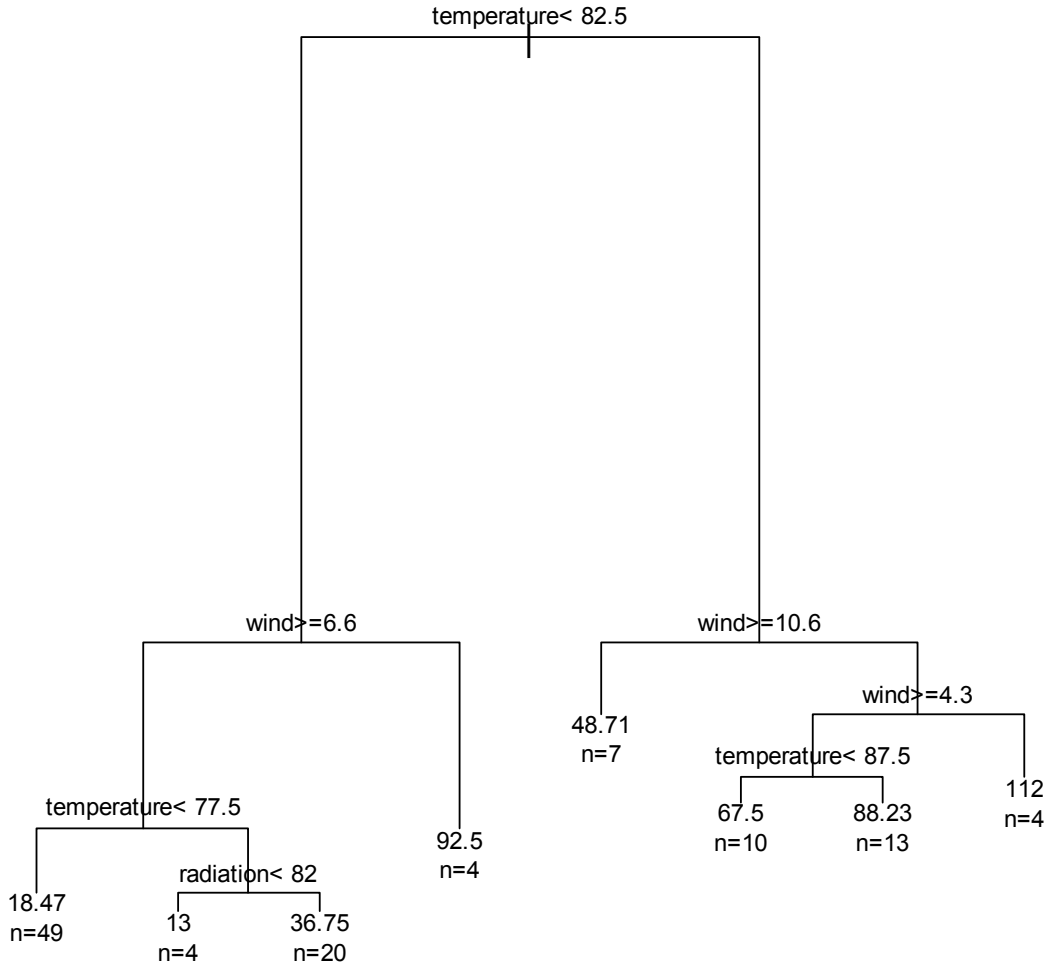
Určení přesnosti regresního stromu

- 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

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

- 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



Je strom vytvořený na základě prediktorů s nejnižší kriteriální statistikou opravdu nejpřesnější?

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

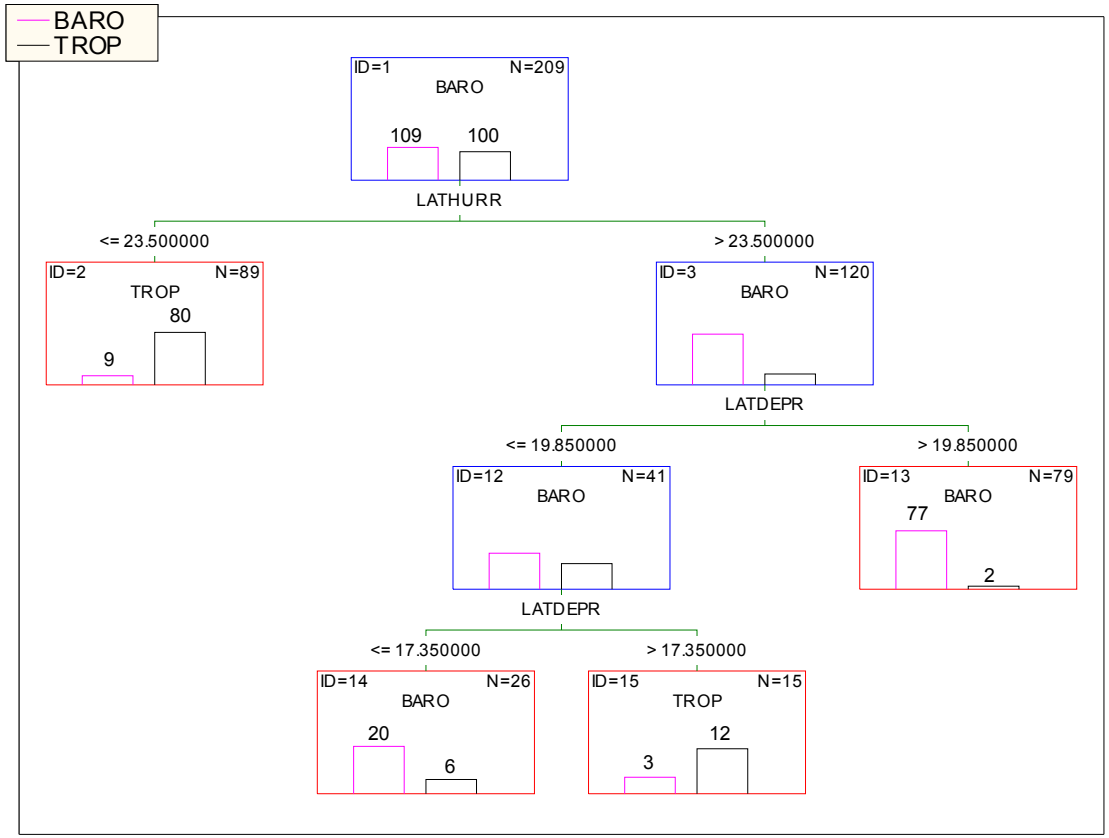
- Na základě hodnot kriteriální statistiky se tak v případě absence primární proměnné rozdělí uzel podle kompetitivní nebo zástupné proměnné.
- Je tedy vybrán jiný prediktor s další nejlepší hodnotou kriteriální statistiky.
- Velký význam pro interpretaci

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

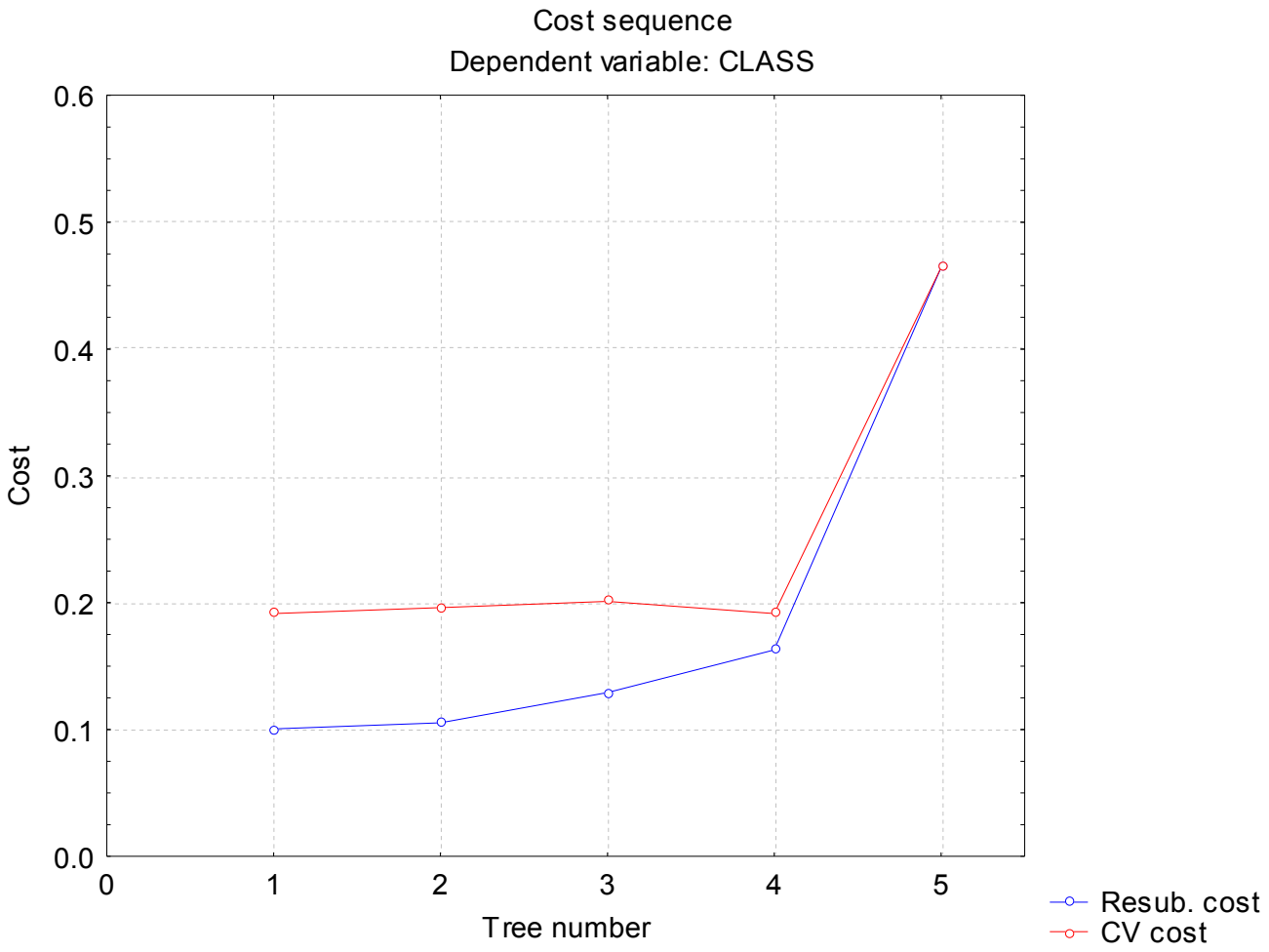
Příklad hurikány



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?

Příklad

- atlantické hurikány BARO a TROP jsme schopni klasifikovat v obou případech s vysokou přesností
  - OA_{BARO} = 97/109 = 0,89 a OA_{TROP} = 92/100 = 0,92
- V tomto případě by se jednalo o optimistický odhad chyby
  - $e'(t)$ = 1- OA_{tren} na trénovacím souboru
- Stejný výpočet bychom však mohli provést pro pozorování z testovacího souboru a získat objektivnější měření chyby
  - $e'(t)$ = 1- OA_{test}

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)

Příklad III: Regresní strom CART

- V tomto příkladu budeme sledovat závislost denního měření koncentrace ozónu (ppb) na rychlosti větru (míle/h), teplotě vzduchu (denní maximum ve stupních Fahrenheita) a intenzitě slunečního záření (cal/cm^2) v New Yorku. Soubor obsahuje celkem 111 měření, která proběhla od května do září v roce 1973.
- Přízemní ozón je součástí tzv. fotochemického smogu, který se vyskytuje v místech s intenzivní automobilovou dopravou. Jeho původcem jsou oxidy dusíku emitované jako součást spalin ze spalovacích motorů. Působením slunečního záření se tyto oxidy štěpí a vzniklé radikály reagují s kyslíkem za vzniku ozónu. Jeho zvýšené koncentrace můžeme tedy očekávat v letních měsících při vyšších teplotách. Určitý nárůst koncentrací ozónu lze ale očekávat i za slunečného počasí v chladnějších měsících, pokud jsou zhoršené rozptylové podmínky. Podíváme se, zdali jsou tato očekávání ověřitelná pomocí výše zmíněných měření.