Příklad I - Regresní strom

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

package *rpart* –knihovna pro CART

```
> library(lattice) /načtení knihovny - soubor
> library(rpart)  /načtení knihovny - CART
//> soubor<-read.delim("clipboard",row.names=1) /načtení souboru kopírováním

> data(environmental)
> summary(environmental) /popisná statistika souboru
> summary(environmental$ozone) /popisná statistika parametru Ozón
> names(environmental) /názvy proměnných
> dim(environmental) /dimenze matice
> hist(environmental $promenna) /histogram
> par(mfrow=c(2,2)) /vytvoření pole pro grafy
> plot(zavisla~ promenna,data= environmental) /graf – závislost y na x
> strom1<-rpart(ozone~.,data= environmental,minsplit=7,minbucket=3)
        / výstup zapsán do souboru strom1
        / vzorec v modelu pojmenovává vysvětlovanou proměnnou a použití tečky na
        pravé straně značí, že všechny ostatní proměnné v souboru jsou prediktory
        / minsplit – min. počet pozorování, při kterém nedojde k oddělení do dalšího
        uzlu (pozor na hledání odlehlých hodnot)
        /minbucket – počet pozorování v koncovém uzlu, kdy se již strom dále nedělí

> plot(strom1) /zobrazí strom
> text(strom1,cex=0.67) /zobrazení popisu parametrů, cex – zmenšení velikosti
znaků
> plot(strom1,margin=0.05) /zobrazí strom; margin – zvětšení okraje stromu, aby se
vešly popisky

> plot(strom1,uniform=T,margin=0.05) /stejná délka větví (T=True)
> text(strom1,cex=0.67,use.n=T) / zobrazí počet pozorování ve výsledném uzlu
```

Validace

Funkce *rpart* automaticky krosvalidaci provádí (můžeme zrušit nastavením argumentu xval=0, který určuje hodnotu *k*)

```
>plotcp(strom1,col="blue") /výsledek krosvalidace
>printcp(strom1) /výsledek krosvalidace v textové podobě

> par(mfrow=c(1,2))
> rsq.rpart(strom1) /výsledek krosvalidace II
```

prořezávání

```
> strom2<-prune(strom1,cp=0.029) /cp je potřeba vybírat z grafu!
> par(mfrow=c(1,2))
> plot(strom1);text(strom1,cex=0.5)
> plot(strom2);text(strom2,cex=0.5)
> par(mfrow=c(1,1))
> strom2 /strom v textové podobě
> summary(strom2) /Zástupné a primární proměnné (surogáty)
```

Příklad II – klasifikační strom (převzato z MRM – Šmilauer)

Datový soubor v package MASS

```
> data(shuttle,package="MASS")
> summary(shuttle)
```

Jde o reálná data, představující shromážděná doporučení expertní komise při přípravě letů raketoplánu. Tato doporučení popisují okolnosti, za kterých by měla posádka nechat přistávací manévr na počítači (faktor use má pak hodnotu ***auto***) a za kterých má být přistání provedeno ručně (use s hodnotou ***noauto***). Rozhodování je ovlivněno stabilitou raketoplánu na předem naplánované dráze sestupu (stability), velikostí (error) a směrem (sign) odchylky od dráhy, směrem větru v místě přistání (wind – čelní resp. v zádech), sílou větru (magn) a také viditelností v oblasti přistání (vis). Ačkoliv poskytnutá data popisují vpodstatě všechny možné kombinace podmínek (28 = 256 kombinací), pro jejich efektivní použití, a také pro ověření jejich konzistentnosti, byl tento návod převeden do soustavy pravidel

```
> strom3<-rpart(use~.,data=shuttle,minsplit=2,minbucket=1)
> plotcp(strom3)

> plot(strom3,margin=0.05)    /grafické znázornění
> text(strom3,cex=3/4, use.n=T)
> strom3      /textové znázornění
> text(strom3,cex=3/4, use.n=T, pretty = 0)
```

Příklad III - PRIM

Datový soubor Boston. Obsahuje údaje z domácností z 506 měst v Bostonu. Jako prediktory použijeme proměnné koncentrace oxidů dusíku (μg) (nox) a průměrný počet pokojů v domě (rm). Závisle proměnná je kriminalita (průměrná na osobu) (crim). Zajímá nás charakteristika lokalit s vysokou kriminalitou.

```
> library(prim)
> library(MASS)
> data(Boston)
> x <- Boston[, 5:6]
```

```
> y <- Boston[, 1]

> boston.prim <- prim.box(x = x, y = y, threshold.type = 1)
```

původní nastavení *prim.box* jsou
- peeling quantile: peel.alpha=0.05
- pasting is carried out: pasting=TRUE
- pasting quantile: paste.alpha=0.01
- minimum box mass (proportion of points inside a box): mass.min=0.05
- threshold is the overall mean of the response variable y
- threshold.type=0

threshold.type: threshold direction indicator: 1 = ">= threshold", -1 =
"<= threshold", 0 = ">= threshold[1] & <= threshold[2]"

Protože nás zajímají jen oblasti s vyšší kriminalitou, nastavení je threshold.type=1

```
> summary(boston.prim, print.box = TRUE)
```
Zobrazí se tři sloupečky: the box mean, the box mass, and the threshold type.
Hvězdička obsahuje zbytek datového souboru, na který již nebyl použit algoritmus.

```
> plot(boston.prim, col = "transparent")
```