

Pokročilé neparametrické metody

Klára Komprdová



evropský
sociální
fond v ČR



EVROPSKÁ UNIE



MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY



OP Vzdělávání
pro konkurenceschopnost



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

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

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 - *spojité, diskrétní*

NE Statické modely - *nezávislé na čase*

Popisují prostorovou strukturu?

ANO Prostorově heterogenní - *diskrétní, spojitě*

NE Prostorově homogenní modely

Zahrnuje náhodnou složku?

ANO Stochastické modely

NE Deterministické modely

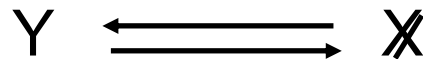
Typy proměnných

- 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

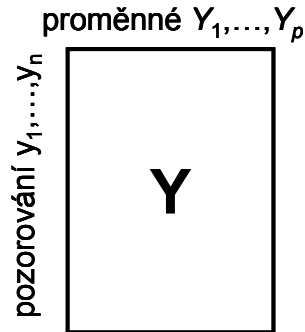


Rozdělení stochastických metod

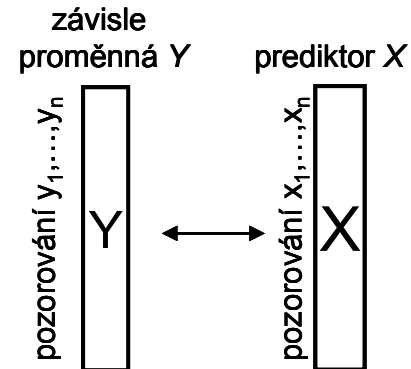
- Parametrické x Neparametrické
 - *Parametrické* – předpoklady o rozdělení dat
 - Klasické lineární modely, zobecněné lineární modely, lineární diskriminační analýza
 - *Neparametrické* – nemají předpoklady o rozložení dat
 - Rozhodovací stromy, lesy, neuronové sítě...
 - *Semiparametrické* – *Zobecněné 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

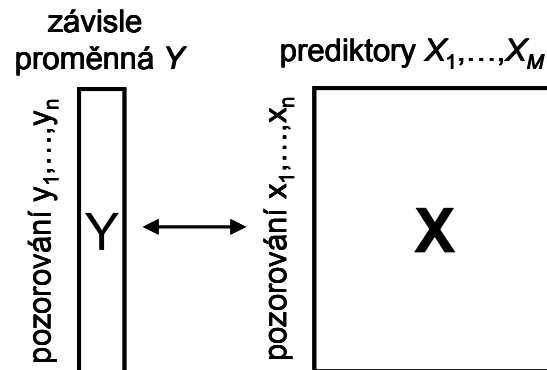
nepřímé ordinační techniky,
shlukovací metody
(vícerozměrné)



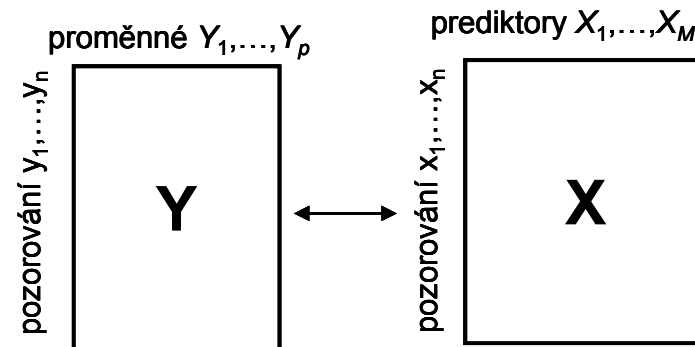
regrese, klasifikace
(jednorozměrná)



regrese a klasifikace
(vícerozměrná)



přímé ordinační techniky
(vícerozměrné)



Z jiného pohledu - živočichové x rostliny x proměnné prostředí



Procesově orientované modely (deterministické)

X



Stochastické modely

X

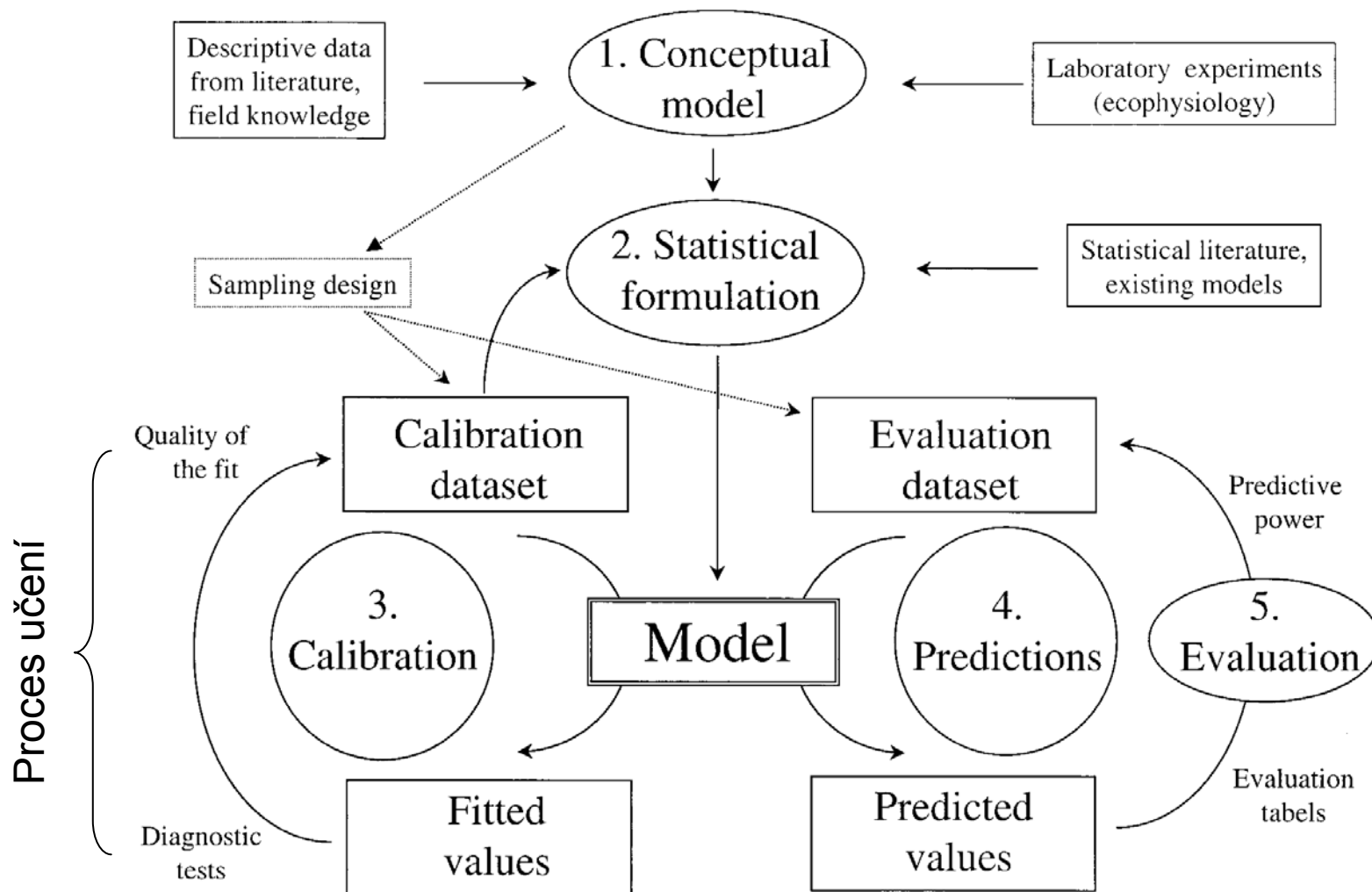


Procesově orientované modely (deterministické)
Stochastické modely, interpolační

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	●	●	●	●	●	●	●	●	●
Interpreovatelnost 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, jackknifing

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



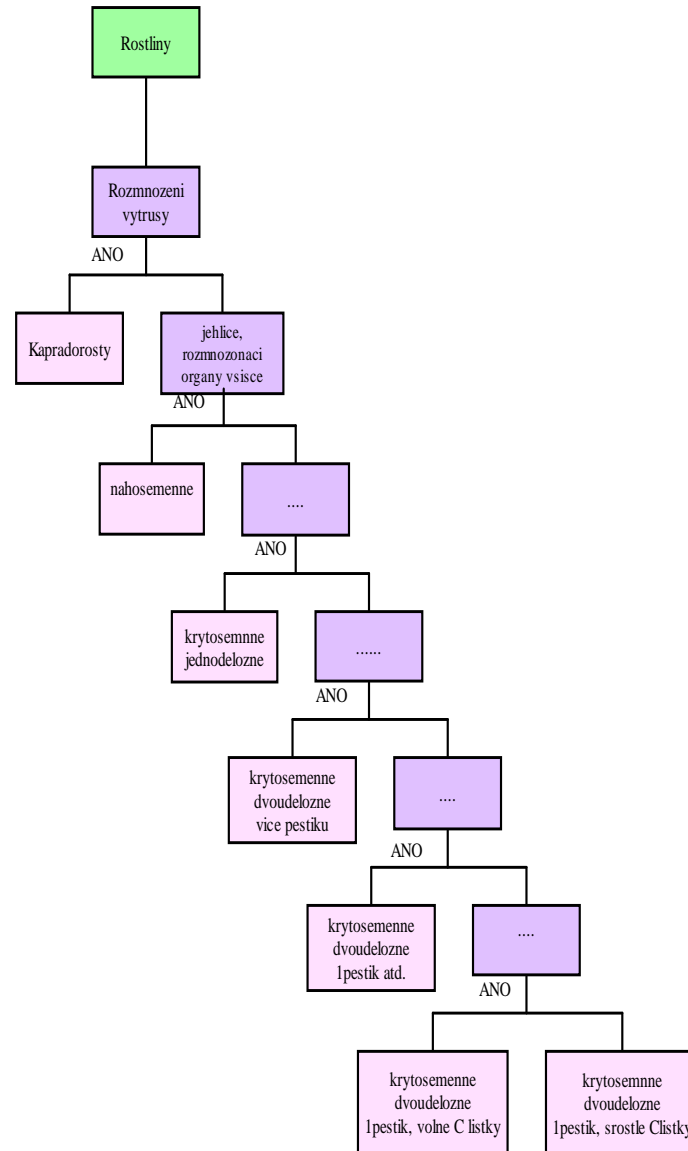
Regresní a klasifikační stromy

(Regression and Classification Trees)

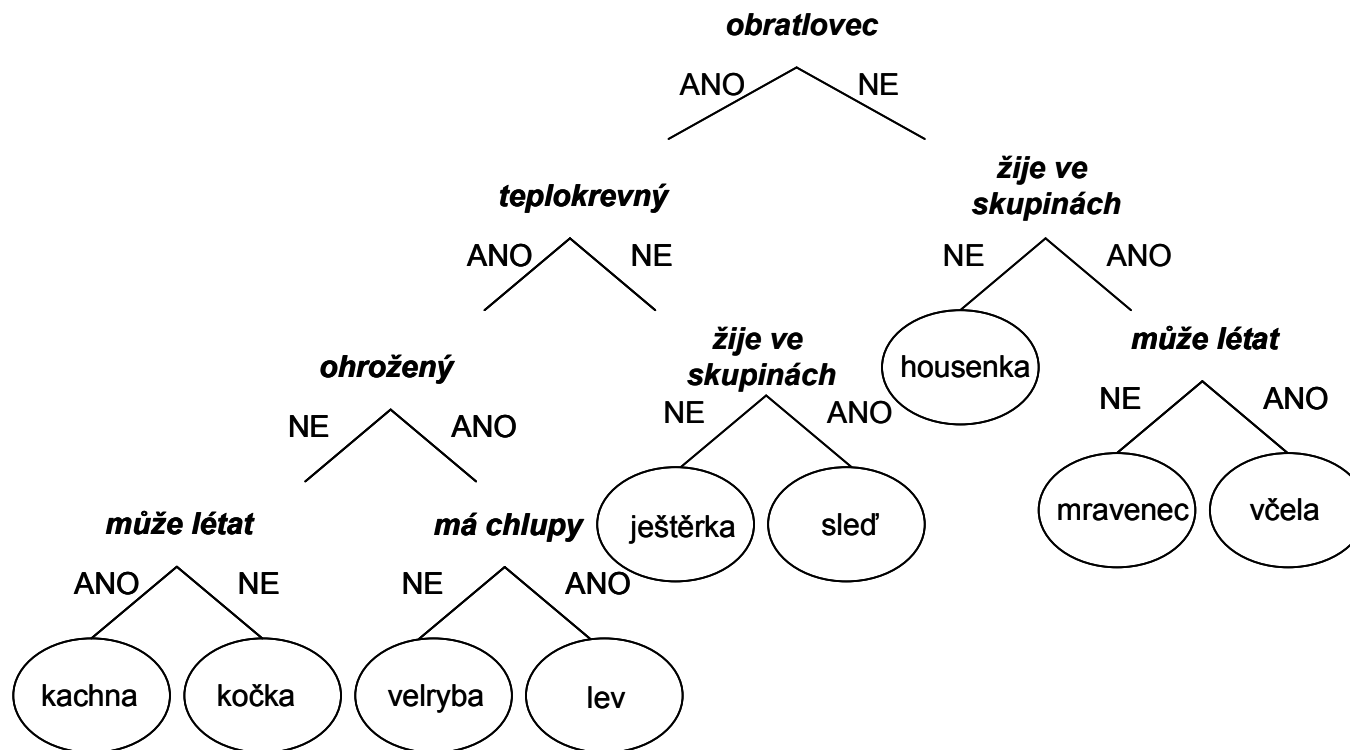
- jsou nejméně formální a nejméně parametrickou skupinou **statistických modelů**
- model – popisuje vzájemné vztahy mezi pozorovanými veličinami
- sada hierarchicky uspořádaných rozhodovacích pravidel
- se stromovou strukturou se setkáváme poměrně často, neboť je přehledná a snadno interpretovatelná - 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

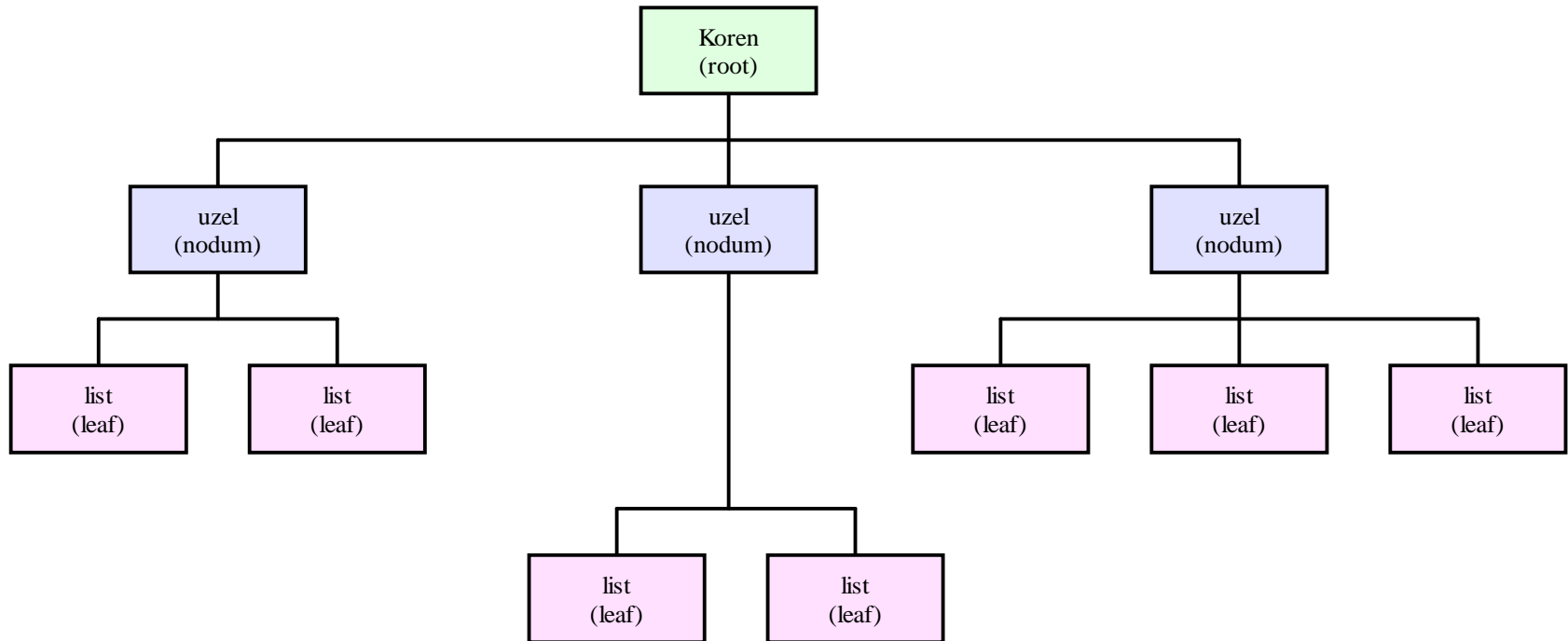


Rozdělení živočichů podle vlastností



	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

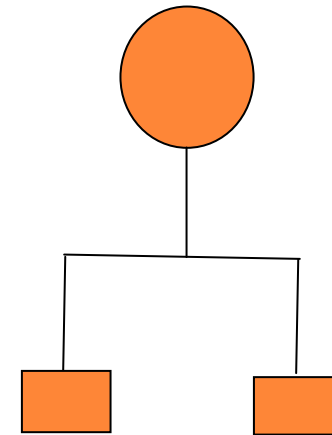
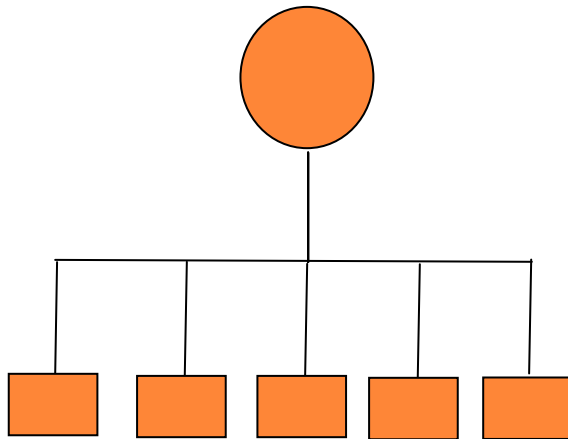


- rozhodovací strom se skládá z **kořene a uzlů** - v každém neterminálním uzlu se strom větví
 - **uzly**
 - **Terminální**
 - **Neterminální (list)**
 - **Mateřské x dceřiné**

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



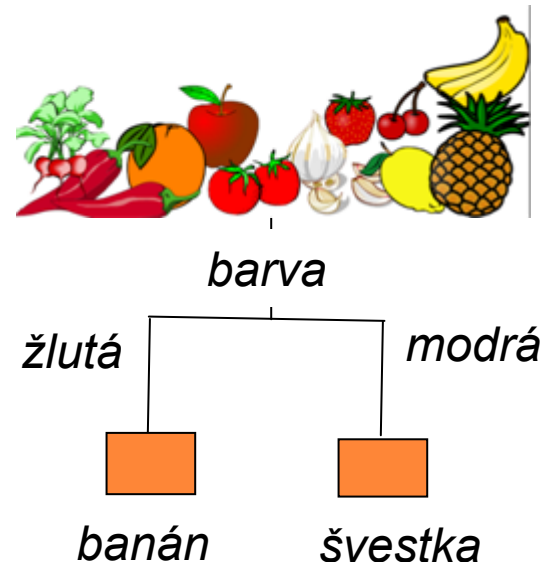
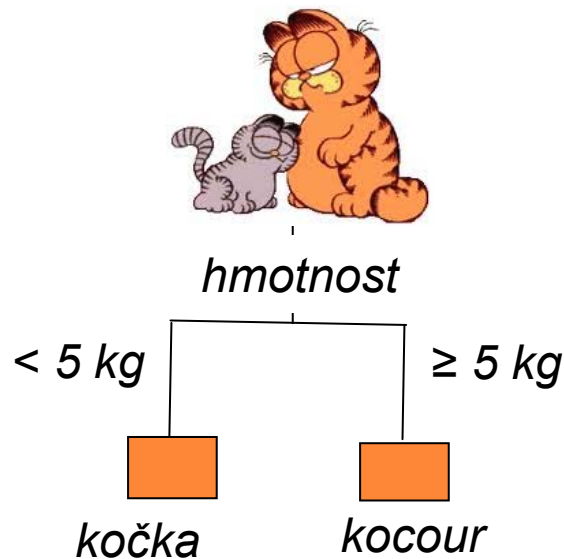
Regresní a klasifikační strom

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

- **klasifikační strom** - pozorování kategoriální závisle proměnné Y s J kategoriemi jsou zařazeny do některé z kategorií $c = (c_1, \dots, c_J)$, 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 $Y = (y_1, \dots, y_n)$, 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

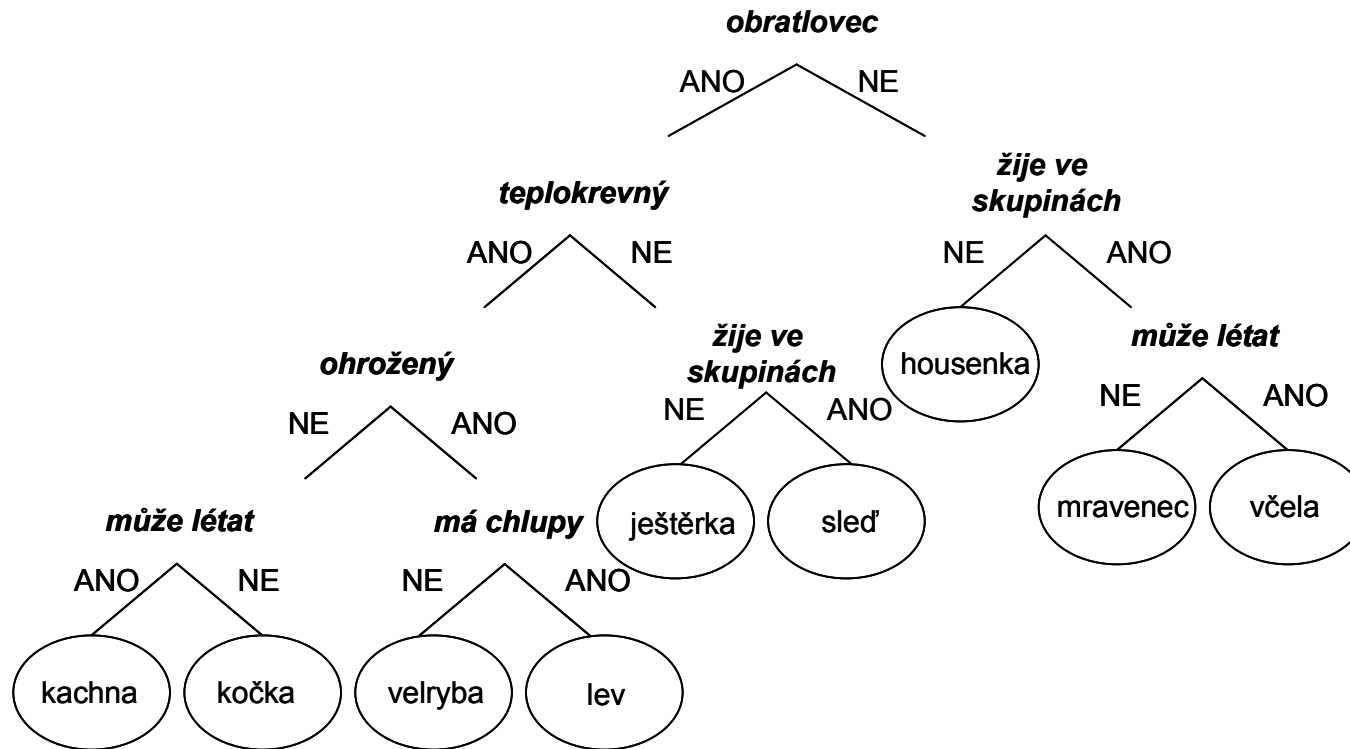
- Pozorování proměnné Y jsou rozdělena do uzlů hodnotami vysvětlujících proměnných (prediktorů) X_1, \dots, X_M .
- Rozdělení je znázorněno graficky pomocí větví stromu.
- Pokud jsou prediktory kategoriální, hodnoty y_i jsou rozděleny podle kategorií prediktoru X - odpovídáme na otázku, které pozorování y_i patří do množiny kde $x_i \in A$, přičemž A je neprázdňá 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 !**

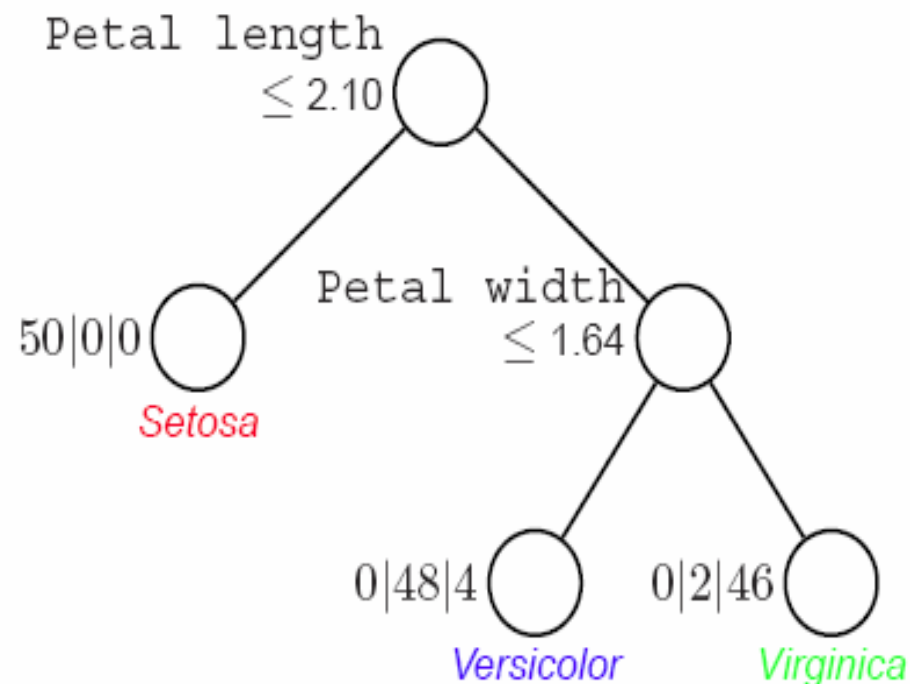
Typ stromu? Typ prediktoru?



	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

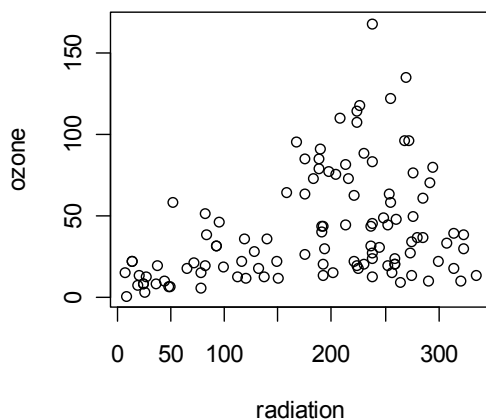
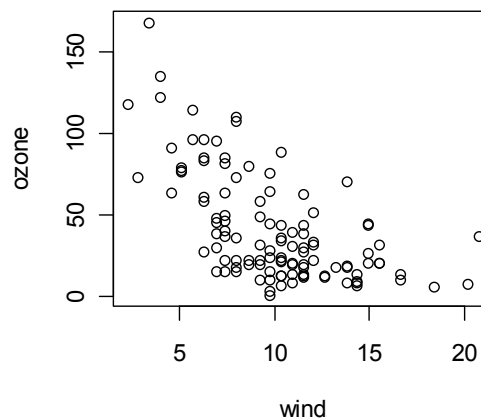
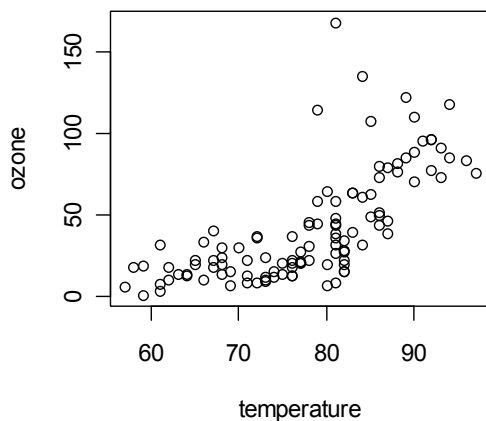
Př: Rozhodovací strom pro kosatce

- 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



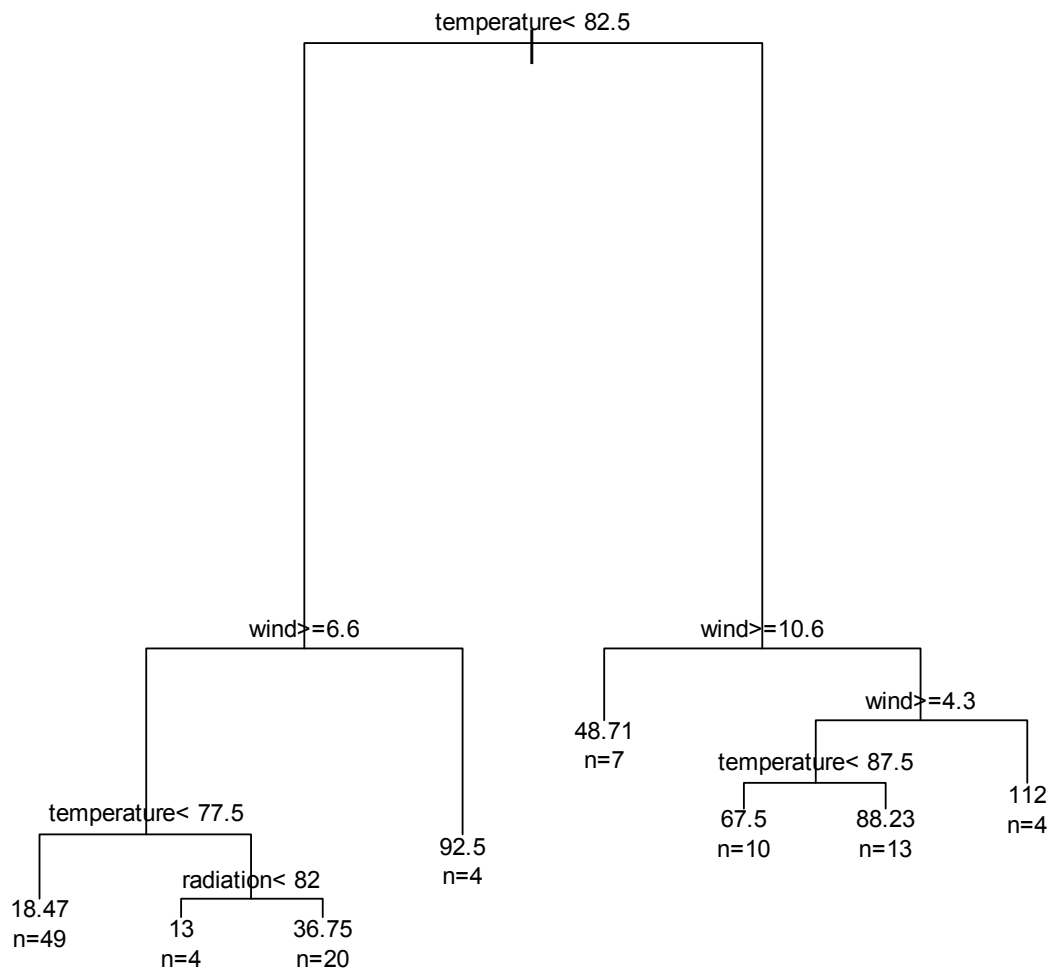
Příklad -ozón

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