Příklad I: Regresní strom CART

Ukázkový příklad ke cvičení v programu R.

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 [7][8].

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.

```
> library(lattice) /načtení knihovny - soubor
> library(rpart)  /načtení knihovny - CART

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

Načteme datový soubor a zobrazíme zdrojovou tabulku:

```
> data(environmental) /načtení souboru
> environmental
> names(environmental) /názvy proměnných
> dim(environmental) /dimenze matice
> summary(environmental) /popisná statistika celého souboru
```

Stejným způsobem lze získat výstup pro každou proměnnou:

```
> summary(environmental$ozone) /popisná statistika parametru Ozón
```

Podíváme se na rozložení jednotlivých proměnných.

Pokud budou hodnoty odlehlé – transformace proměnných? Rozdělíme si pole na čtyři plochy pomocí příkazu *par*, do kterých se zobrazí histogramy proměnných. Histogramy vytvoříme pomocí funkce *hist*:

```
> par(mfrow=c(2,2))
//> hist(environmental $promenna)
> hist(environmental $ozone, main = paste("histogram ozón"), xlab =
   "koncentrace ozónu", ylab = "počet měření", cex.lab=1.2)
```

Vytvoříme regresní strom, který zapíšeme do proměnné *strom_ozon*. Nastavíme parametry funkce *rpart*, jako je *minsplit* - minimální počet pozorování, při kterém nedojde k oddělení do dalšího uzlu (pozor na hledání odlehlých hodnot) a *minbucket* – minimální počet pozorování v terminálním uzlu. Pokud jeden z těchto parametrů není zadaný, druhý z nich je nastaven: *minbucket = minsplit/3; minulit = 3* minbucket.*

Dalším důležitým parametrem je *cp*, který určuje kritérium složitosti

a parametr *xval*, což je počet podsouborů použitých při krosvalidaci neboli *k*.

Ostatní parametry se vztahují k hledání počtu kompetitivních a zástupných proměnných pro každý uzel: *maxcompete* a *maxsurrogate*, defaultně jsou nastaveny na nulu.

```
> strom_ozon<-rpart(ozone~.,data= environmental,minsplit=10,minbucket=5,
cp =0.01)
```

Pomocí příkazu *plot* strom zobrazíme.

Parametr *use.n = T* zobrazí počet pozorování ve výsledných uzlech, *cex* určuje velikost znaků a *margin* je parametr pro zvětšení/zmenšení okraje grafu:

```
> plot(strom_ozon,margin=0.05);text(strom_ozon, cex=1,use.n=T)
```

Délka větví výsledného stromu ukazuje variabilitu vyčerpanou dělením. V případě regresního stromu je to residuální suma čtverců, v případě stromu klasifikačního suma Gini indexů. Dělící hodnoty proměnných a příslušné operátory se vztahují vždy k levé části stromu. Je proto užitečné zobrazit si strom ještě v textové podobě, kterou obdržíme jednoduše zadáním jeho názvu:

```
> strom_ozon
```

Ve výsledcích je nejdříve zobrazena proměnná, která byla vybrána pro dělení, následuje hodnota rozdělení, počet pozorování v uzlu, rozptyl v daném uzlu a jeho hodnota (průměrná koncentrace ozónu). Hvězdičky označují terminální uzly, jejichž výsledné hodnoty nás zajímají nejvíce.

Validace

Krosvalidaci nastavíme ve funkci *rpart* pomocí argumentu *xval*, který určuje hodnotu *k*. Pomocí funkce *plotcp* zobrazíme závislost velikosti stromu na chybě z krosvalidace. Jde o závislost geometrických průměrů z intervalů hodnot *cp* na chybě testovacích souborů při krosvalidaci. Hodnota *cp* odpovídá α/T_1 z rovnice kritéria složitosti stromu a chyba na testovacím souboru je rovna $1\text{-}R^2_{test}$.

Můžeme specifikovat parametr *minlin*, který zobrazí (*T*) nebo nezobrazí (*F*) vodorovnou referenční čáru a parametr *upper*, který umožňuje měnit popis horní osy na počet uzlů (*size*), počet dělení (*splits*) nebo bez popisu (*none*).

```
>  strom_ozon1<-rpart(ozone~.,data= environmental, minsplit=10,
minbucket=5, cp =0.01, xval=5)
> plotcp(strom_ozon1, minlin=T, cex.axis=1.2, cex=1.5, cex.lab=1.2,
upper =  'splits')
```

Vodorovná čára je minimální krosvalidovaná chyba plus 1*SE* (standardní chyba odhadu).

Výsledek krosvalidace zobrazíme také v textové podobě pomocí funkce *printcp* nebo *rsq.rpart*. Použitím funkce *rsq.rpart* navíc získáme dva grafy, první zobrazuje hodnoty *e*(*t*) pro trénovací a testovací soubor při různé velikosti stromu a druhý je totožný s grafem z funkce *plotcp*, ovšem bez referenční čáry.

```
> par(mfrow=c(2,1))
> rsq.rpart(strom_ozon1)
```

Prořezání

K prořezání stromu použijeme funkci *prune*. Optimální hodnotu *cp* vybereme z grafu z funkce *plotcp* nebo z textového výstupu funkce *rsq.rpart*.

```
> strom_ozon2<-prune(strom_ozon1,cp=0.02)
```

Nyní zobrazíme do jednoho obrázku původní a prořezaný strom. Parametr *uniform=T* zobrazí stromy se stejnou délkou jednotlivých větví:

```
> par(mfrow=c(1,2))
> plot(strom_ozon1, uniform=T,margin=0.1);text(strom_ozon1,
cex=0.64,use.n=T)
> plot(strom_ozon2, uniform=T,margin=0.1);text(strom_ozon2,
cex=0.64,use.n=T)
```

Zástupné a kompetitivní proměnné

Zadáním příkazu *summary* s názvem vytvořeného stromu získáme, mimo chyby stromu pro testovací a trénovací soubor, také informaci o kompetitivních a zástupných proměnných:

```
> summary(strom_ozon3)
```

parametr *improve* udává procentuální změnu v součtu čtvercových odchylek y_i od průměru (SS_{uzel}) pro dané rozdělení: 1-($SS_{pravý\_uzel}$ + $SS_{levý\_uzel}$)/$SS_{mateřský\_uzel}$). Proměnné jsou tedy uvedeny v pořadí, v jakém by byly na základě kriteriální statistiky $Q_t(T)$ vybrány pro rozdělení mateřského uzlu. Výstup nerozlišuje kompetitivní proměnné, ty se však dají snadno poznat. Pokud je proměnná zároveň uvedená u *surrogate splits*, je zástupná, jinak kompetitivní. Parametr *agree* u zástupné proměnné udává pravděpodobnost, s jakou jsou pozorování rozdělena stejně, jako u primární proměnné.

Příklad II – Klasifikační strom CART

Datový soubor obsahuje informace o pasažérech lodi Titanic, která se potopila v roce 1912 na cestě ze Southamptonu do New Yorku čtyři dny po vyplutí, když narazila do ledovce. Na palubě bylo přes 2200 pasažérů a členů posádky a přežila pouze necelá třetina cestujících. Informace byly shromážděny Britskou obchodní komorou *(British Board of Trade*) při šetření potopení lodi. Celkem 2202 shromážděných záznamů se týkají třídy, kterou cestovali (první, druhá a třetí třída a členové posádky), pohlaví, přežití a věku (pouze rozdělení na dospělé a děti) pasažérů. Je dobře známým faktem, že zejména u přežití žen a dětí hrála důležitou roli třída, kterou cestovaly [13].

```
>titanic<-read.delim("clipboard",row.names=1)
> summary(titanic)

> library(datasets)
> data(Titanic)
> tit<-as.data.frame(Titanic)
> summary(tit)
```