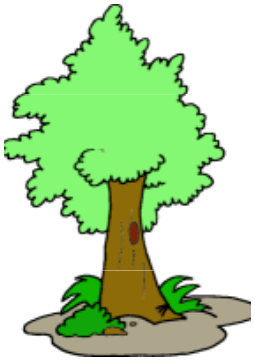
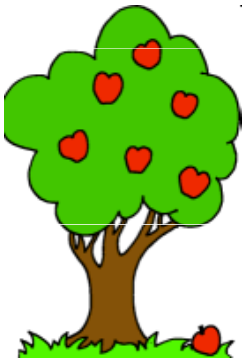
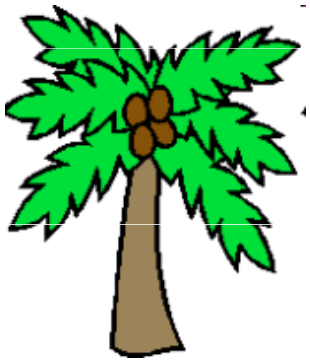
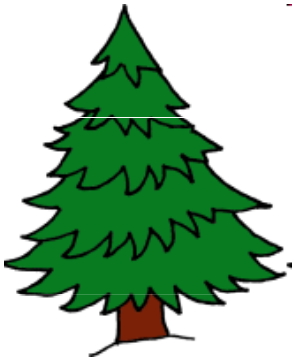
Pokročilé neparametrické metody

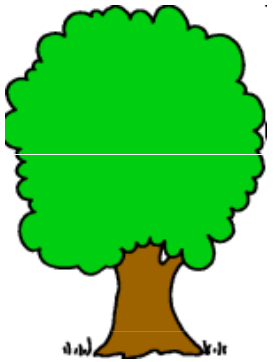
Klára Komprdová









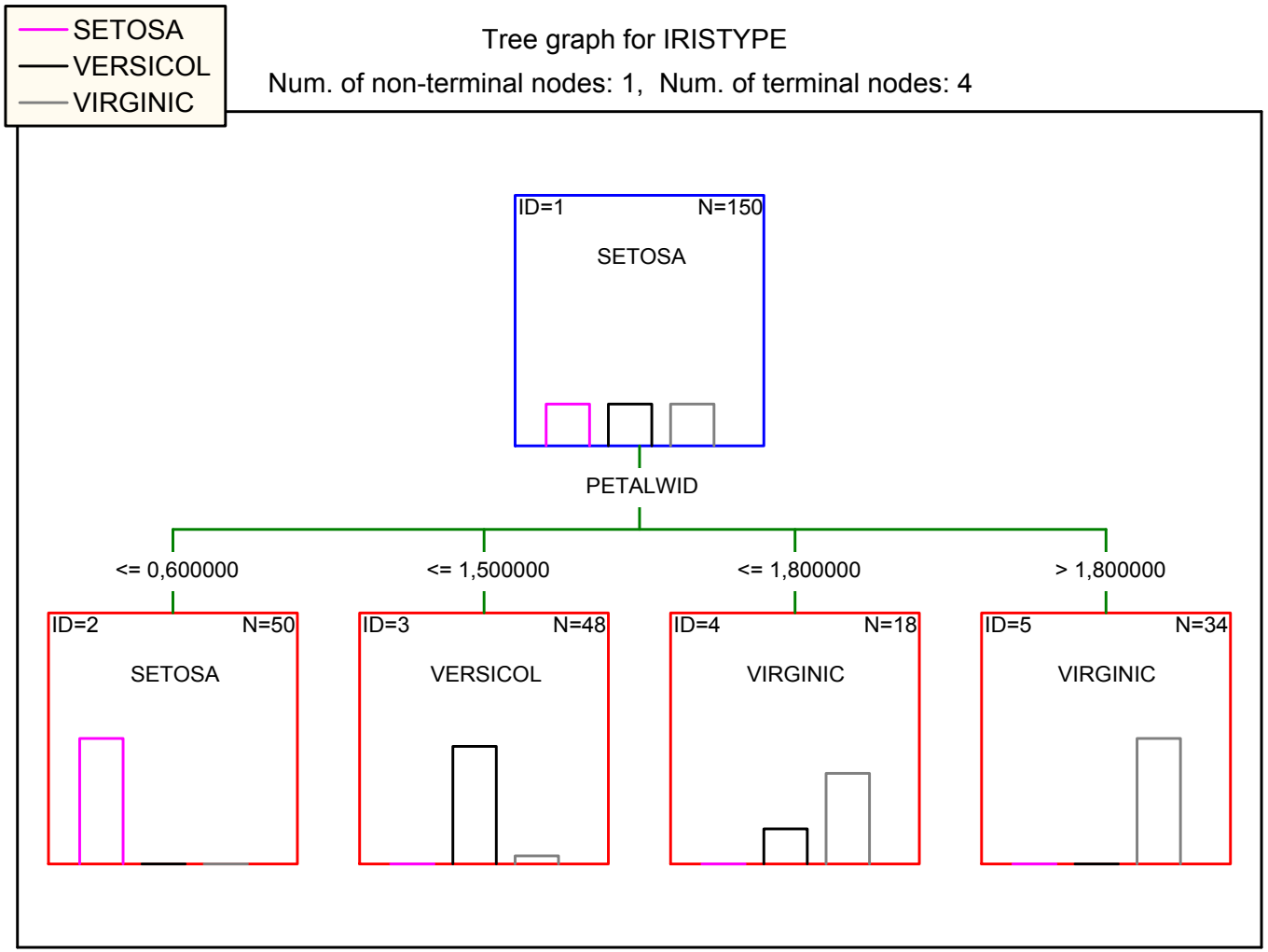


Další typy stromů
CHAID, PRIM, MARS

CHAID - Chi-squared Automatic Interaction Detector

- G.V.Kass (1980)
- Strom pro kategoriální proměnné→ převod spojitých proměnných na ordinální
- Je často využíván v komerčních sférách, především v marketingu a průzkumech veřejného mínění, má ale použití i v přírodovědných oborech.
- nebinárního typu
  - Po prvním dělení nemusí zbývat dostatek pozorování na vytvoření dalších „pater“ stromu →vhodnější pro větší datové soubory.
- Jako kriteriální statistika pro větvení se používá χ^2 –test.

Příklad - kosatce



χ^2 –test - opakování

- χ^2 –test je použit pro zjištění nezávislosti v kontingenční tabulce, která je tvořena kombinací kategorií závisle proměnné a prediktoru
- Jsou-li *Y* a *X* nezávislé, má testová statistika přibližně Pearsonovo χ^2 rozdělení s $u = (r-1)(s-1)$ stupni volnosti, kde *r* je počet řádků a *s* je počet sloupců v kontingenční tabulce.
- Nezávislost v kontingenční tabulce znamená, že se obě proměnné navzájem neovlivňují v hodnotách, které nabývají.
- Hypotéza nezávislosti jevů je zde nulovou hypotézou H_0.
- Pearsonův χ^2 –test je často označován jako test dobré shody.

Kontingenční tabulka

		kategorie prediktoru X				
kategorie Y	*j* / *i*	**1**	**2**	**…**	***s***	**Celkem**
	1	p_{11}	p_{12}	…	p_{1s}	R_1
	2	p_{21}	p_{22}	…	p_{2s}	R_2
	…	…	…	…	…	…
	r	p_{r1}	p_{r2}	…	p_{rs}	R_r
	Celkem	S_1	S_2	…	S_s	n

porovnáváme očekávané četnosti v kontingenční tabulce s jejich skutečnými četnostmi

χ^2 –test

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(p_{ij} - o_{ij}\right)^2}{o_{ij}}$$

$$o_{ij} = \frac{R_i S_j}{n}$$

- kde *i* a *j* je označení řádků (resp. sloupců) v kontingenční tabulce, p_{ij} je pozorovaná frekvence, o_{ij} očekávaná frekvence, *n* je celkový počet pozorování, R_i je počet pozorování v řádku *i*, S_j je počet pozorování ve sloupci *j*.

Příklad - Rozdělení semen dvou příbuzných rostlin podle barvy a tvaru

- Bylo zkoumáno celkem 160 semen dvou druhů příbuzných rostlin. Semena byla roztříděna do následujících kategorií: žluté/hladké; žluté/vrásčité; zelené/hladké; zelené/vrásčité

	žluté/hladké	žluté/vrásčité	zelené/hladké	zelené/vrásčité	Celkový součet
Druh1	10	25	10	15	60
Očekávaný počet					
Druh2	20	30	20	30	100
Očekávaný počet					
Celkový součet	30	55	30	45	160

Algoritmus růstu stromu CHAID

- **Krok1:** pro každý prediktor X_i: Vytvoř kontingenční tabulku kategorií závisle proměnné a prediktoru.

- **Krok 2:** mohou nastat tři případy:
  - Pokud je počet kategorií prediktoru > 2, utvoří se dvojice z kategorií prediktoru→ kategoriální x ordinální. Najde se taková dvojice, která si je co do hodnot závisle proměnné *Y* nejvíce podobná →dvojice, jejíž χ^2 - test má nejvyšší *p* hodnotu.
  - Pokud má prediktor 2 kategorie → algoritmus pokračuje krokem 5
  - Pokud má prediktor *X* pouze jednu kategorii → *p* hodnota je nastavena na 1

- **Krok 3:** Dvojice s nejvyšší *p* hodnotou, která není statisticky významná nebo větší než *alpha2*, se sloučí do jedné skupiny.
  - u ordinálního prediktoru se spojují pouze sousední kategorie
  - u kategoriálního jsou dvojice vytvořeny kombinací všech kategorií.

  Prediktor *X* je dále používán s novými již sloučenými kategoriemi
  - Pokud je i po sloučení počet kategorií > 2, algoritmus se vrátí do kroku 2. Pokud ne, algoritmus pokračuje krokem 4 nebo 5.

- Pozn: *alpha2,* 3 a 4 jsou hodnoty zadané uživatelem

Algoritmus růstu stromu CHAID

- **Krok 4:** Sloučené kategorie mohou být zpětně rozděleny. Jestliže se nově vytvořené skupiny kategorií skládají ze tří nebo více původních kategorií, najde se nejlepší binární rozdělení mezi sloučenými kategoriemi (s nejnižší *p* hodnotou). Pokud je *p* hodnota významná nebo větší než *alpha3*, dojde k rozdělení.

- **Krok 5:** Každá kategorie, která má velmi málo pozorování (minimum je definováno uživatelem), je spojena s nejpodobnější kategorií (opět určeno na základě největší *p* hodnoty)

  pozn: toto nastavení je volitelné a bývá dostupné jen v některých softwarech.

Výše popsaným postupem jsme získali optimální sloučení pro každý prediktor.

- **Krok 6:** V posledním kroku je spočítána adjustovaná *p* hodnota $\chi 2$ testu pro sloučené kategorie každého z prediktorů pomocí Bonferroniho korekce. Vybere se prediktor s nejmenší adjustovanou *p* hodnotou nebo hodnotou větší než *alpha4*. Tento prediktor s optimálně sloučenými kategoriemi je použit k rozdělení uzlu. Pokud významný prediktor nelze nalézt, uzel se již dále nedělí a je považován za terminální.

Algoritmus růstu stromu CHAID – ilustrační příklad

- Zajímá nás klasifikace potravních strategií druhů makrozoobentosu podle různých kategorií nadmořské výšky. Pro jednoduchost se budeme zabývat pouze jedním prediktorem.

Krok 1

Kontingenční tabulka -v buňkách by byly počty jednotlivých druhů

	N-nížinné	S - střední	P - podhorské	H - horské
sběrači				
spásači				
filtrátoři				
dravci				

Algoritmus růstu stromu CHAID – ilustrační příklad

- Pro každou podtabulku je spočítán Pearsonův χ^2 -test nezávislosti. Najdeme největší *p* hodnotu testu, pokud není signifikantní (menší než zvolené α), kategorie spojíme. Protože je nadmořská výška ordinální parametr, můžeme sloučit pouze vedlejší kategorie.

Krok 2 a 3

	N	S
sběrači		
spásači		
filtrátoři		
dravci		

χ_1^2 $p = 0{,}01$

	S	P
sběrači		
spásači		
filtrátoři		
dravci		

χ_1^2 $p = 0{,}05$

	P	H
sběrači		
spásači		
filtrátoři		
dravci		

χ_1^2 $p = 0{,}1$

	N	S	P + H
sběrači			
spásači			
filtrátoři			
dravci			

Algoritmus růstu stromu CHAID – ilustrační příklad

Test sloučených kategorií:

Opět spočítáme Pearsonův χ^2-test nezávislosti pro každou podtabulku, nyní již sloučených kategorií. Obě *p* hodnoty byly statisticky významné pro zvolené α=0,05 a k dalšímu sloučení již nedochází. Přecházíme rovnou do kroku 6, neboť jsme získali optimální sloučení prediktoru → krok 4 a 5 není v našem příkladu potřeba.

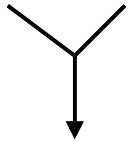
Krok 2 a 3

	N	S
sběrači		
spásači		
filtrátoři		
dravci		

χ_1^2 p = 0,01

	S	P + H
sběrači		
spásači		
filtrátoři		
dravci		

χ_1^2 p = 0,001



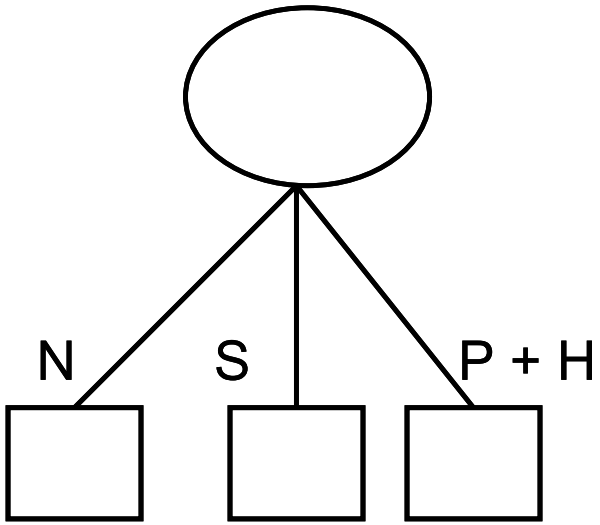
	N	S	P + H
sběrači			
spásači			
filtrátoři			
dravci			

***p**B**

Algoritmus růstu stromu CHAID – ilustrační příklad

- Finální rozdělení uzlu:
  - Za předpokladu, že je nadmořská výška prediktorem s nejnižší adjustovanou *p* hodnotou, původní uzel obsahující celý datový soubor bude rozdělen na tři dceřiné uzly, podle sloučených kategorií nadmořské výšky.

Krok 6



Bonferroniho korekce

- V algoritmu dochází k současnému testování více hypotéz →v našem příkladu bylo třeba učinit celkem čtyři testy pro možné sloučení kategorií.
- Při mnohonásobném testování však vzrůstá pravděpodobnost, že zamítneme nulovou hypotézu H_0, přestože platí.
- Počet prováděných testů u metody CHAID roste s počtem kategorií závisle proměnné a prediktorů.
- Použitím Bonferroniho korekce je možné zmírnit vliv mnohonásobného testování a získat porovnatelné *p* hodnoty pro jednotlivé prediktory s různým počtem kategorií.
- Výsledná *p* hodnota pro kontingenční tabulku kategorií závisle proměnné a optimálně sloučeného prediktoru je vynásobena *B* koeficientem, čímž získáme adjustovanou *p* hodnotu pro daný prediktor.

Bonferroniho korekce - Koeficient *B*

- ordinální proměnná → slučování sousedních kategorií
- kategoriální proměnná→ slučování všech možných kombinací

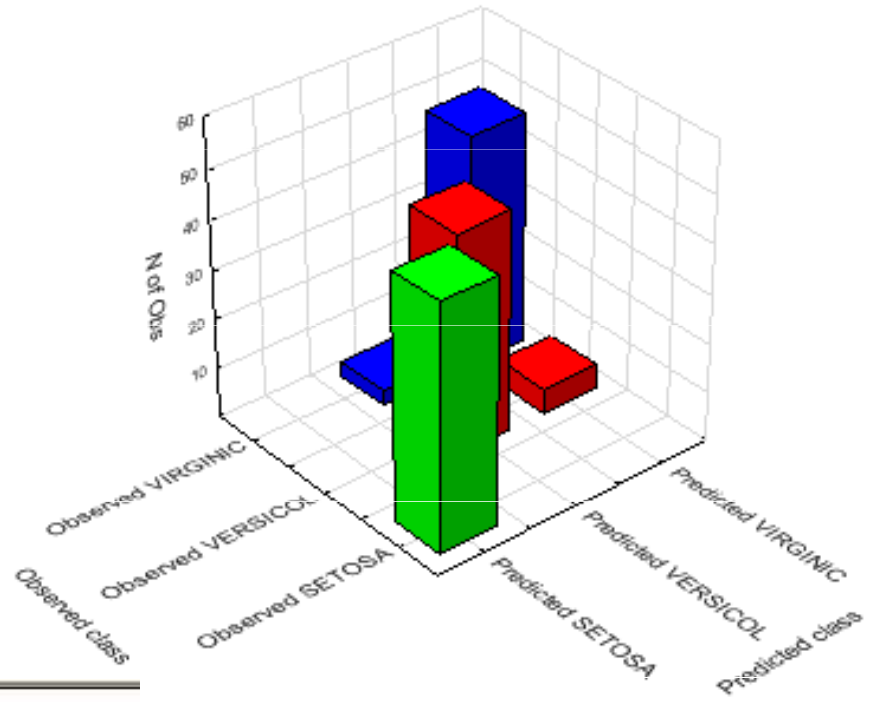
$$B_{ordinal} = \binom{s-1}{r-1}$$

$$B_{kategorial} = \sum_{i=0}^{r} (-1)^i \frac{(r-i)^s}{i!(r-i)!}$$

- kde *r* je počet řádků a *s* je počet sloupců kontingenční tabulky kategorií závisle proměnné a prediktoru.

Příklad- kosatce

Classification matrix 1
Dependent variable: IRISTYPE
Options: Categorical response, Analysis sample

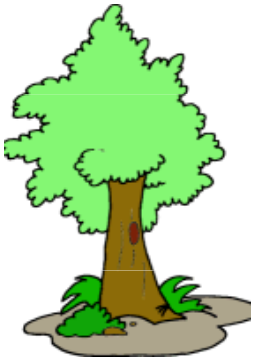


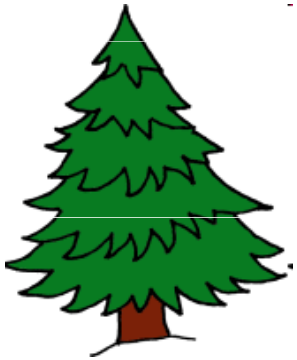
	Classification matrix 1 (Irisdat) Dependent variable: IRISTYPE Options: Categorical response, Analysis sample				
	Observed	Predicted SETOSA	Predicted VERSICOL	Predicted VIRGINIC	Row Total
Number	SETOSA	50			50
Column Percentage		100.00%	0.00%	0.00%	
Row Percentage		100.00%	0.00%	0.00%	
Total Percentage		33.33%	0.00%	0.00%	33.33%
Number	VERSICOL		45	5	50
Column Percentage		0.00%	93.75%	9.62%	
Row Percentage		0.00%	90.00%	10.00%	
Total Percentage		0.00%	30.00%	3.33%	33.33%
Number	VIRGINIC		3	47	50
Column Percentage		0.00%	6.25%	90.38%	
Row Percentage		0.00%	6.00%	94.00%	
Total Percentage		0.00%	2.00%	31.33%	33.33%
Count	All Groups	50	48	52	150
Total Percent		33.33%	32.00%	34.67%	

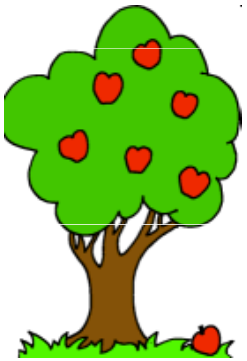
Strom CHAID

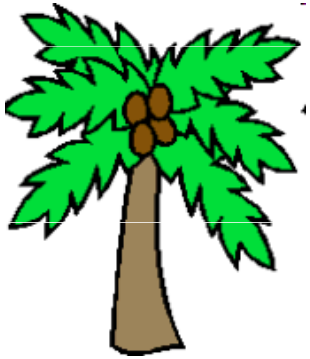
- Růst stromu se zastaví, pokud je dosaženo následujících pravidel:
  - není možné nalézt žádné významné rozdělení.
  - Všechna pozorování závisle proměnné v uzlu mají stejnou hodnotu nebo identickou hodnotu pro každý prediktor.
  - Pokud je dosaženo uživatelem definovaných nastavení, která se týkají:
    - parametrů velikosti stromu jako je nastavení počtu terminálních uzlů nebo větví;
    - počtu pozorování v uzlu, které je menší než minimum stanovené uživatelem nebo počtu pozorování, které by po rozdělení vedlo k dceřiným uzlům s menším počtem pozorování, než je definováno uživatelem.

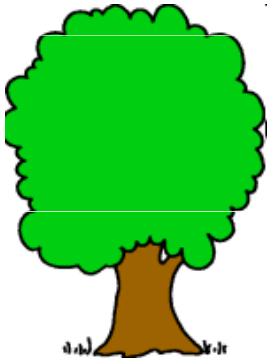
- Celkovou správnost stromu OA_{kateg} určujeme stejně jako v případě stromu CART. K odhadu obecné chyby $e(t)$ je možné opět použít k-testovacích souborů z krosvalidace.









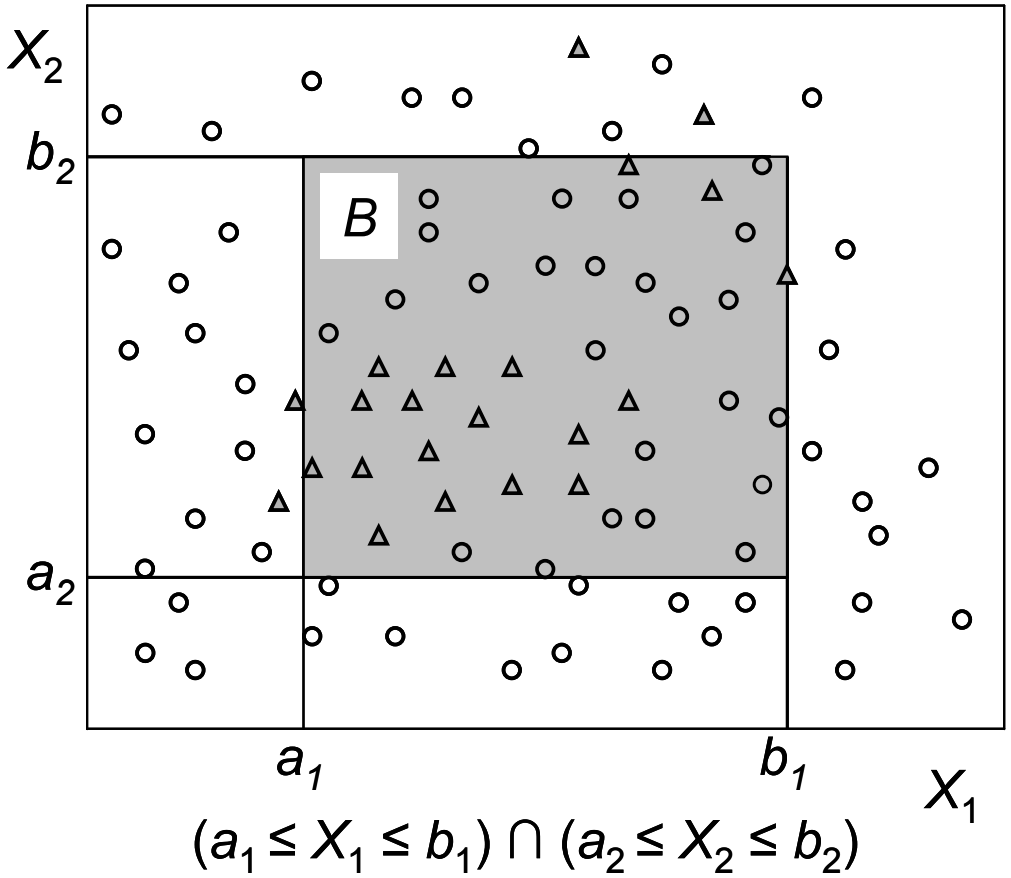


PRIM - Patient Rule Induction Method

PRIM - Patient Rule Induction Method

- **PRIM** (Friedman & Fisher, 1999) - metoda primárně určena pro regresi.
- PRIM podobně jako ostatní rozhodovací stromy rozděluje pozorování závisle proměnné Y pomocí hodnot prediktorů do uzlů $t_1,..,t_N$, → označovaných jako okna $B_1,\ldots, B_K$
- Graficky můžeme okna znázornit jako jednotlivé regiony v prostoru prediktorů $X_1,\ldots, X_M$.
- V případě metody PRIM se však vyhledávají takové regiony, ve kterých je průměr hodnot závisle proměnné Y nejvyšší (nebo nejnižší).
- Výsledkem je sada jednoduchých pravidel, která definují jednotlivá okna a rozdělují pozorování závisle proměnné

PRIM



Mějme 100 pozorování. Závisle proměnná Y označuje presenci $y_i = 1$ (trojúhelníky) nebo absenci $y_i = 0$ (kolečka) určitého druhu rostliny. Pro jednoduchost uvažujme pouze dva spojité prediktory: teplotu X_1 a srážky X_2. Rostlina bude přítomna s větší pravděpodobností v podmínkách daných rozsahem prediktorů $(a_1 \leq X_1 \leq b_1) \cap (a_2 \leq X_2 \leq b_2)$, které jsou zde znázorněny pomocí okna B.

PRIM - algoritmus

- 1. Soubor se rozdělí na testovací a trénovací (v poměru zadaném uživatelem). Seřadí se hodnoty prediktorů od nejmenší po největší. Okno obsahuje celý datový soubor (trénovací)

- 2. Okno se zmenšuje podél jedné hrany o malé množství pozorování (často α=0.1 nebo α=0.05) – tak aby výsledný průměr ve zmenšeném okně byl co největší (nejmenší)

Krok 1 a 2 se opakuje dokud okno neobsahuje předem stanovené minimum pozorování (např. 10)

- 3. dochází k reverznímu procesu-okno je zpětně rozšiřováno do všech směrů, ale jen pokud se zvýší průměr v okně

Z kroku 1-3 se získá sekvence oken o různém počtu pozorování

- 4. použije se krosvalidace k vybrání optimálního okna B_1- testovací soubor!
- 5. Odstraní se vzorky z okna B_1 -pozorování která jsou odstraněna z okna mají nejvyšší (nejnižší) hodnoty prediktoru X_j

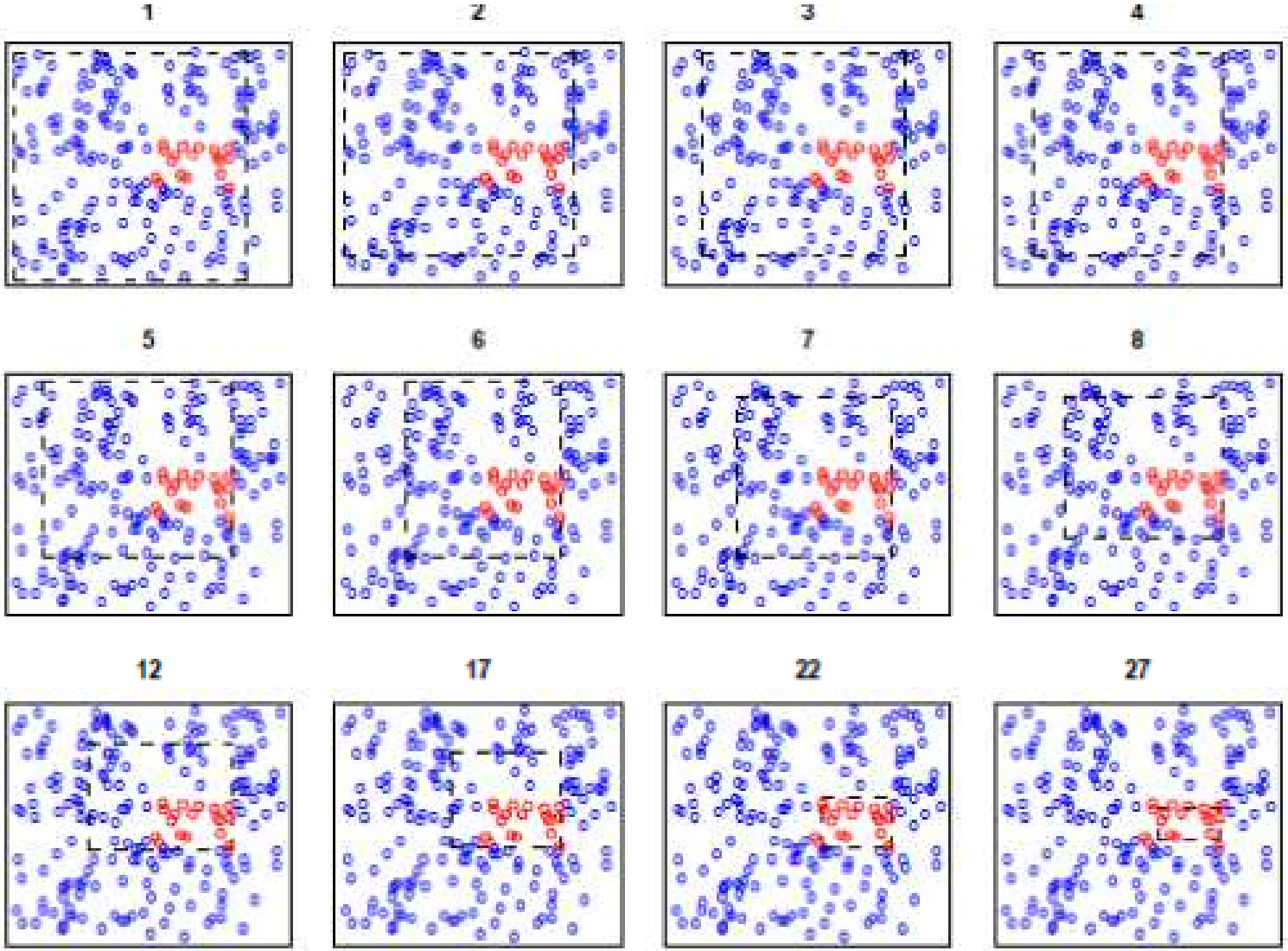
Krok 2-5 se opakuje, dokud není dosaženo konečného počtu oken $B_1, B_2, \ldots B_K$

- Okna jsou dána rozhodovacími pravidly
- Stejně jako v CART lze použít kategoriální prediktor

PRIM - algoritmus

- 200 bodů, rovnoměrně rozdělených do jednotkového čtverce
- Závisle proměnná Y má hodnotu 1 (červená barva) pokud je $0.5 < X_1 < 0.8$ a $0.4 < X_2 < 0.6$
- Závisle proměnná Y má hodnotu 0, modrá barva
- Proporce bodů o které se okno posune $\alpha = 0.1$

PRIM - algoritmus



(Hastie et. al, 2009)

Algoritmus je hierarchický a používáme krosvalidaci

PRIM

- Stejně jako v CART lze použít kategoriální prediktor
- Oproti CART je výhodou, že se probere větší škála pravidel a můžeme najít optimální řešení
- Nevýhoda- není k dispozici stromová struktura → okna jsou dána rozhodovacími pravidly

- PRIM je velmi vhodný pro případy, kdy nás zajímá nalezení skupin v datech s nejvyšší nebo nejnižší hodnotou závisle proměnné – např. při různých ochranných opatření, kdy výsledky mohou sloužit ke stanovení vhodné velikosti území podle pravděpodobnosti výskytu druhu nebo ke zjištění klimatických podmínek, při kterých dochází k největšímu znečištění ovzduší