Pokročilé neparametrické metody

Klára Komprdová

Pokročilé neparametrické metody

Výuka

- 11 přednášek doplněných o praktické cvičení v SW

- **Úvod do neparametrických metod + princip rozhodovacích stromů**
- **Klasifikační a regresní stromy typu CART**
- **Další typy stromů (MARS, PRIM, CHAID)**
- **Náhodné lesy -** Bagging, Boosting, Arcing, Random forest
- **Měření přesnosti modelů**
- **Validační techniky**
- **Příklady použití neparametrických metod**

průběžné testy z probírané látky (každou druhou hodinu)

Ukončení

- písemná zkouška (příklady; minimum 60% bodů) + ústní zkouška
- z průběžných testů lze získat 10% bodů do celkového testu!

Úvod do neparametrických metod
Princip rozhodovacích stromů

Rozdělení modelů

Popisuje budoucí stav systému nebo jeho podmínek?

ANO Dynamické modely - závislé na čase - s*pojité, diskrétní*
NE Statické modely - nezávislé na čase

Popisují prostorovou strukturu?

ANO Prostorově heterogenní - *diskrétní*, *spojité*
NE Prostorově homogenní modely

Zahrnuje náhodnou složku?

ANO Stochastické modely
NE Deterministické modely

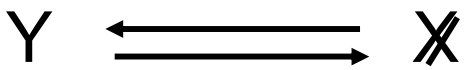
Typy proměnných

- Kvalitativní (kategoriální)
  - lze pouze určit, zda jsou dvě „hodnoty“ stejné nebo se liší
  - typ půdy, barvy, typ habitatu
- Semikvantitativní (ordinální)
  - Lze určit rovněž pořadí hodnot
  - abundanční třídy, řády toku, teplota po stupních
- Kvantitativní (spojité)
  - lze provádět všechny matematické operace
  - Intervalové, poměrové
  - Výška, váha, počty druhů
- *binární*
  - lze ji považovat za kvantitativní, semikvantitativní i kvalitativní proměnnou
  - výskyt/ nevýskyt druhu, odpověď pacientů na léčbu, výsledky dotazníků typu ANO/NE

Typy proměnných

- Ze statistického hlediska
  - závisle proměnná (vysvětlovaná) – proměnná, jejíž hodnoty chceme vysvětlit a/nebo předpovědět pomocí jiných proměnných, na kterých závisí
  - vysvětlující proměnné, nezávisle proměnné, prediktory – proměnné, pomocí nichž se snažíme vysvětlit závisle proměnnou

Vztah – lineární, nelineární

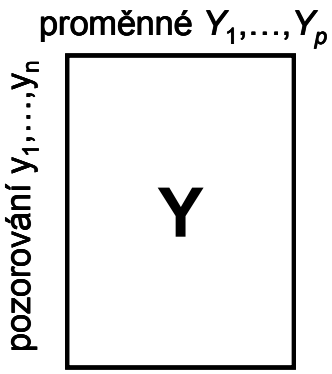


Rozdělení stochastických metod

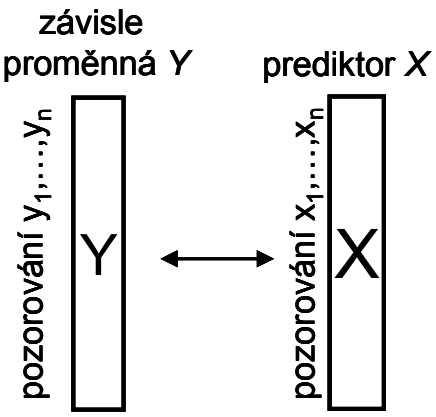
- Parametrické x Neparametrické
  - *Parametrické* – předpoklady o rozdělení dat
    - Klasické lineární modely, zobecněné lineární modely, lineární diskriminační analýza
  - *Neparametrické* – nemají předpoklady o rozložení dat
    - Rozhodovací stromy, lesy, neuronové sítě…
  - *Semiparametrické – Zobecněné aditivní modely, metoda podpůrných vektorů*
- Regresní x Klasifikační
  - *Regresní* - modelujeme závislost *spojité* závisle proměnné na jedné či více nezávislých proměnných
  - *Klasifikační* - modelujeme závislost *kategoriální* závisle proměnné na jedné či více nezávislých proměnných
- Lineární x Nelineární
- Jednorozměrné x Vícerozměrné

Rozdělení metod podle počtu závisle proměnných a prediktorů

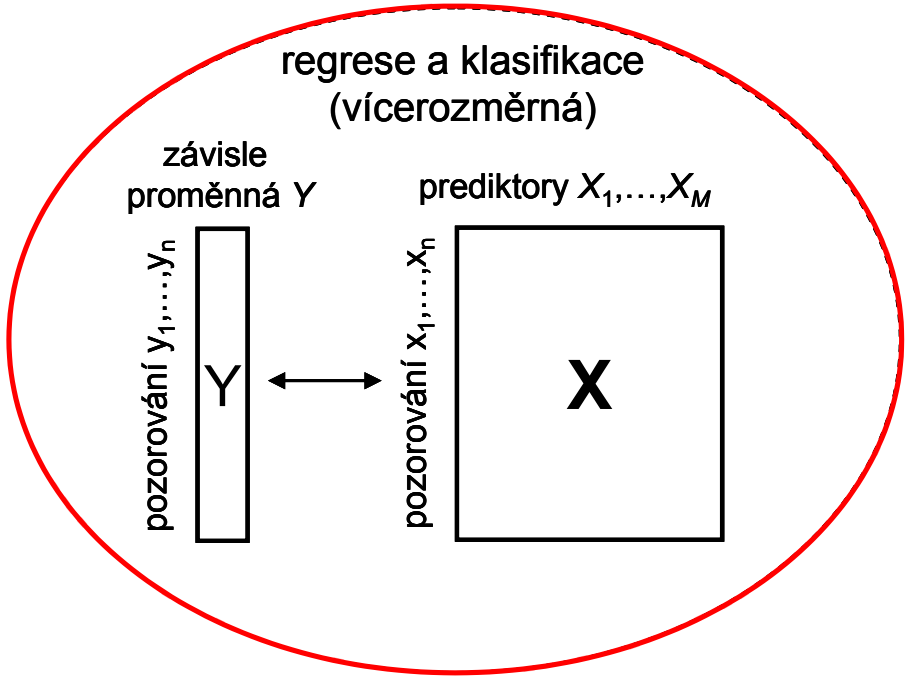
nepřímé ordinační techniky,
shlukovací metody
(vícerozměrné)



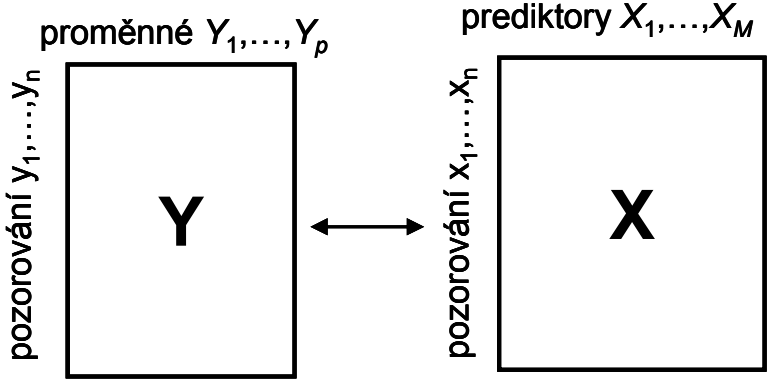
regrese, klasifikace
(jednorozměrná)



regrese a klasifikace
(vícerozměrná)



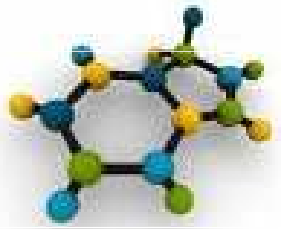
přímé ordinační techniky
(vícerozměrné)



Z jiného pohledu - živočichové x rostliny x proměnné prostředí

x

x



Procesově orientované modely (deterministické)

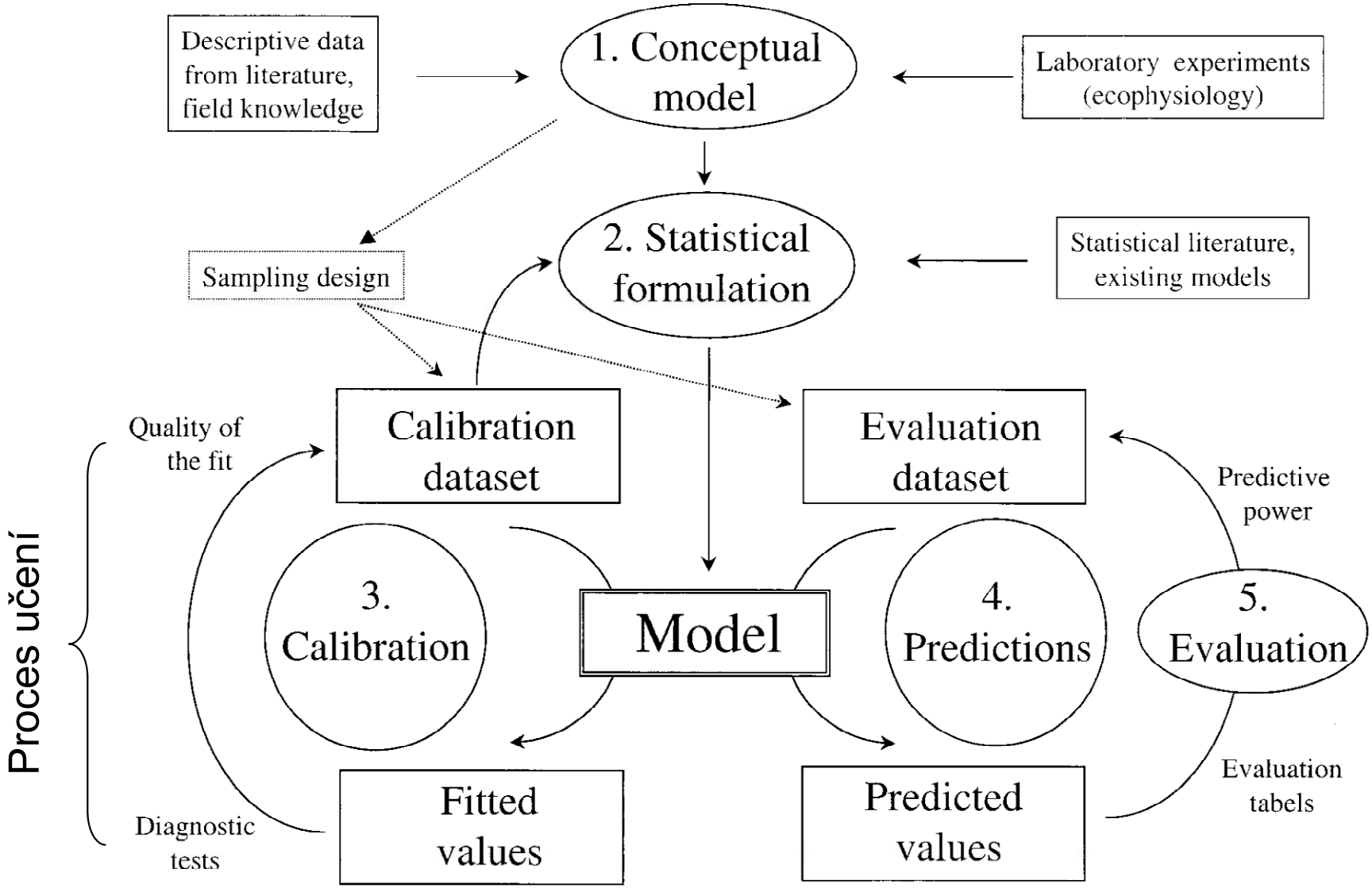
Stochastické modely

Procesově orientované modely (deterministické) Stochastické modely, interpolační

Proces modelování I

- Design vzorkování a zpracování dat (z literatury, předešlých experimentů)
- Terénní sběr dat a laboratorní analýzy
- Analýza datového souboru a tvorba modelu
- Kalibrace a validace modelu
- Interpretace modelu, jeho srovnání s realitou
- použití modelu

Proces modelování II



Proces modelování III

- **simulace -** použití modelu na libovolném datovém souboru, i uměle vytvořeném. Simulace může sloužit k hlubšímu pochopení modelovaných procesů a zjištění chování modelu při limitním nastavení jeho parametrů
- **validace** - porovnání výsledků modelu s nezávislým datovým souborem, (např. získaným experimentálně na jiné lokalitě, nebo v jiném roce). Parametry modelu jsou již pevně stanoveny předchozí kalibrací. Pro pojem validace se velmi často používá také obecnější pojem testování
- **robustnost** - ověření funkčnosti modelu při opakované aplikaci např. za různých environmentálních podmínek a na různých lokalitách
- **post audit** - srovnání předpovědi výsledku modelu s experimentální činností prováděnou v budoucnosti
- **analýza citlivosti** - zjištění efektu malých změn parametrů modelu na jeho výsledek
- **analýza nejistot** - stanovení standardní odchylky predikované proměnné (jejího průměru) na základě nejistot ve vstupních parametrech modelu
- **expertní posouzení** - odborné zhodnocení, zda model obsahuje všechny důležité procesy a závislosti, jestli jsou správně matematicky formulovány a zdali model správně popisuje modelovaný problém
- **tolerance k šumu** - tolerance k irelevantním neboli odlehlým pozorováním.
- **stabilita** – model je stabilní, pokud při malé změně dat nedojde k rozdílným výsledkům modelu
- **predikce** – předpověď nových hodnot pomocí modelu

Srovnání vlastností metod

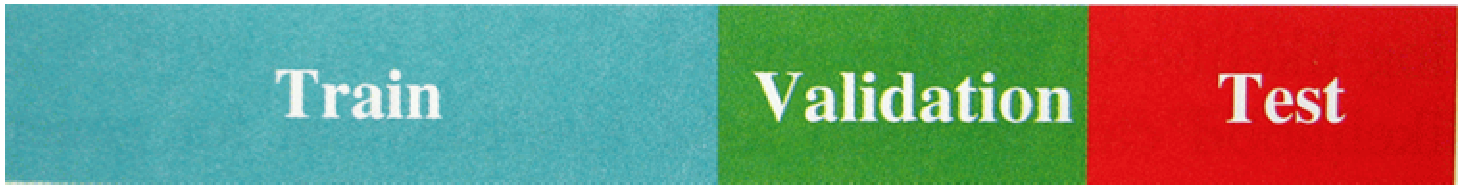
KLM - Klasický lineární model, GLM – Zobecněné lineární modely, GAM – Zobecněné aditivní modely, LDA – Lineární diskriminační analýza, CART- Klasifikační a regresní stromy, RF – Random forest, SVM – Metoda podpůrných vektorů, NNs – Neuronové sítě, Naivní bayes. – Naivní bayesovský klasifikátor, k-NN – metoda nejbližšího souseda

	KLM	GLM, GAM	LDA	CART	RF	SVM	NNs	Naivní bayes.	k-NN
Použití pro klasifikaci	🔴	🔴	🟢	🟢	🟢	🟢	🟢	🟢	🟢
Použití pro regresi	🟢	🟢	🔴	🟡	🟡	🟢	🟢	🔴	🟡
Distribuční předpoklady	🔴	🟡	🔴	🟢	🟢	🟢	🟢	🟢	🟢
Celková přesnost predikce	🟢	🟢	🟢	🟡	🟢	🟢	🟢	🟡	🟡
Použití prediktorů různých typů	🔴	🔴	🔴	🟢	🟢	🔴	🔴	🔴	🔴
Tolerance k velkému počtu prediktorů	🔴	🔴	🔴	🟢	🟢	🟢	🟢	🟡	🟢
Tolerance k redundantním proměnným	🔴	🔴	🔴	🟢	🟢	🟢	🟢	🔴	🟡
Tolerance k odlehlým hodnotám	🔴	🔴	🔴	🟢	🟢	🔴	🔴	🔴	🟡
Metoda vhodná pro malý počet pozorování	🟢	🟢	🟡	🟡	🟢	🔴	🔴	🟡	🟡
Metoda vhodná pro velký počet pozorování	🟡	🟡	🟡	🟢	🟢	🟢	🟢	🟢	🟢
Tolerance k nerelevantním proměnným	🟢	🟢	🟢	🟢	🟢	🟡	🔴	🟢	🔴
Tolerance k šumu	🟢	🟢	🟢	🟢	🟢	🟡	🟡	🟢	🔴
Stabilita	🟡	🟡	🟡	🔴	🟢	🟢	🟢	🟢	🔴
Interpretovatelnost modelu	🟢	🟢	🟡	🟢	🔴	🔴	🔴	🟡	🔴
Náročnost nastavení parametrů modelu	🟢	🟢	🟢	🟡	🔴	🔴	🔴	🟢	🟢

Legenda: 🟢 výborné 🟡 dobré 🔴 problematické

Validace modelu

- validace modelu je jedním z nejdůležitějších bodů v procesu modelování
- probíhá s použitím různých datových souborů



- Trénovací - soubor k tvorbě modelu
- Testovací – soubor ke kalibraci modelu
- Validační – nezávislý soubor k validaci modelu (např. jiné území, skup. pacientů…)
- Ve skutečnosti většinou nenastává takto ideální situace a nezávislý testovací soubor nemusí být k dispozici. Pro tyto případy se používají různé **validační techniky**.

- **!vybrat „nejjednodušší“ model, vysvětlující největší množství informace!**

Validace modelu

Validační techniky:

- Analytické - zahrnující například informační kritéria (AIC, BIC)
- Založené na opakovaném použití pozorování - krosvalidace, jednoduché rozdělení, bootstrap, jacknifing

Odhady celkové chyby pomocí validačních technik jsou používány:

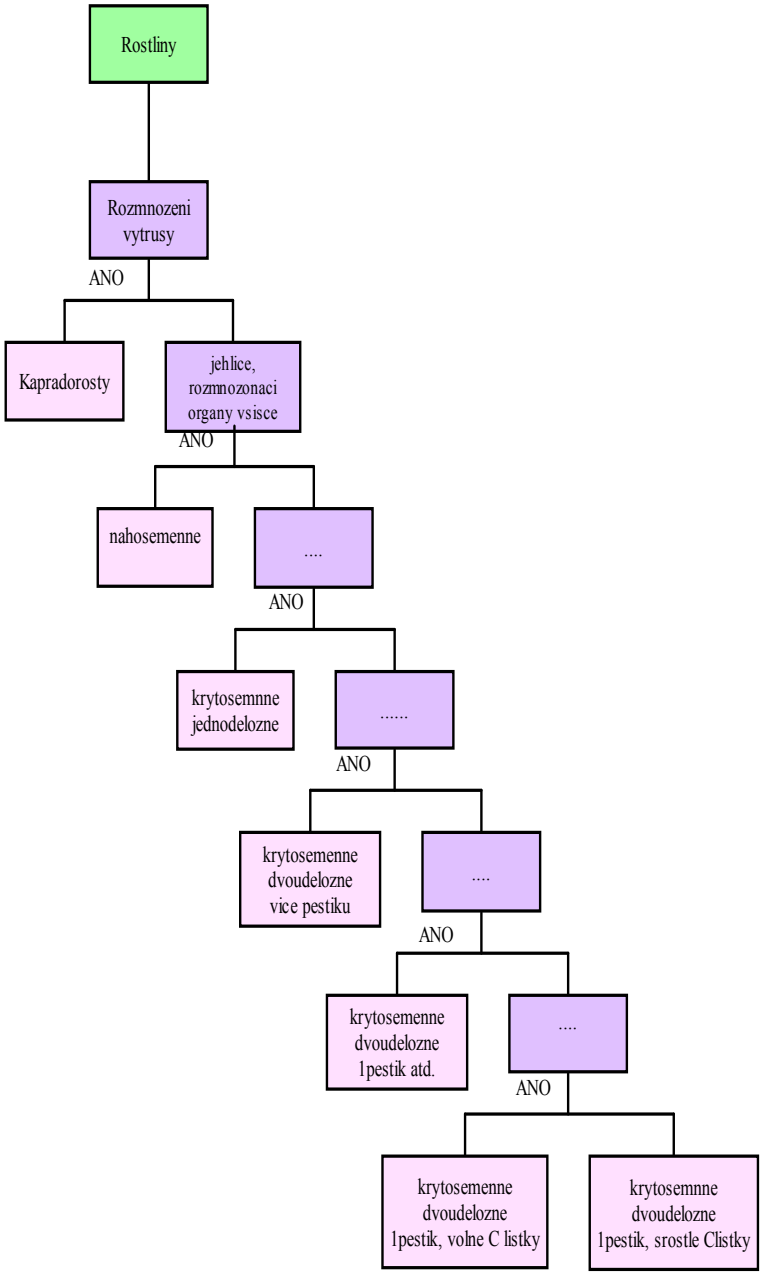
- pro výběr mezi různými modely
- k odhadu stability modelu
- k zjištění obecné platnosti modelu
- k určení složitosti modelu
- k výběru proměnných do modelu

Rozhodovací stromy (*Decision Trees*)
Úvod

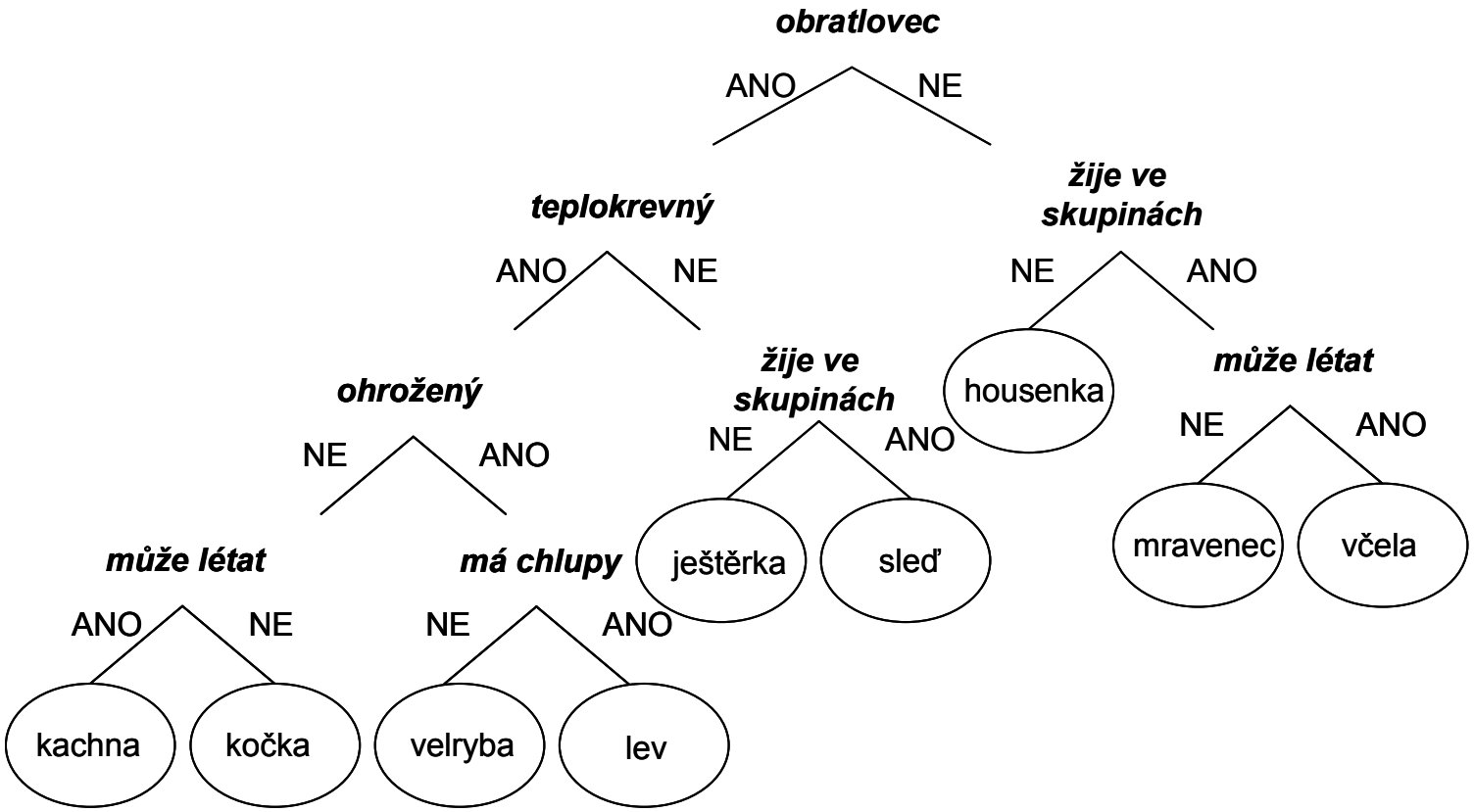
Regresní a klasifikační stromy (*Regression and Classification Trees*)

- jsou nejméně formální a nejméně parametrickou skupinou **statistických modelů**
- model – popisuje vzájemné vztahy mezi pozorovanými veličinami
- sada hierarchicky uspořádaných rozhodovacích pravidel
- se stromovou strukturou se setkáváme poměrně často, neboť je přehledná a snadno interpretovatelná - rodokmeny, fylogenetické (evoluční) stromy, botanické klíče nebo zobrazení adresářů a jejich podsložek v počítači...
- terminologie – analogie se stromy v přírodě → stromy rostou, větví se, prořezávají

Botanický klíč – určení skupin
Klíč ke Květeně České republiky, str.48

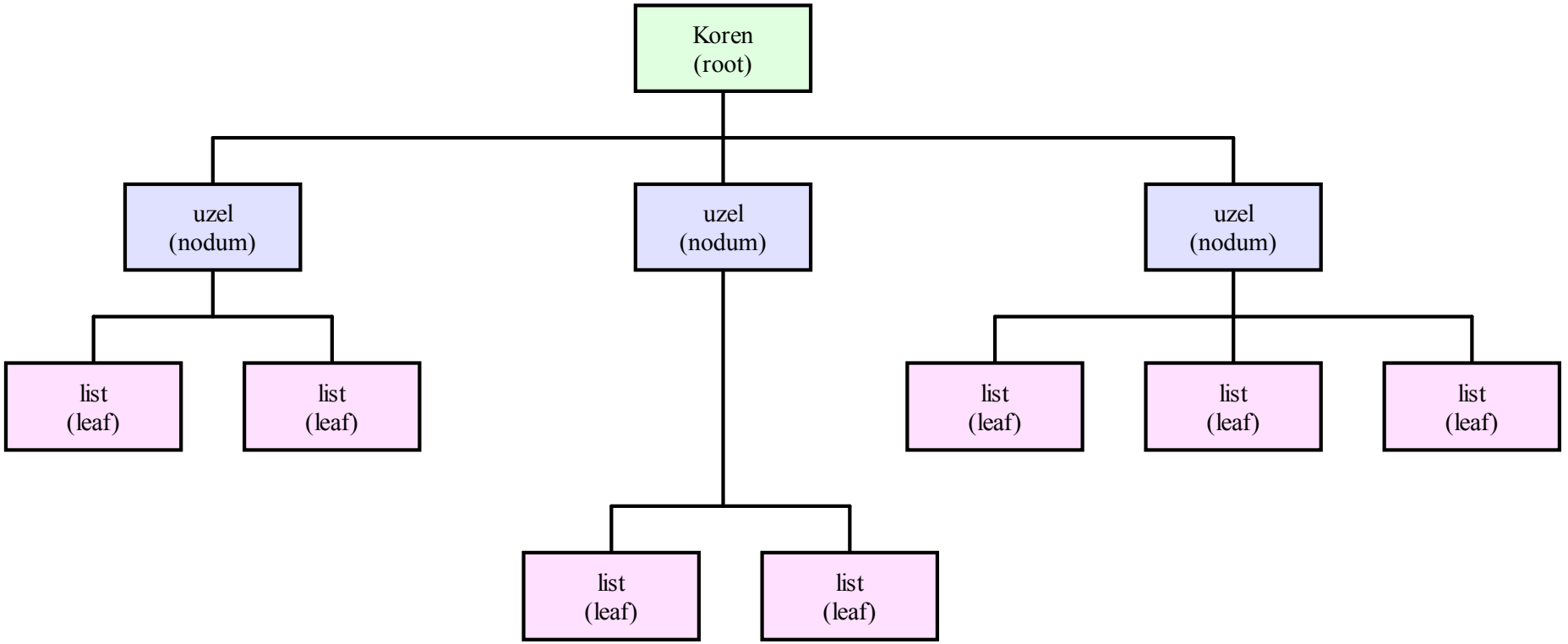


Rozdělení živočichů podle vlastností



	teplokrevný	může létat	obratlovec	ohrožený	žije ve skupinách	má chlupy
kočka	ANO	NE	ANO	NE	NE	ANO
kachna	ANO	ANO	ANO	NE	ANO	NE
sleď	NE	NE	ANO	NE	ANO	NE
lev	ANO	NE	ANO	ANO	ANO	ANO
ještěrka	NE	NE	ANO	NE	NE	NE
velryba	ANO	NE	ANO	ANO	ANO	NE
mravenec	NE	NE	NE	NE	ANO	NE
včela	NE	ANO	NE	NE	ANO	ANO
housenka	NE	NE	NE	NE	NE	ANO

Struktura stromu

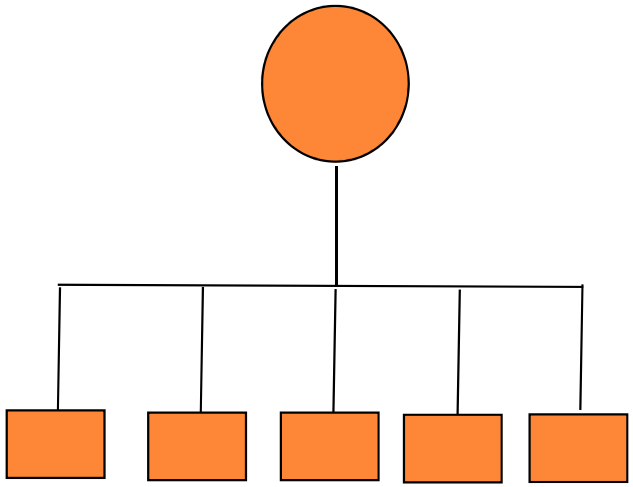


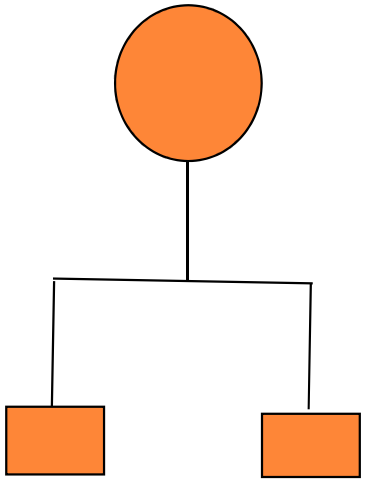
- rozhodovací strom se skládá z **kořene a uzlů** - v každém neterminálním uzlu se strom větví
  - **uzly**
    - **Terminální**
    - **Neterminální (list)**
    - **Mateřské x dceřiné**

kořen představuje celý soubor a postupně probíhá větvení do dalších uzlů → strom roste
- uzly, které se již dále nedělí, se označují jako terminální uzly nebo také listy

typy stromů – binární x nebinární

- **Binární stromy** – z jednoho uzlu vyrůstají právě dvě větve
- **Nebinární stromy** – z jednoho uzlu vyrůstají dvě a více větví





Regresní a klasifikační strom

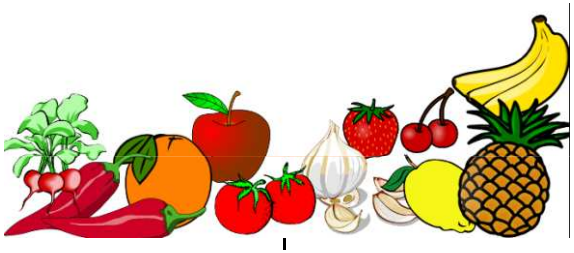
Mějme strom T s uzly $t = (t_1,...,t_N)$.

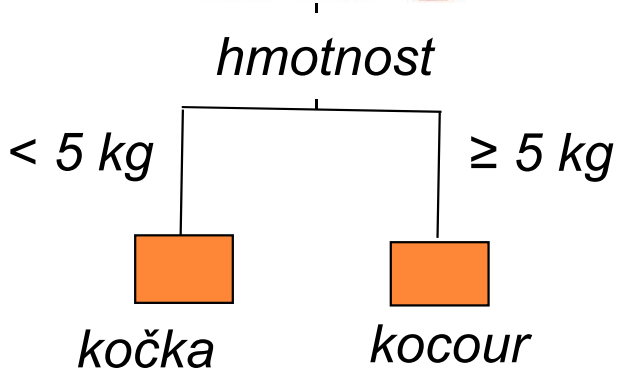
- **klasifikační strom** - pozorování kategoriální závisle proměnné Y s J kategoriemi jsou zařazeny do některé z kategorií $c = (c_1,..,c_J)$, kde $J \geq 2$.
  - Spamy – určení, který doručený e-mail je spam a který není spam.
  - Kosatce – třídění kostaců do jednotlivých druhů na základě velikosti jejich okvětních a kališních lístků

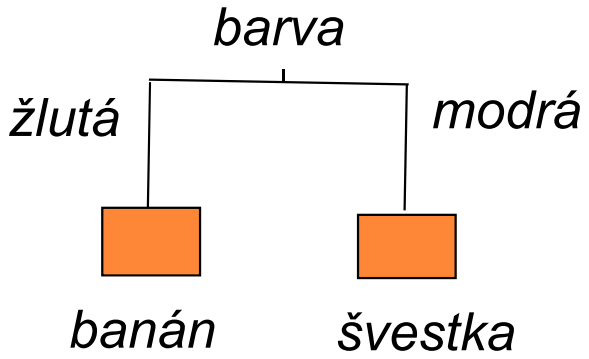
- **regresní strom** - Pokud je závisle proměnná spojitá $Y = (y_1,...,y_n)$, pozorováním je přiřazena hodnota predikovaná modelem $\hat{y}_i$ a výsledný strom bude regresní.
  - Ozón – modelování množství ozonu v závislosti na nadmořské výšce, teplotě a rychlosti větru
  - Závislost spotřeby plynu na venkovní teplotě

Prediktory

- Pozorování proměnné Y jsou rozdělena do uzlů hodnotami vysvětlujících proměnných (prediktorů) $X_1,...,X_M$.
- Rozdělení je znázorněno graficky pomocí větví stromu.
- Pokud jsou prediktory kategoriální, hodnoty y_i jsou rozděleny podle kategorií prediktoru X - odpovídáme na otázku, které pozorování y_i patří do množiny kde $x_i \in A$, přičemž A je neprázdná vlastní podmnožina množiny všech hodnot veličiny X.
  - př. Rozdělení ovoce na základě barev
- V případě spojitého prediktoru rozdělujeme Y pomocí hodnoty a daného prediktoru X - pozorování y_i patří do prvního uzlu, pokud je $x_i \geq a$ a do druhého uzlu pokud je $x_i < a$.
  - př. určení pohlaví dospělých koček (závisle proměnná) na základě jejich hmotnosti (prediktor).







Obecně…

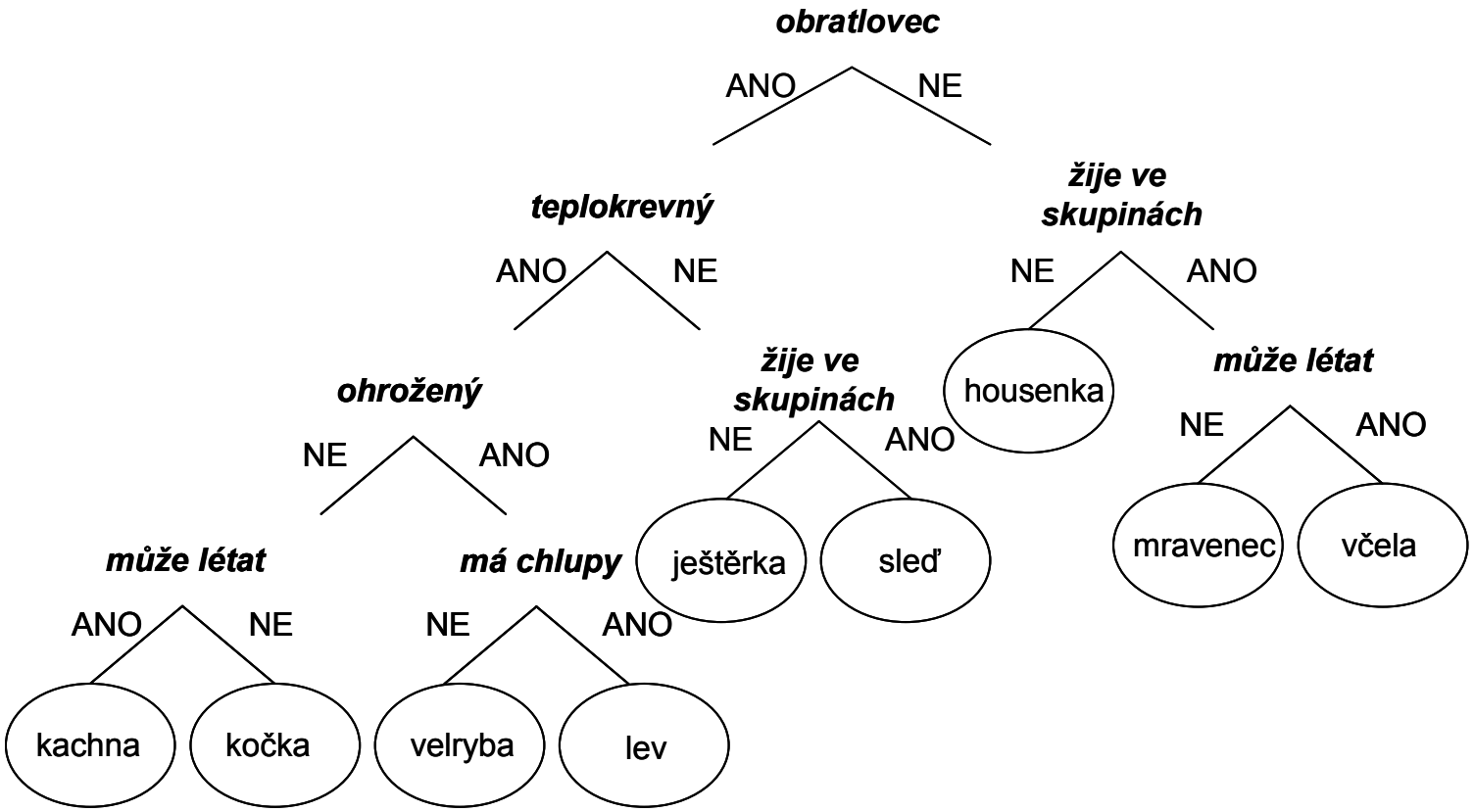
- k danému větvení stromu je použito vždy jen jednoho prediktoru
- stejný prediktor však může být využit v dalším větvení
- každé pozorování tak patří pouze do jednoho terminálního uzlu
  - je mu přiřazena kategorie (klasifikační strom)
  - nebo průměr hodnot (regresní strom) závisle proměnné *Y* tohoto uzlu

- stromy nekladou nároky na rozložení dat, jako například konstantní rozptyl, normální rozložení nebo nezávislost prediktorů…

- parametry algoritmu jsou často určeny experimentálně testováním různých nastavení jejich hodnot -tento postup však skrývá nebezpečí zejména při kalibraci modelu, která může být do jisté míry *subjektivní a závisí na zkušenosti badatele*

  → ! je potřeba opatrnosti při tvorbě a interpretaci modelu !

Typ stromu? Typ prediktoru?

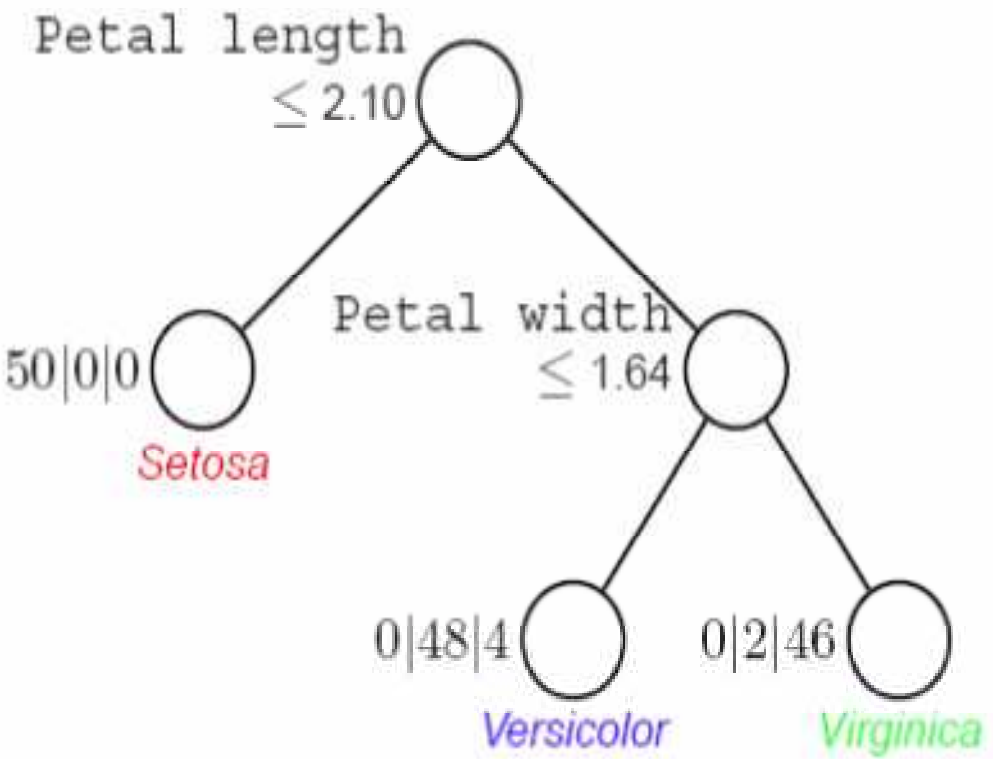


	teplokrevný	může létat	obratlovec	ohrožený	žije ve skupinách	má chlupy
kočka	ANO	NE	ANO	NE	NE	ANO
kachna	ANO	ANO	ANO	NE	ANO	NE
sleď	NE	NE	ANO	NE	ANO	NE
lev	ANO	NE	ANO	ANO	ANO	ANO
ještěrka	NE	NE	ANO	NE	NE	NE
velryba	ANO	NE	ANO	ANO	ANO	NE
mravenec	NE	NE	NE	NE	ANO	NE
včela	NE	ANO	NE	NE	ANO	ANO
housenka	NE	NE	NE	NE	NE	ANO

Př: Rozhodovací strom pro kosatce

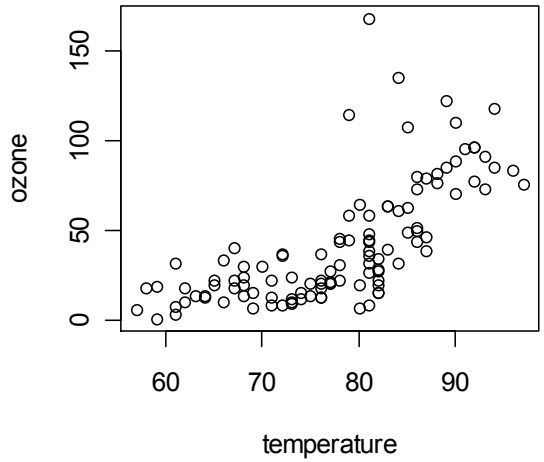
- 150 případů, vždy 50 případů ve skupině
- 3 skupiny – druhy kosatců: Setosa, Versicolour, Virginica
- 4 prediktory: délka a šířka korunních a kališních lístků

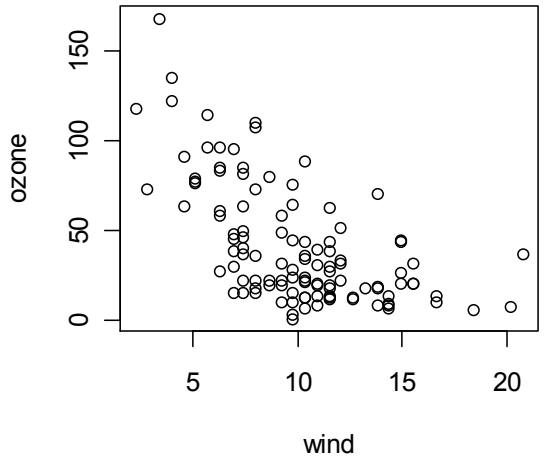
- Zdroj příkladu: Yu-Shan Shih - Tree-structured methods - IRIS data

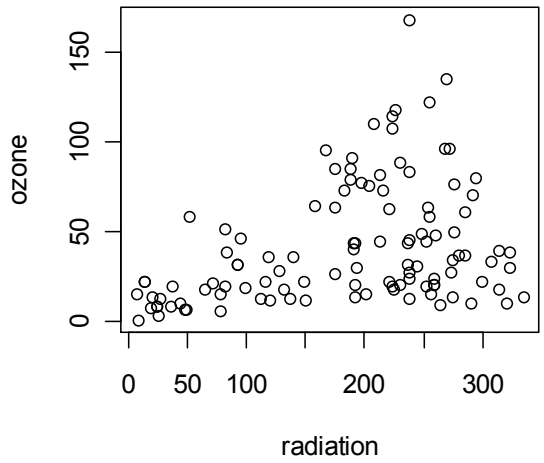


Příklad -ozón

denní měření koncentrace ozónu (%) v závislosti na rychlosti větru, teplotě vzduchu a intenzitě slunečního záření v New Yorku

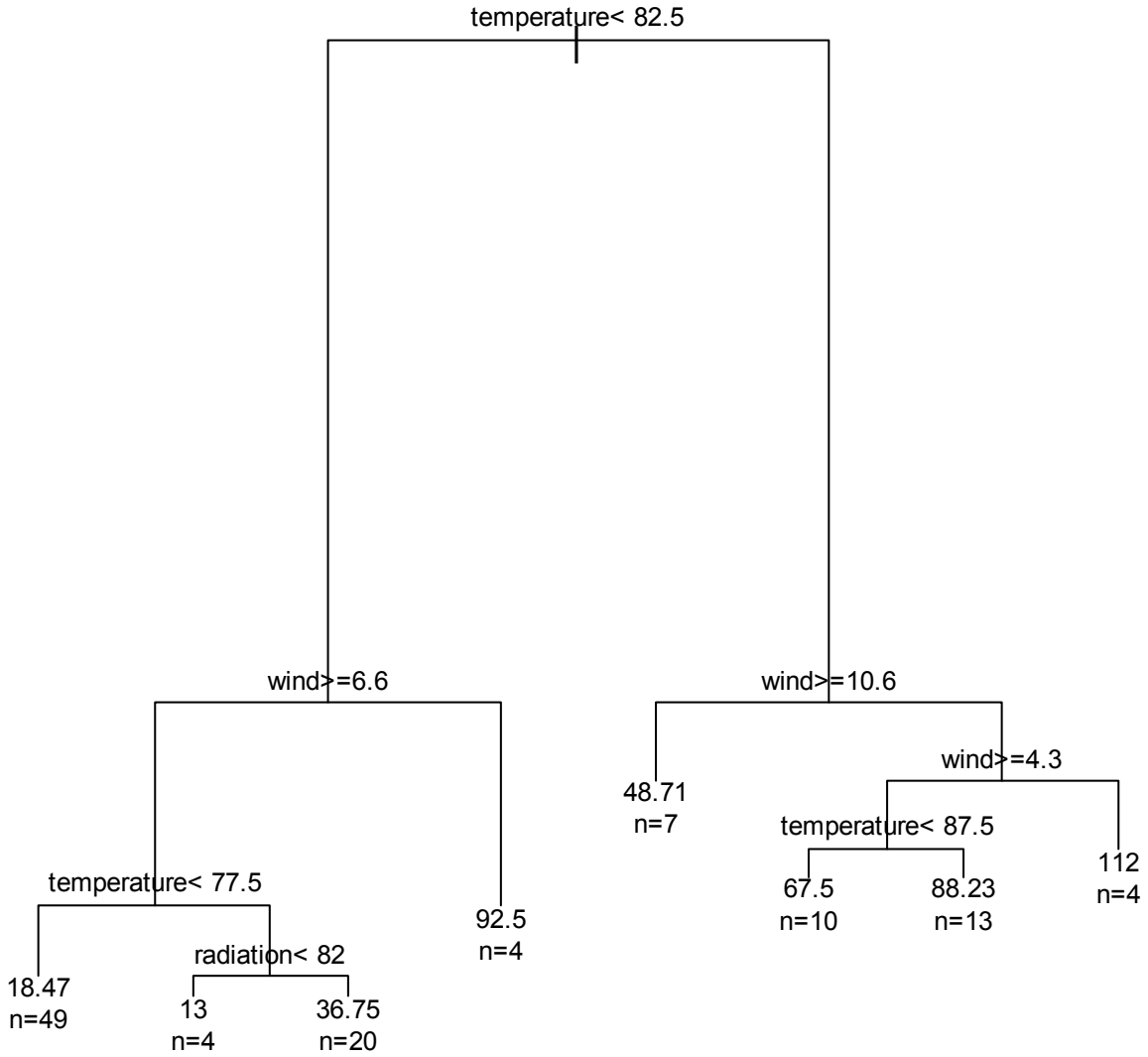






n = 111

Příklad – ozón



Typy stromů

- Existuje celá řada algoritmů pro vytváření stromů
-
- CART a C4.5 - nejznámější a nejpoužívanější

- CHAID pro kategoriální a ordinální proměnné
- stromy určené pro regresní problémy PRIM a MARS
  - nedají se zobrazit pomocí stromové struktury
  - PRIM - sada rozhodovacích
  - MARS – výstupem je regresní rovnice

- princip tvorby stromu je pro všechny algoritmy velmi podobný
- liší se především v nalezení vhodného prediktoru *X* pro každou hierarchickou úroveň stromu a hodnoty prediktoru *a* pro rozdělení proměnné *Y*

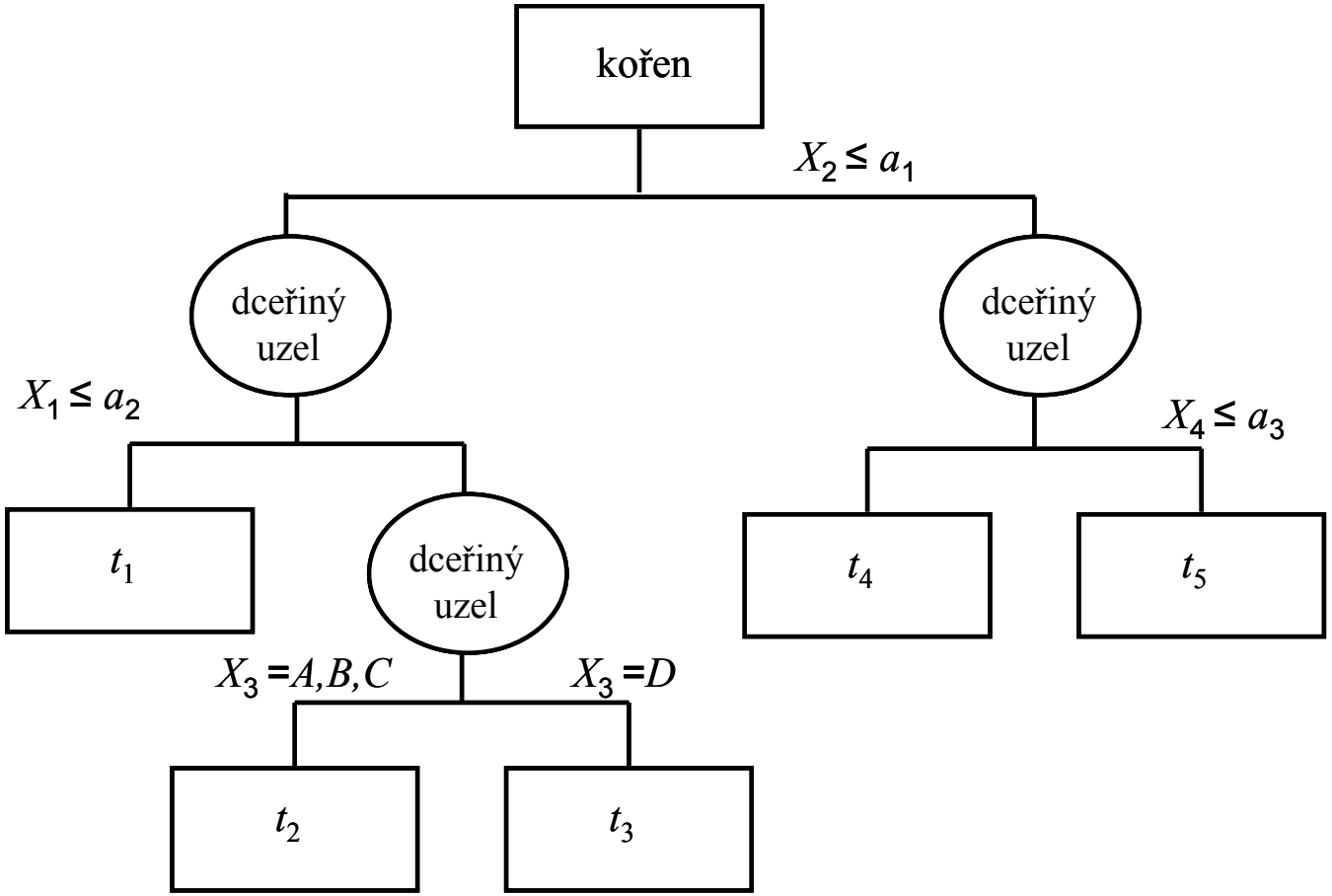
K čemu budeme stromy využívat?

- zajímá nás struktura těchto dat, postižení vzájemných vztahů **– explanatorní technika**
- klasifikace nebo predikce dosud neznámých případů

Stromy typu CART

Strom typu CART

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

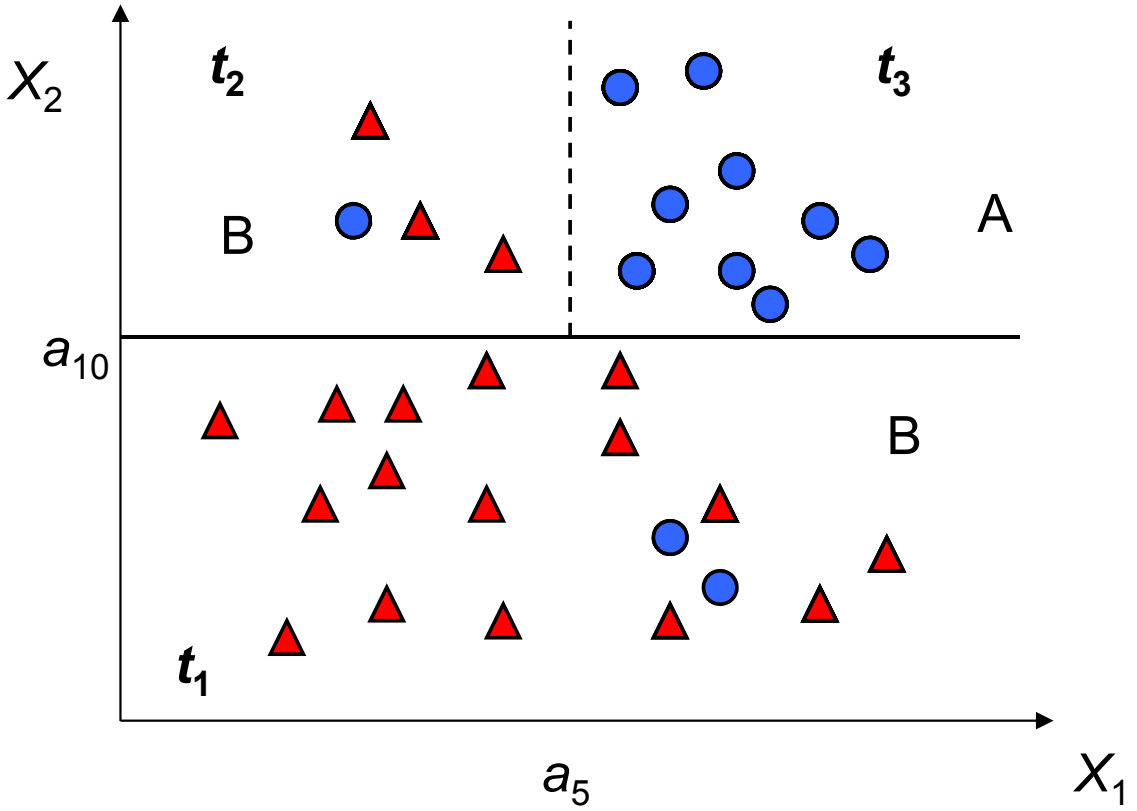


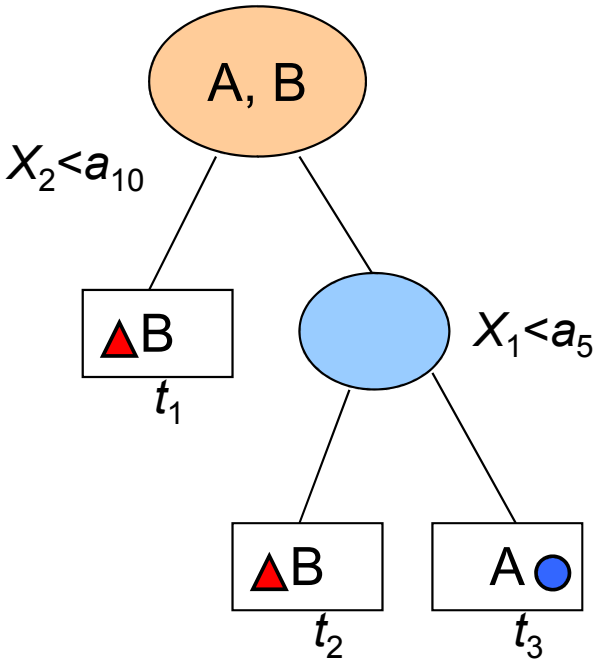
Jak roste strom CART?

- 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

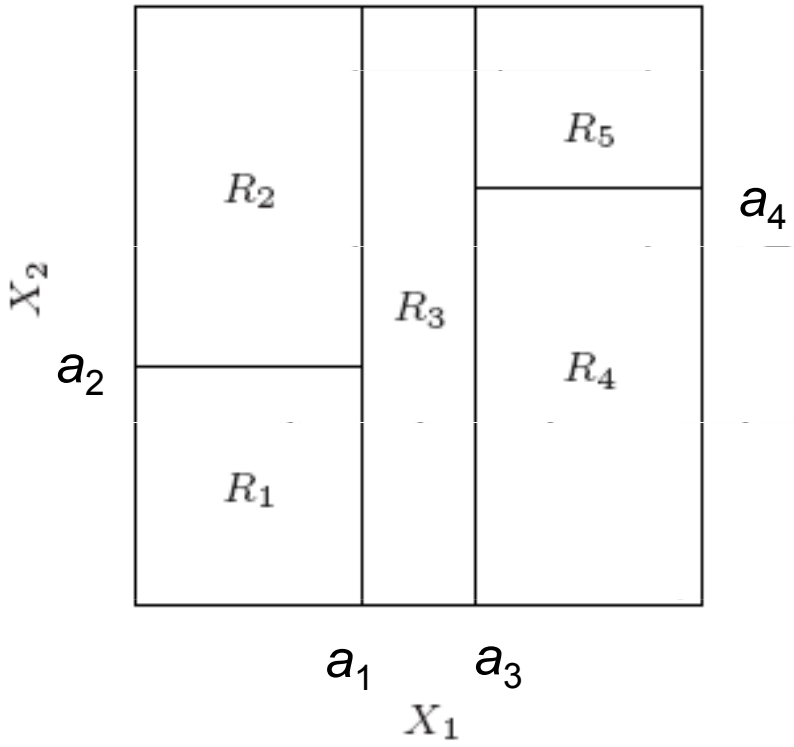
Grafické znázornění stromu CART

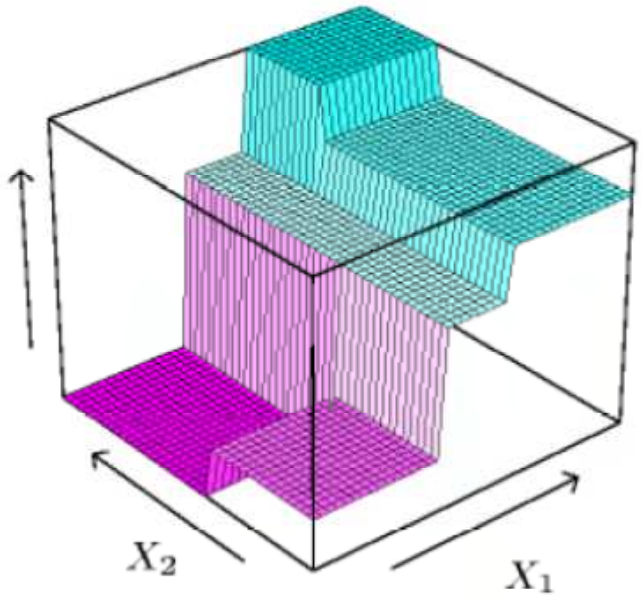
- rozdělení pozorování do kategorií A a B závisle proměnné *Y* s použitím dvou spojitých prediktorů X_1, X_2





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: $$GI = \sum_{c=1}^{J} p_{tc}(1 - p_{tc}) = 1 - \sum_{c=1}^{J} p_{tc}^2$$
- Entropie: $$H = -\sum_{c=1}^{J} p_{tc} \log_2 p_{tc}$$
- Klasifikační chyba: $$ME = 1 - \max\{p_{tc}\}$$

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

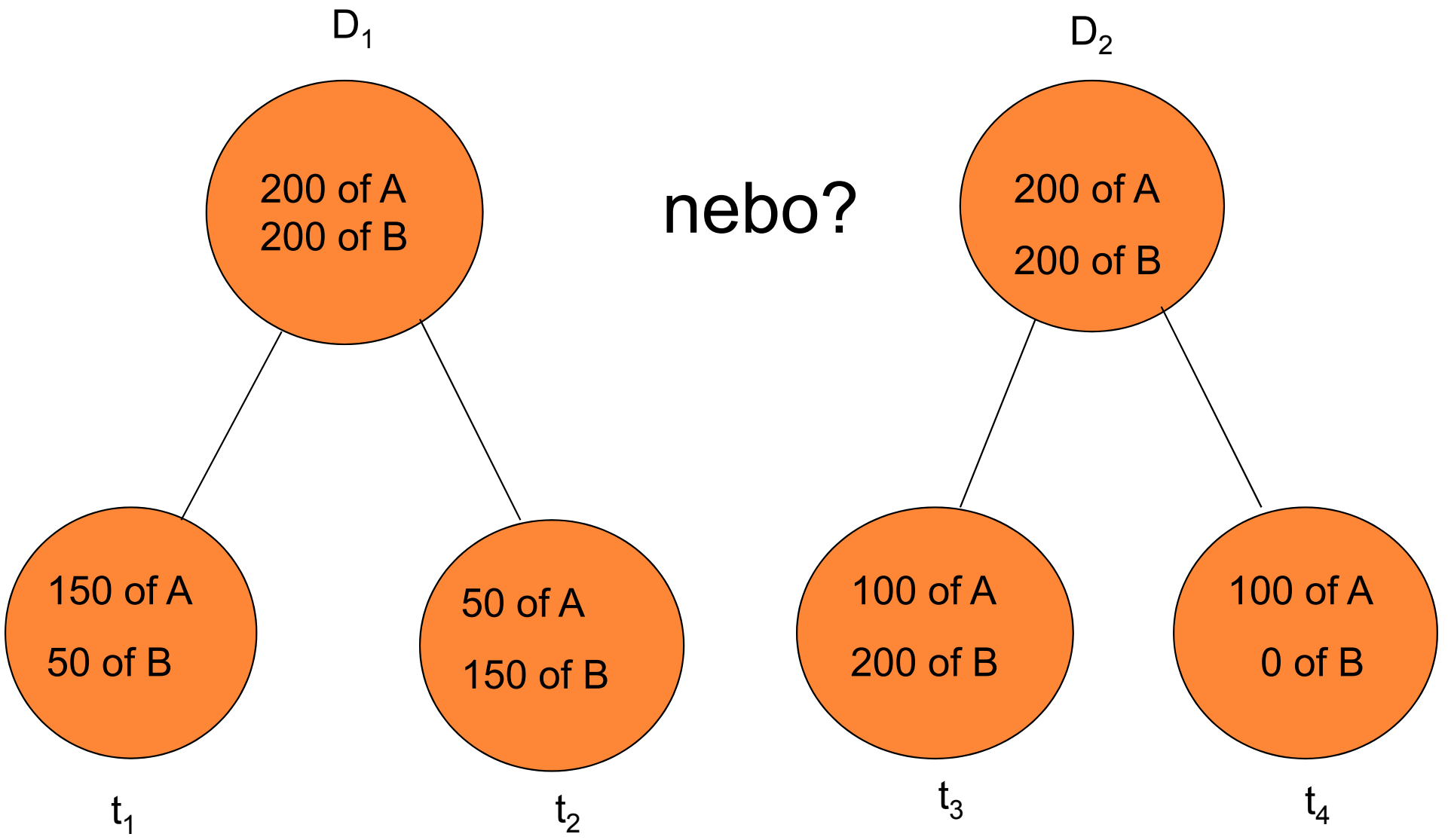
$$ME_{celk} = \sum_{i=1}^{k} \frac{N_i}{N_t} \, ME(i)$$

- 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

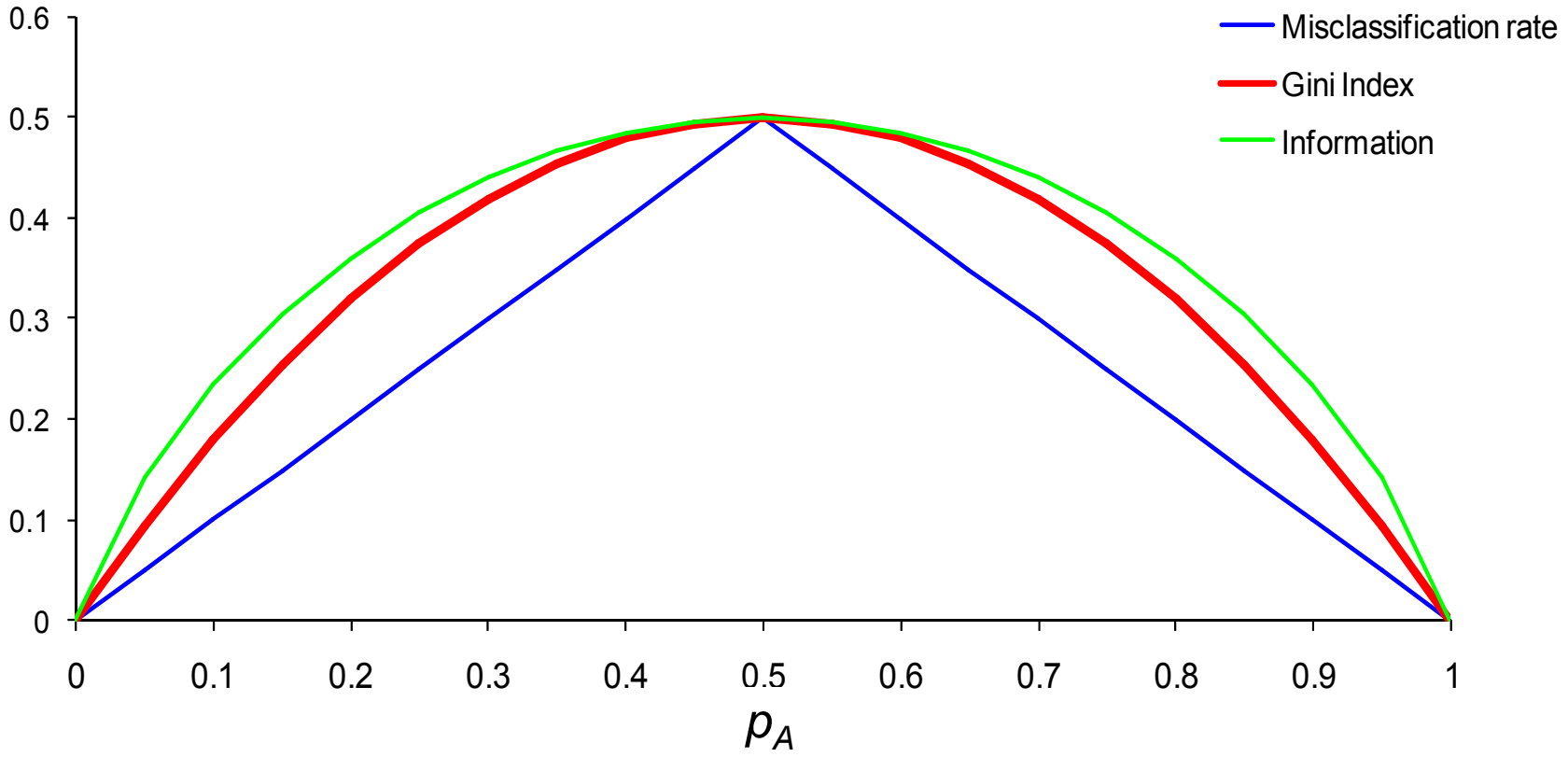
- Klasifikační chyba je obvykle používána k finálnímu měření přesnosti klasifikačního stromu, proto je logické její použití jako kriteriální statistiky

- 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

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).

Přiřazení hodnoty terminálnímu uzlu

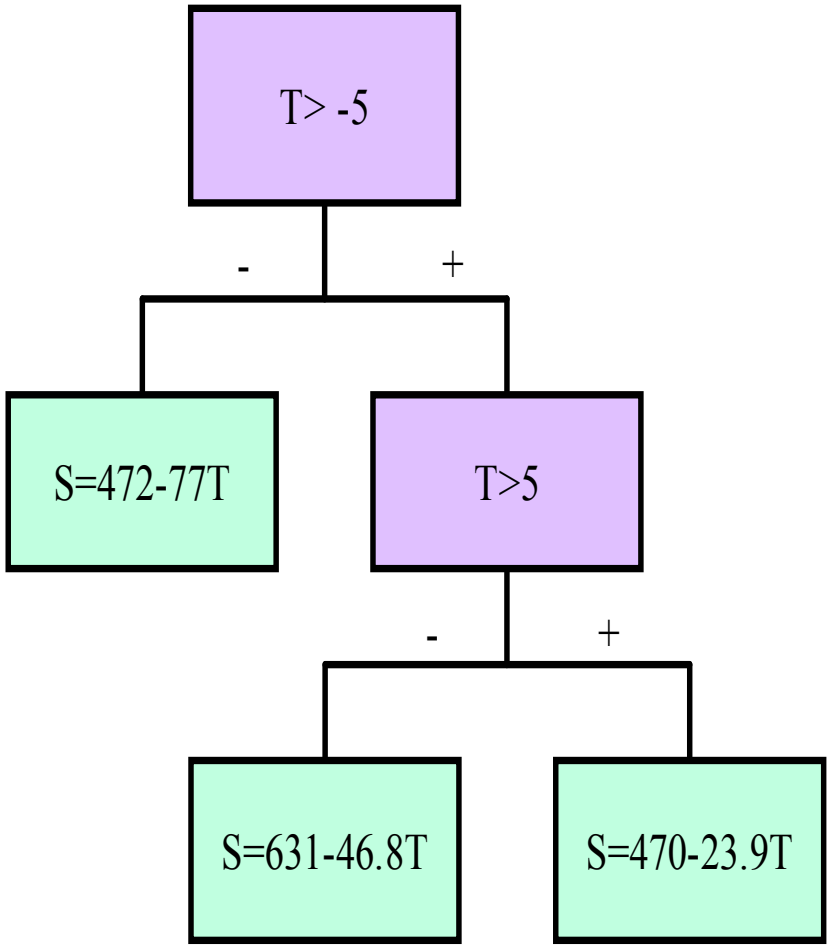
- klasifikační strom - každému uzlu, včetně kořenového, je přiřazena výsledná kategorie závisle proměnné
- výsledná kategorie - má v daném uzlu největší zastoupení
- nové pozorování je klasifikováno podle kategorie uzlu, do kterého je stromem zařazeno
- může se stát, že po rozdělení do dvou terminálních 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

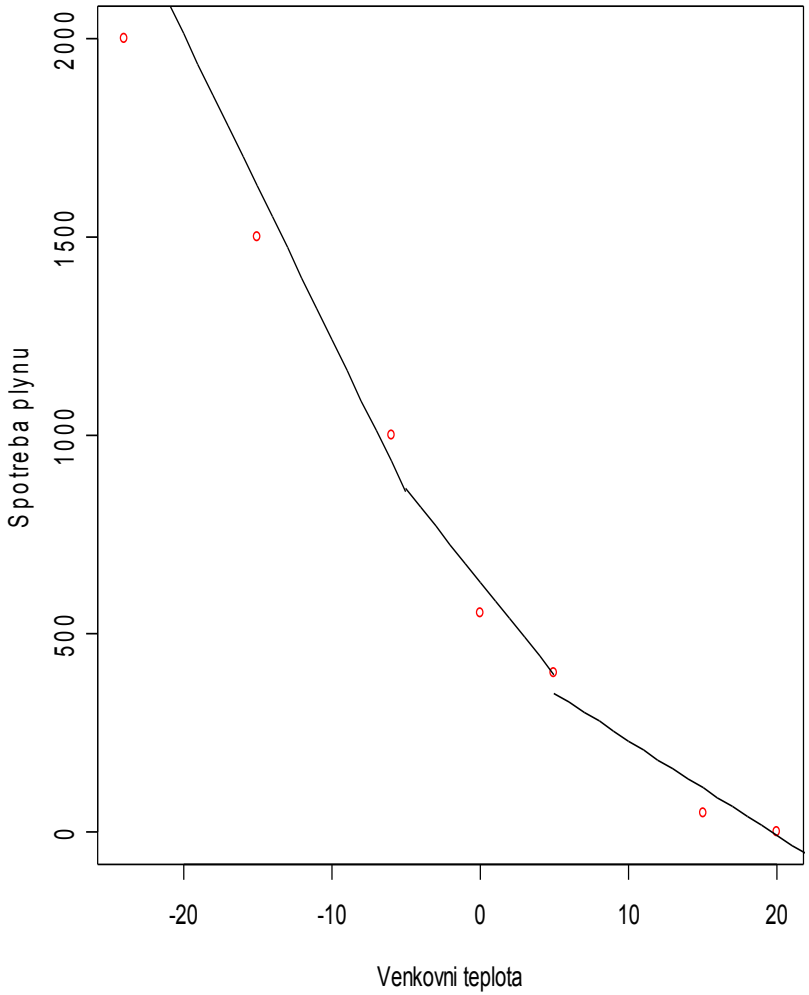
Výsledná hodnota predikce – regresní strom

- 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: Ukázka regresního stromu

Závislost spotřeby plynu na venkovní teplotě

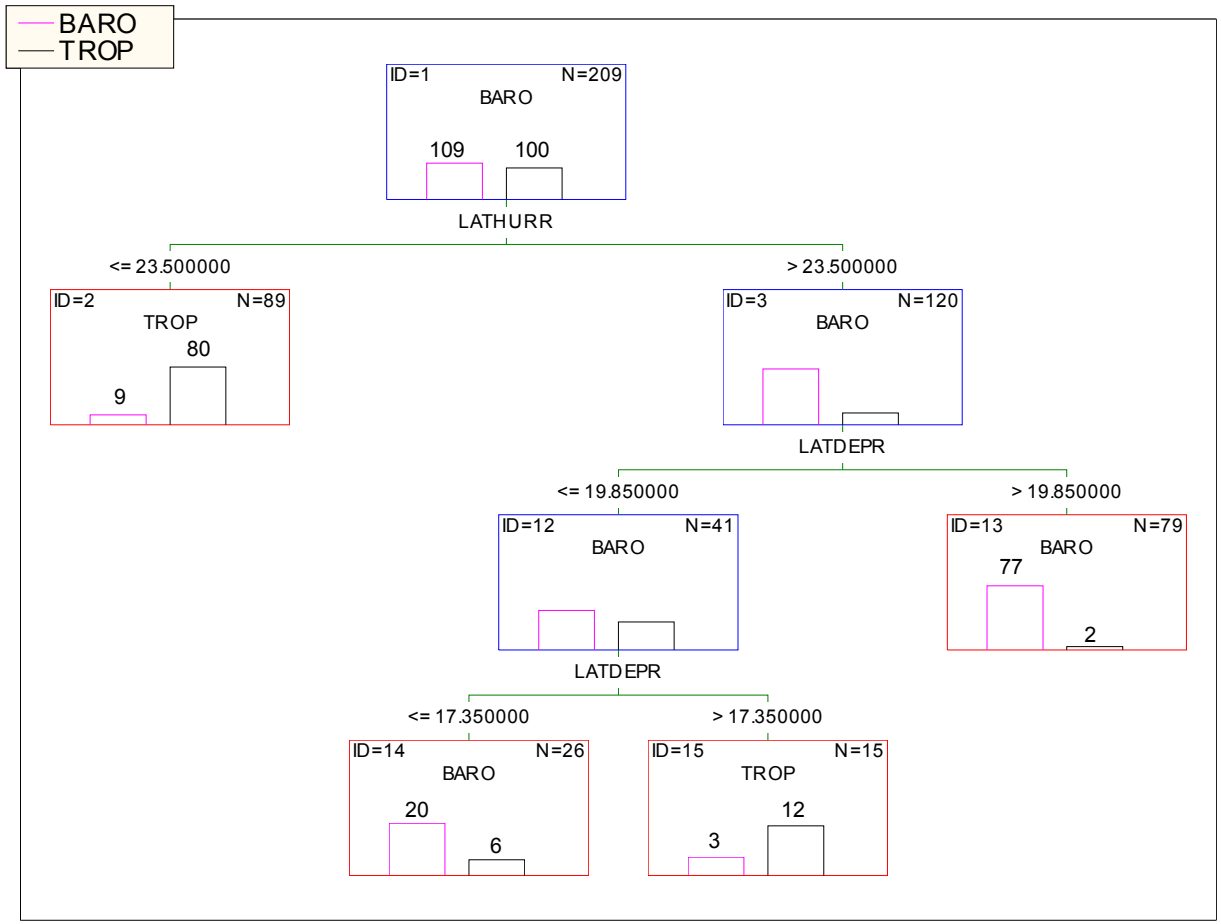




Příklad hurikány

- 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 ?

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.