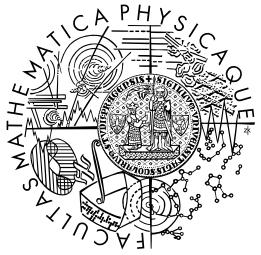


MATEMATICKO-FYZIKÁLNÍ FAKULTA
Univerzita Karlova

BAKALÁŘSKÁ PRÁCE

Martin Dvořák

Bayesovské klasifikační a regresní stromy

Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: prof. RNDr. Jaromír Antoch, CSc.
Studijní program: Matematika
Studijní obor: Finanční matematika

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ....

Podpis autora

Rád bych na tomto místě poděkoval prof. RNDr. Jaromíru Antochovi, CSc., který mě vedl při mé práci. Děkuji za jeho čas a rady. Dále děkuji rodině za pomoc a opravu gramatických chyb. Následně děkuji Lukáši Krylovi za motivaci práci dokončit.

Název práce: Bayesovské klasifikační a regresní stromy

Autor: Martin Dvořák

Katedra: Katedra pravděpodobnosti a matematické statistiky

Vedoucí bakalářské práce: prof. RNDr. Jaromír Antoch, CSc., Katedra pravděpodobnosti a matematické statistiky

Abstrakt: Bakalářská práce se věnuje klasifikačním a regresním stromům, jejich stavbě a interpretaci. V první části se čtenář seznámí se strukturou rozhodovacích stromů, základními definicemi a metodikou. V druhé části jsou představeny pokročilejší a efektivnější metody pro tvorbu takových stromů využívající Bayesovský přístup k celému problému. Poslední část práce je zaměřená na praktickou úlohu, kde jsou využity poznatky z této práce. Celý text je doplněn obrázky, vysvětleními a odvozeními, aby bylo pro čtenáře jednodušší celý problém pochopit více do hloubky. Práce Bayesovské klasifikační a regresní stromy může posloužit všem zájemcům, kteří chtějí blíže poznat problematiku rozhodovacích stromů.

Klíčová slova: klasifikační stromy, regresní stromy, CART

Title: Bayesian classification and regression trees

Author: Martin Dvořák

Department: Department of Probability and Mathematical Statistics

Supervisor: prof. RNDr. Jaromír Antoch, CSc., Department of Probability and Mathematical Statistics

Abstract: The bachelor's thesis is devoted to classification and regression trees, their construction, and interpretation. In the first part, the reader gets acquainted with the structure of decision trees, basic definitions, and methodology. In the second part, more advanced and efficient methods for creating such trees using a Bayesian approach to the whole problem are presented. The last part of the work is focused on a practical task, where knowledge from this work is used. The entire text is accompanied by pictures, explanations, and derivations to make it easier for the reader to understand the whole problem in more depth. The thesis Bayesian classification and regression trees can serve all those interested who want to learn more about the issue of decision trees.

Keywords: classification trees, regression trees, CART

Obsah

Úvod

Tato práce se zaměřuje na binární rozhodovací stromové struktury, neboli klasifikační a regresní stromy. Cílem této práce je seznámit čtenáře s touto problematikou a vysvětlit, jak takové stromy vznikají. Stromová metoda je velmi jednoduše interpretovatelná a používá se k řešení široké škály problémů. Jejich popularita roste díky schopnosti snadno pochopit a vizualizovat výsledky, což umožňuje jak odborníkům, tak i laikům efektivně využívat tyto modely.

Čtenář se seznámí se se strukturou rozhodovacích stromů, jejich tvorbě a s metodami a algoritmy, které jsou k sestavení potřebné. Tyto postupy vycházejí z knihy Classification and regression trees autora Leo Breimana a jeho týmu, která byla publikována v roce 1988 a dala základ moderním metodám tak, jak je známe dnes. Následně se práce zaměřuje na využití Bayesovské statistiky a pokročilejších metod při tvorbě rozhodovacích stromů.

Práce je psaná jednoduchým a pochopitelným jazykem, obsahuje odvození a vysvětlivky tak, aby čtenář co nejvíce porozuměl této problematice. V textu se objevují obrázky, které jsou vytvořeny v programu Mathematica a statistickém programu *R*. Ty přispívají k vizualizaci a lepšímu pochopení textu.

Práce má následující strukturu:

První kapitola se věnuje klasifikačním stromům. Vysvětluje základy stromových struktur a popisuje stavbu stromové struktury pro klasifikační úlohu, kdy predikovaná proměnná má kategoriální tvar.

V druhé kapitole se čtenář seznámí s regresními stromovými strukturami, které přímo vycházejí ze struktur klasifikačních. obsahují ovšem některé úpravy a modifikace, které jsou potřebné, jelikož predikovaná proměnná je numerického tvaru.

Ve třetí kapitole je představen Bayesovský přístup k řešení klasifikačních a regresních úloh. Tento přístup je modernější a poskytuje metody, které tvoří mnohdy přesnější stromové struktury.

Ve čtvrté kapitole je předveden praktický příklad na reálných datech. Je zkonstruován a okomentován klasifikační strom pomocí metod předvedených v této práci.

1. Klasifikační stromy

Obecný klasifikační problém si lze představit následovně: Předpokládejme, že těhotné ženy při příchodu do nemocnice vyplňují dotazník, ve kterém jsou otázky týkající se jejich pohlaví, věku, hmotnosti a jejich zdravotního stavu. Odpovědi na takové otázky mohou být buď numerické[1], nebo kategoriální[2]. Podle odpovědí by lékaři následně chtěli určit, zda porod dané ženy bude komplikovaný, či nikoliv. K tomu by potřebovali vyplněné dotazníky žen, které již porodili, spolu s informací, zda byl nebo nebyl jejich porod komplikovaný, a následně predikovat budoucí případy.

Předpokládejme, že byly položeny 4 otázky, potom vektor odpovědí každé ženy, označený jako $\boldsymbol{x}$, je $\boldsymbol{x} = (x_1,x_2,x_3,x_4)$. Hodnota predikované proměnné y odpovídající porodu je rovná 0, pokud byl porod bezproblémový, a 1, pokud byl komplikovaný. Snadno vidíme, že y je kategoriální proměnná mající dvě třídy. Na takto zavedené struktuře je poté na základě historických dat možné sestavit klasifikátor, který podle vektoru $\boldsymbol{x}$ bude predikovat hodnotu y, a tak bude schopný s určitou pravděpodobností předpovědět, jestli daná žena bude mít komplikovaný porod.

Označme $\mathcal{X}$ jako prostor obsahující všechny možné vektory měření $\boldsymbol{x}$ a předpokládejme, že případy z $\mathcal{X}$ spadají do J různých tříd. Očíslujme třídy $1,2,\ldots,J$ a množinu všech J možných tříd označme C, potom $C = \{1,\ldots,J\}$.

Definice 1 (Klasifikátor)**.** *Klasifikátor je funkce $d(\boldsymbol{x})$ definovaná na X splňující:*

$$d(\boldsymbol{x}) : \boldsymbol{x} \to j,\ \forall \boldsymbol{x} \in \mathcal{X},\ j \in C.$$

Druhý pohled na celý problém lze popsat následovně: Předpokládejme, že

$$A_j = \{\boldsymbol{x} : d(\boldsymbol{x}) = j\},\ j \in J. \tag{1.1}$$

Klasifikátor d klasifikuje případy z A_j, podmnožiny $\mathcal{X}$, do j-té třídy. Z výrazu (1.1) tedy vyplývá, že podmnožiny $A_1,\ldots,A_J$ jsou disjunktní a platí

$$\cup_{j=1}^{J} A_j = \mathcal{X}.$$

Pro samotnou konstrukci klasifikátoru je potřeba mít trénovací data. Trénovací výběr $\mathcal{L}$ je množina N dvojic $(\boldsymbol{x}_i,y_i)$, kde $\boldsymbol{x}_i \in X$, $y_i \in C$, $i = 1,\ldots,N$. Vektor $\boldsymbol{x}_i$ je vektor měření i-tého případu a y_i je skutečná třída tohoto případu. Pak má trénovací výběr $\mathcal{L}$ následující tvar:

$$\mathcal{L} = \{(\boldsymbol{x}_1,y_1),\ldots,(\boldsymbol{x}_N,y_N)\}.$$

[1]Hodnoty numerických proměnných mají samy o sobě nějaký číselný význam, např. věk, výška.

[2]Kategoriální proměnné zpravidla rozdělujeme na ordinální a nominální. V případě ordinálních dat (např. známky ve škole) lze hodnoty přirozeně uspořádat, zatímco v případě proměnných nominálních (např. pohlaví, rasa) tomu tak nelze. Proto kategoriální proměnné nabývají hodnot (tříd) v konečné množině.

1.1 Odhady přesnosti

Pro každý klasifikátor je klíčové, abychom znali alespoň přibližnou přesnost jeho predikce. Pokud má klasifikátor malou přesnost, pak není účinný a může být vhodnější místo něho klasifikovat případy náhodně. Proto se zaměříme na odhady přesnosti. K tomu budeme ovšem potřebovat nejdříve zavést pravděpodobnostní model.

Nechť $R^*(d)$ označuje skutečnou míru chybné klasifikace klasifikátoru $d(\boldsymbol{x})$. Definujme prostor $\mathcal{X} \times C$ jako množinu všech možných dvojic $(\boldsymbol{x},j)$, kde $\boldsymbol{x} \in \mathcal{X}$ a $j \in C$. Poté pravděpodobnost $P[A,\, j]$ definovanou na $\mathcal{X} \times C$ lze interpretovat jako pravděpodobnost, že náhodně vybraný případ z populace má pravděpodobnost $P[A,\, j]$, že jeho vektor měření $\boldsymbol{x}$ je v A a jeho třída je j. Trénovací výběr $\mathcal{L}$ je tedy náhodný výběr N případů vybraných z rozdělení $P[A,\, j]$.

Definice 2 (Míra chybné klasifikace)**.** *Nechť je dán klasifikátor d zkonstruovaný pomocí trénovacího výběru $\mathcal{L}$. Vezměme $(\boldsymbol{X},Y)$, $\boldsymbol{X} \in \mathcal{X}$, $Y \in C$ jako nový výběr z rozdělení pravděpodobnosti $P[A,\, j]$. Platí*

- $P[\boldsymbol{X} \in A,\, Y = j] = P[A,\, j]$ *a*
- $(\boldsymbol{X},Y)$ *je nezávislý na* $\mathcal{L}$*.*

Pak je míra chybné klasifikace klasifikátoru d definována jako

$$R^*(d) = P[d(\boldsymbol{X}) \neq Y|\mathcal{L}]$$

Míra chybné klasifikace $R^*(d)$ je teoretická hodnota, kterou je možné určit pouze za podmínky, že známe celý prostor $\mathcal{X} \times C$. To ovšem velmi často není v reálných případech možné, a tak se spokojíme alespoň s odhady míry chybné klasifikace $R^*(d)$, které lze vypočítat pomocí trénovacího a testovacího výběru.

Resubstituční odhad

První z odhadů míry chybné klasifikace je takzvaný resubstituční odhad.

Definice 3 (Resubstituční odhad)**.** *Nechť je dán klasifikátor d zkonstruovaný pomocí trénovacího výběru $\mathcal{L}$. Potom má resubstituční odhad $R(d)$ míry chybné klasifikace $R^*(d)$ tvar*

$$R(d) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(d(\boldsymbol{x}_i) \neq j_i).$$

Z definice je patrné, že resubstituční odhad je vypočítán jako podíl chybně klasifikovaných případů z $\mathcal{L}$. Takový odhad ovšem dává zkreslený pohled na přesnost klasifikátoru d, jelikož výběr $\mathcal{L}$ je použitý jak na konstrukci klasifikátoru d, tak pro výpočet odhadu jeho kvality. Proto dává zkreslený pohled a je zpravidla příliš optimistický. Ovšem i takový odhad bude mít později v této práci své využití.

Odhad pomocí testovacího výběru

Druhým odhadem je odhad pomocí testovacího výběru. Takový odhad nám již bude dávat lepší informaci o přesnosti klasifikátoru, jelikož bude sestavený na výběru nezávislém na výběru trénovacím.

Definice 4 (Odhad pomocí testovacího výběru)**.** *Nechť je výběr $\mathcal{L}$ nezávisle rozdělen na dvě disjunktní podmnožiny $\mathcal{L}_1$ a $\mathcal{L}_2$. Nechť je následně sestaven klasifikátor d za pomoci podvýběru $\mathcal{L}_1$. Potom se odhad pomocí testovacího výběru vypočítá jako*

$$R^{TV}(d) = \frac{1}{N_2} \sum_{(\boldsymbol{x}_i, j_i) \in \mathcal{L}_2} \mathbb{1}(d(\boldsymbol{x}_i) \neq j_i),$$

kde N_2 je počet případů testovacího výběru $\mathcal{L}_2$.

Nevýhodou takového odhadu je ovšem fakt, že trénovací výběr $\mathcal{L}$ musí být dostatečně velký, aby bylo možné si dovolit zmenšit jeho velikost a část případů použít pouze pro odhad míry chybné klasifikace.

Odhad metodou křížové validace

Třetím, a posledním odhadem, je odhad metodou křížové validace. Tento odhad má výhodu, že je možné ho použít i u menších výběrů.

Definice 5 (Odhad metodou křížové validace)**.** *Nechť je trénovací výběr $\mathcal{L}$ rozdělen na V podobně velkých disjunktních podmnožin $\mathcal{L}_1, \ldots, \mathcal{L}_V$. Potom pro každé v, $v = 1, \ldots, V$, je sestrojen klasifikátor $d^{(v)}$ pouze s použitím trénovacího výběru $\mathcal{L} \setminus \mathcal{L}_v$. Potom odhad odhadu $R^* d^{(v)}$ má tvar*

$$R^{TV}(d^{(v)}) = \frac{1}{N_v} \sum_{(\boldsymbol{x}_i, j_i) \in \mathcal{L}_v} \mathbb{1}(d^{(v)}(\boldsymbol{x}_i) \neq j_i), \tag{1.2}$$

kde N_v je počet případů v $\mathcal{L}_v$. Nechť je následně sestaven klasifikátor d za pomoci celého trénovacího výběru $\mathcal{L}$. Pak je odhad metodou křížové validace $R^{KV}(d)$ vypočítán jako

$$R^{KV}(d) = \frac{1}{V} \sum_{v=1}^{V} R^{TV}(d^{(v)}). \tag{1.3}$$

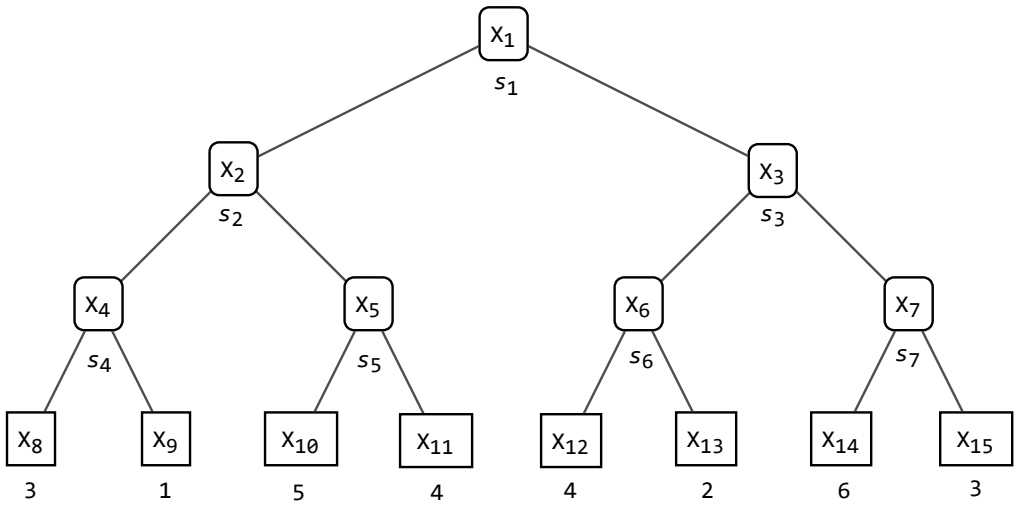
Nejprve si všimněme, že pro každé v, $v = 1, \ldots, V$, je $\mathcal{L}_v$ nezávislé na výběru $\mathcal{L} - \mathcal{L}_v$ a proto je možné v rovnosti (1.2) použít odhad pomocí testovacího výběru pro $d^{(v)}$. Vzhledem k tomu, že každý z V klasifikátorů je zkonstruován pomocí výběru s $N(1 - \frac{1}{V})$ případy, který je pro vhodně zvolené V přibližně velký jako základní trénovací výběr $\mathcal{L}$, je tato metoda poměrně stabilní. Pro $V = N$ se daná metoda nazývá N-násobná křížová validace (one-out validation). Pro každé n, $n = 1, \ldots, N$, se při konstrukci $d^{(n)}$ nepoužije vždy pouze jeden případ. Tento jeden případ se následně použije jako testovací a konečný odhad se vypočítá podle (1.3).

1.2 Struktura klasifikačního stromu

Binární stromově strukturovaný klasifikátor, dále označovaný jako klasifikační strom, je soubor dělících pravidel, a pravidla přiřazení do tříd uspořádaných do stromové struktury, díky které je možné klasifikovat na základě vektoru měření $\boldsymbol{x}$ případ do jedné z možných tříd proměnné y, také označované jako predikovaná proměnná y. Každé dělení, označované písmenem s_i, dělí vhodnou nekoncovou množinu X_i do dvou dceřiných disjunktních podmnožin X_{2i} a X_{2i+1}. Takové značení nám zaručí přehlednost stromu, jak uvidíme později. To, že jsou množiny disjunktní, nám zajistí, že případ vždy skončí pouze v jedné koncové podmnožině. O predikci proměnné y daného případu se následně rozhodne na základě třídy přiřazené příslušné koncové množině.

Na obrázku 1.1 vidíme klasifikační strom tak, jak se typicky znázorňuje. První dělení s_1 dělí prostor $\mathcal{X} = X_1$ obsahující všechny vektory $\boldsymbol{x}$ na dvě disjunktní podmnožiny X_2 a X_3 a platí, že $X_1 = X_2 \cup X_3$. X_2 se dále dělí na X_4 a X_5 a tak dále. Jak strom pokračuje dolů, vytvářejí se stále menší a menší podmnožiny prostoru $\mathcal{X}$. Podmnožiny $X_8, \ldots, X_{15}$ jsou koncové podmnožiny s označením přiřazené třídy, které je uvedeno na obrázku 1.1 pod nimi. Tedy platí

$$A_1 = X_9,\ A_2 = X_{13},\ A_3 = X_8 \cup X_{15}, \ldots$$



Obrázek 1.1: Struktura stromu

Podmnožiny X_i budou označovány jako uzly a značené písmenem t. Každý strom má koncové uzly, které odpovídají koncovým množinám, a nekoncové uzly, které odpovídají nekoncovým množinám. Prvnímu nekoncovému uzlu t_1 se také říká počáteční uzel. Nakonec je potřeba ještě definovat vztahy mezi uzly.

Definice 6 (dceřiný uzel)**.** *Nechť je dán uzel t. Potom uzel t' nazveme dceřiným uzlem uzlu t, jestliže existuje cesta z uzlu t dolů do uzlu t'.*

Jak lze vidět na obrázku 1.1, dceřinými uzly uzlu $t_2 = X_2$ jsou například uzly $t_5 = X_5$ a $t_8 = X_8$. Počáteční uzel t_1 má vždy všechny ostatní uzly dceřiné.

Definice 7 (Otcovský uzel)**.** *Je-li dán uzel t', nazýváme uzel t otcovským uzlem uzlu t', jestliže t je ve stromu výše a jsou spojeny cestou.*

Na obrázku 1.1 má uzel $t_8 = X_8$ otcovské uzly $t_4 = X_4$, $t_2 = X_2$ a $t_1 = X_1$. Je jednoduché si uvědomit, že uzel t' je dceřiný uzel uzlu t právě tehdy, když t je otcovský uzel uzlu t'.

Definice 8 (Větev)**.** *Větev T_t stromu T s počátečním uzlem $t \in T$ se skládá z uzlu t a všech jeho dceřiných uzlů.*

Na obrázku 1.1 můžeme vidět například větev $T_{t_4} = T_{X_4}$, která obsahuje uzly $t_4 = X_4$, $t_8 = X_8$ a $t_9 = X_9$.

1.3 Dělení uzlu

Ještě než se podíváme na to, jak v každém nekoncovém uzlu najít dělení s, je potřeba vysvětlit problémy spojené s tím, jak vlastně určit koncový uzel. Prvotní myšlenka by mohla vést k tomu, že by se strom v každém uzlu dělil dělením s do té doby, než by bylo uspokojeno nějaké pravidlo pro zastavení dělení, které by daný uzel prohlásilo za koncový, a následně by mu byla podle určitého pravidla přiřazena třída. Takový postup použijeme s jednou změnou, a tou je, že žádné pravidlo pro zastavení nebude potřeba. Místo toho se zkonstruuje opravdu velký strom a následně se bude za určitých pravidel zmenšovat tak, aby byl co nejefektivnější. Nakonec se každému koncovému uzlu takového stromu přiřadí vhodná třída.

Standardní sada otázek

Předpokládejme standardní strukturu dat, tj. konečnou dimenzi případů, kdy $\boldsymbol{x} = (x_1, \ldots, x_M)$ je směsí numerických a kategoriálních prediktorů. Ve standardní sadě otázek závisí každé dělení s pouze na jednom zvoleném prediktoru. Od této chvíle bude Ω označovat množinu všech možných standardních otázek. Pro každý kategoriální prediktor x_m, kde $x_m \in \{x_1, \ldots, x_M\}$ nabývá kategoriálních hodnot v $\{b_1, \ldots, b_L\}$, obsahuje Ω všechny otázky tvaru $\{\text{Je } x_m \in B?\}$, kde B probíhá přes všechny podmnožiny $\{b_1, \ldots, b_L\}$. Je-li x_m numerický prediktor, $x_m \in \{x_1, \ldots, x_M\}$, potom Ω obsahuje všechny otázky tvaru $\{\text{Je } x_m \leq c?\}$ pro $c \in (-\infty, +\infty)$. Přestože Ω je nekonečná množina otázek, existuje konečný počet dělení dat. Především je třeba si uvědomit, že pokud je x_m numerický prediktor, existuje maximálně N různých hodnot x_m. Pak existuje maximálně $(N-1)$ hodnot c_n, přičemž hodnoty c_n jsou voleny tak, aby byly přesně v polovině intervalů po sobě jdoucích hodnot x_m. Proto mají otázky numerického prediktoru tvar $\{\text{Je } x_m \leq c_n?\}$ a je jich konečně mnoho. Pokud je x_m kategorický prediktor, potom nabývá hodnot například v $\{b_1, \ldots, b_L\}$. Pak existuje přesně $(2^{L-1} - 1)$ různých smysluplných[3] dělení s pro daný prediktor. Je důležité si uvědomit, že na každou otázku je možné odpovědět pouze „Ano"či „Ne", proto se taková dělení nazývají binární. Množinu všech možných binárních dělení budeme značit jako S.

[3]Nezahrnuji dělení na celou množinu C a prázdnou množinu, jelikož takové dělení nemá smysl.

1.3.1 Funkce nečistoty

V této podkapitole bude zavedena funkce nečistoty, která nám v každém uzlu t při použití dělení s řekne, jak efektivní dané dělení je a jestli ho použít, nebo je třeba hledat nějaké jiné. Nejdříve bude ovšem potřeba představit myšlenku takzvaných proporcí uzlu, která nám pomůže tuto funkci definovat.

Proporce uzlu

Předpokládejme, že je daný trénovací výběr $\mathcal{L}$ a jednotlivé případy mohou být klasifikovány do J tříd, tedy $C = \{1, \ldots, J\}$. Nechť N_j představuje počet případů z trénovacího výběru $\mathcal{L}$ ve třídě j. Pak lze apriorní pravděpodobnosti příslušnosti do třídy j, které značíme $\pi(j)$ pro všechna $j \in C$, buď odhadnout jako $\pi(j) = N_j/N$, nebo jsou dodány analytikem externě podle uvažovaných předpokladů. Nechť t je uzel, pak $N(t)$ označuje počet případů v uzlu t a $N_j(t)$ počet případů třídy j v t. Potom $N_j(t)/N_j$ je podíl případů j-té třídy spadajících do uzlu t. Odhadovaná pravděpodobnost, že případ je v j-té třídě a zároveň v uzlu t, se proto označuje $p(j, t)$ a splňuje rovnost

$$p(j, t) = \pi(j)\frac{N_j(t)}{N_j} = \frac{N_j(t)}{N}.$$

Poslední rovnost platí pouze tehdy, pokud jsou apriorní pravděpodobnosti $\pi(j)$ odhadnuty z dat v trénovacím výběru $\mathcal{L}$. Jakmile je definována pravděpodobnost $p(j, t)$, je možné definovat odhad pravděpodobnosti spadnutí případu do uzlu t jako

$$p(t) = \sum_{j=1}^{J} p(j, t).$$

Poté je odhad pravděpodobnosti, že případ patří do třídy j za podmínky, že je v uzlu t, definován následovně

$$p(j|t) = \frac{p(j, t)}{p(t)} = \frac{N_j(t)}{N_j},$$

kde poslední rovnost platí pouze v případě, že jsou apriory $\pi(j)$ odhadnuty z dat. Samozřejmě platí $p(j|t) \geq 0$, $\forall j \in C$, a také $p(1|t) + p(2|t) + \ldots + p(J|t) = 1$.

Funkce nečistoty

Předpokládejme, že v uzlu t skončily nějaké případy. Cílem dobrého dělení s je, aby případy ve dvou následujících dceřiných uzlech, označených t_L (levý dceřiný uzel) a t_R (pravý dceřiný uzel), byly „čistší". Nechť $p(j|t)$, $j = 1, \ldots, J$, je proporce všech případů v uzlu t patřících do třídy j. Potom definujeme nečistotu uzlu t jako funkci jeho proporcí:

Definice 9 (Nečistota uzlu)**.** *Míra nečistoty $i(t)$ uzlu t je definována jako nezáporná funkce ϕ proporcí uzlu $p(1|t), \ldots, p(J|t)$ splňující*

- $\phi(\frac{1}{J}, \ldots, \frac{1}{J}) = maximum.$
- $\phi(1,0,\ldots,0) = \phi(0,1,0,\ldots,0) = \ldots = 0.$

- *ϕ je symetrická funkce $p(1|t), \ldots, p(J|t)$.*

Tato definice říká, jak funkce nečistoty měří nečistotu uzlu t. Nečistota uzlu je největší, když jsou všechny třídy $1, \ldots, J$ zastoupeny v uzlu rovnoměrně. Naopak funkce ϕ je nulová, když je uzel t absolutně čistý. To znamená, že obsahuje pouze případy z právě jedné třídy.

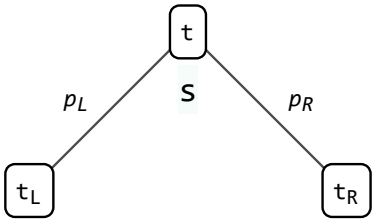
Předpokládejme, že S je množina všech binárních dělení s. Je-li dáno dělení $s \in S$, pak kvalita dělení s na uzlu t je určena poklesem nečistoty definovaným jako

$$\Delta i(s,t) = i(t) - p_L i(t_L) - p_R i(t_R),$$

kde t_R a t_L jsou dříve definované dceřiné uzly uzlu t. Navíc p_L a p_R jsou hodnoty takové, že podíl p_L případů ze všech případů v t jde do t_L a podíl p_R případů jde do t_R (obrázek 1.2). Zřejmě platí, že $p_L + p_R = 1$.

S takovou definicí nečistoty uzlu je možné rozhodovat o kvalitě dělení s, a tedy volit nejlepší dělení. Proto v každém uzlu t zvolíme jako nejlepší dělení s^* to s největším poklesem nečistoty, tedy

$$\Delta i(s^*, t) = \max_{s \in S} \Delta i(s,t).$$



Obrázek 1.2: dělení uzlu

Z definice 9 je zřejmé, jak má funkce nečistoty vypadat, ale není vůbec triviální takovou funkci najít a ověřit, že použít takovou funkci opravdu dává smysl. Proto se nyní podíváme na problém pouze dvou tříd a odvodíme takovou funkci, kterou bude následně možné použít i pro problém více tříd.

Předpokládejme nyní $C = \{1,2\}$. Nechť je dán uzel t, potom má proporce $p(1|t)$ a $p(2|t)$ a platí

$$p(1|t) = 1 - p(2|t).$$

Proto dává smysl zavést funkci nečistoty $i(t)$ tak, že se bude rovnat pravděpodobnosti chybné klasifikace v daném uzlu t. Položme

$$\begin{aligned} i(t) = \phi(p(1|t), p(2|t)) &= 1 - \max(p(1|t), p(2|t)) = \\ &= \min(p(1|t), p(2|t)) = \min(p(1|t), 1 - p(1|t)). \end{aligned}$$

Tato funkce je definovaná na intervalu [0,1] s maximem v 1/2 a splňuje vlastnosti definice 9. „Neodměňuje“ ovšem čistější uzly více, jelikož na intervalech $[0, 1/2]$ a $[1/2, 1]$ je tato funkce lineární. Proto bychom potřebovali takovou funkci nečistoty, aby na intervalu [0,1] byla spojitá, symetrická, s maximem v 1/2, s minimem v 0 a v 1 a navíc aby byla konkávní. To zaručí, že budou upřednostňovány čistější uzly. Konkávní bude samozřejmě tehdy, když její druhá derivace bude spojitá a zároveň záporná na intervalu [0,1]. Proto definujeme třídu takových funkcí.

Definice 10. *Nechť je dán uzel t. Potom je funkce $\phi(p(1|t))$ pro $p(1|t) \in [0,1]$ se spojitou derivací druhého řádu na intervalu $[0,1]$ třídy $\mathcal{F}$ právě tehdy, když splňuje:*

1. $\phi(0) = \phi(1) = 0$.
2. $\phi(p(1|t)) = \phi(1 - p(1|t))$.
3. $\phi''(p(1|t) < 0$ *na intervalu* $[0,1]$.

Nyní uvedeme tvrzení, které říká, že funkce z třídy $\mathcal{F}$ jsou vhodné, jelikož jsou vždy nenulové a jsou nulové pouze tehdy, když jsou totožné proporce uzlů t, t_L a t_R.

Tvrzení 1. *Nechť $\phi \in \mathcal{F}$ a $i(t) = \phi(p(1|t))$. Potom pro libovolný uzel $t \in T$ a libovolné dělení $s \in S$ platí*

$$\Delta i(s,t) \geq 0 \tag{1.4}$$

s rovností právě tehdy, když platí

$$p(1|t) = p(1|t_L) = p(1|t_R). \tag{1.5}$$

Důkaz. $\phi(p(1|t))$ je z třídy funkcí $\mathcal{F}$, proto je to konkávní funkce na $[0,1]$. Z konkavity vyplývá

$$\begin{aligned} i(t_L)p_L + i(t_R)p_R = \phi(p(1|t_L))p_L + \phi(p(1|t_R))p_R \leq \\ \leq \phi(p(1|t_L)p_L + p(1|t_R)p_R) \end{aligned} \tag{1.6}$$

s rovností tehdy a jen tehdy, když $p(1|t_L) = p(1|t_R)$. Nyní

$$\begin{aligned} p(1|t_L)p_L + p(1|t_R)p_R &= \frac{p(1,\, t_L)}{p(t_L)}\frac{p(t_L)}{p(t)} + \frac{p(1,\, t_R)}{p(t_R)}\frac{p(t_R)}{p(t)} = \\ &= \frac{p(1,\, t_L) + p(1,\, t_R)}{p(t)} \\ &= p(1|t). \end{aligned} \tag{1.7}$$

(1.6) a (1.7) proto implikují

$$i(t_L)p_L + i(t_R)p_R \leq \phi(p(1|t)) = i(t),$$

což můžeme přepsat jako

$$0 \leq i(t) - i(t_L)p_L - i(t_R)p_R = \Delta i(s,t) \tag{1.8}$$

a nerovnost (1.4) je dokázána. Rovnost v (1.8) podle (1.6) nastává tehdy a jen tehdy, pokud $p(1|t_L) = p(1|t_R)$. Označme $\gamma = p(1|t_L) = p(1|t_R)$ a dosaďme do rovnosti (1.7). Potom platí:

$$p(1|t) = \gamma p_L + \gamma p_R = \gamma(p_L + p_R) = \gamma.$$

Tím je dokázána rovnost (1.5).

□

Funkce ϕ z $\mathcal{F}$ jsou konkávní, což více „odměňuje“ čisté uzly. Navíc tvrzení 1

říká, že je $\Delta i(s,t)$ nulová právě tehdy, když $p(1|t) = p(1|t_L) = p(1|t_R)$. To se ovšem stává vzácně, aby měly všechny tři uzly totožné proporce pro každou třídu. Zároveň se také stává ojediněle, že více než jedno dělení s maximalizuje $\Delta i(s,t)$. Nyní stačí už jen nějakou takovou funkci ϕ najít.

Pro problém dvou tříd je možné takovou funkci najít například jako kvadratický polynom. Po jednoduchém dosazení podmínek z 10 dostaneme funkci

$$\phi(x) = b(x - x^2),\ b \geq 0. \tag{1.9}$$

Bez újmy na obecnosti tedy zvolme $b = 1$. Potom má naše výsledná funkce ϕ tvar

$$\phi(x) = x(1 - x),$$

a tedy

$$\phi(p(1|t)) = p(1|t)(1 - p(1|t)) = p(1|t)p(2|t).$$

Následně definujeme míru nečistoty libovolného uzlu t jako

$$i(t) = p(1|t)p(2|t). \tag{1.10}$$

Funkce 1.10je v $\mathcal{F}$ a tak splňuje všechny vlastnosti definice 10. Výhoda této funkce je ta, že lze jednoduše převést do klasifikačního problému s více třídami.

Giniho kritérium

Giniho index diverzity je zobecnění funkce nečistoty definované v (1.10) pro více tříd. Nechť je dán uzel t a jeho proporce $p(1|t), \ldots, p(J|t)$. Potom definujeme Giniho index diverzity jako funkci nečistoty $i(t)$ vztahem

$$i(t) = \sum_{j \neq i} p(j|t)p(i|t). \tag{1.11}$$

Rovnost (1.11) můžeme ještě upravit:

$$\begin{aligned} i_{GI}(t) &= \sum_{j \neq i} p(j|t)p(i|t) = \\ &= \sum_{j \neq i} p(j|t)p(i|t) + \sum_{j=i} p(j|t)p(i|t) - \sum_{j=i} p(j|t)p(i|t) = \\ &= \sum_{j,i} p(j|t)p(i|t) - \sum_{j} p(j|t)^2 = \\ &= \left(\sum_{j} p(j|t)\right)^2 - \sum_{j} p(j|t)^2 = \\ &= 1 - \sum_{j} p(j|t)^2. \end{aligned} \tag{1.12}$$

Z (1.11) vidíme, že funkce nečistoty definovaná pro problém dvou tříd je vlastně Giniho index diverzity, pokud v (1.9) zvolíme $b = 2$. Z konkavity a vlastností této funkce tedy opět platí, že

$$\Delta i(s,t) \geq 0$$

s rovností právě tehdy, když

$$p(j|t) = p(j|t_L) = p(j|t_R) \text{ pro } j = 1, \ldots, J.$$

Twoing a Entropy kritéria

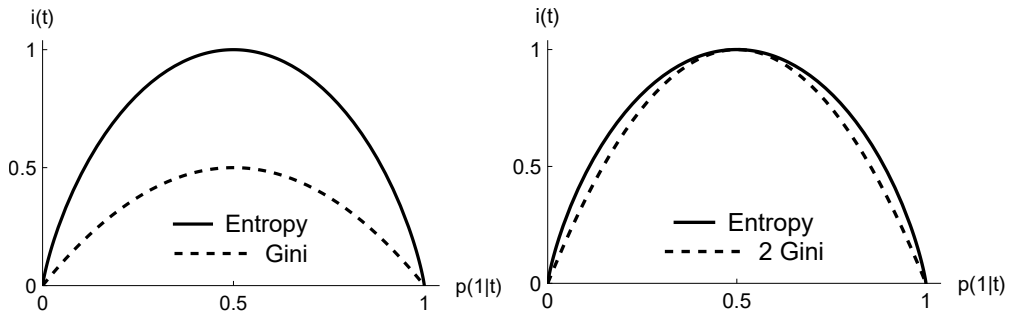
Další možnou volbou funkce nečistoty může být takzvané Twoing kritérium. To je definované přímo jako funkce poklesu nečistoty. Za podmínky, že je dán uzel t a dělení s, má Twoing kritérium tvar

$$\Delta i_{TW}(s,t) = \frac{p_L p_R}{4}\left[\sum_{j=1}^{J} |p(j|t_L) - p(j|t_R)|\right]^2.$$

Taková definice funkce poklesu nečistoty dává smysl proto, že je to normovaný součet absolutních rozdílů pravděpodobností $p(j|t_L)$ a $p(j|t_R)$ pro každou třídu j. Nejlepší dělení bude tuto funkci maximalizovat, přičemž toho dosáhne tehdy, když absolutní hodnoty budou co největší. To nastane právě tehdy, když dělení $s*$ pošle co nejvíce případů z uzlu t náležejících každé třídě společně právě do jednoho ze dvou dceřiných uzlů. Na rozdíl od Giniho kritéria, které mělo spíše tendenci v každém uzlu separovat jednu třídu a tu poslat do jednoho dceřiného uzlu, se Twoing kritérium snaží hledat taková dělení, která dělí případy po třídách, které mají společné vlastnosti. Použití jednoho či druhého kritéria závisí na tom, zda můžeme předpokládat takové vlastnosti tříd, či nikoliv.

Dalším možným kritériem je použití klasické Shannonovy entropie. Toto kritérium má tvar

$$i_{EN}(t) = -\sum_j p(j|t)\log_2 p(j|t).$$



(a) (b)

Obrázek 1.3: Giniho a Entropy kritéria

Na obrázku 1.3(a) lze vidět rozdíl mezi Giniho a Entropy kritériem pro dvě třídy. Entropy kritérium také náleží do třídy $\mathcal{F}$ podle definice 10. Na obrázku 1.3(b) je jejich srovnání, po vynásobení Giniho kritéria dvěma, což nemá žádný vliv na funkci míry nečistoty. Entropy kritérium tedy více penalizuje nečisté uzly než Giniho kritérium. Co se týče výpočetní složitosti, je Entropy kritérium kvůli logaritmu výpočetně náročnější, ale jak lze vidět v článku A. Hershyho (Hershy, 2019), může být často efektivnější.

1.3.2 Pravidlo přiřazení třídy

Předpokládejme nyní, že již máme zkonstruovaný klasifikační strom T. Jak se takový strom dostane bude popsáno v následují kapitole. Nyní ovšem předpokládejme množinu koncových uzlů stromu T a označme ji $\tilde{T}$. Pravidlo přiřazení třídy

nám říká, že každému koncovému uzlu t z množiny $\tilde{T}$ je přiřazena třída $j \in C$, označená výrazem $j(t)$. Je přirozené vybrat tu třídu, pro kterou je v daném koncovém uzlu nejvíce případů. Tedy

$$p(j(t)|t) = \max_{j \in C} p(j|t). \tag{1.13}$$

Taková třída zároveň minimalizuje resubstituční odhad míry chybné klasifikace případu v uzlu t. Proto má takový odhad tvar

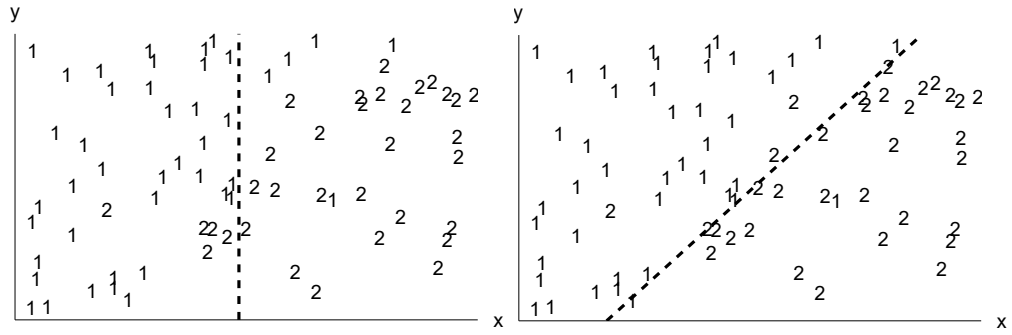
$$r(t) = \sum_{j \neq j(t)} p(j|t) = 1 - \max_{j \in C} p(j|t). \tag{1.14}$$

Pokud existuje více tříd splňujících rovnici (1.13), pak je třída $j(t)$ vybrána libovolně, tedy jako jedna z maximalizujících tříd.

1.3.3 kombinace prediktorů

Dosud jsme předpokládali, že každé dělení $s \in S$ dělí uzel pouze na základě jednoho prediktoru. Nyní se v krátkosti podíváme na to, jak lze využít více prediktorů k efektivnějšímu dělení.

Nyní uvažujme pouze numerické prediktory a předpokládejme, že vektor $\boldsymbol{x}$ má dimenzi 2, tedy $\boldsymbol{x} = (x,y)$. Problém dělení podle jednoho prediktoru, dále označovaného jako jednoduché dělení, může být ten, že každé dělení tvaru s :{Je $x \leq c$?} nebo s :{Je $y \leq c$?} dělí rovinu na dvě poloviny přímkou rovnoběžnou s jednou ze dvou os, kdy druhou osu protíná v bodě c. Data ovšem můžou mít závislost, kterou rovnoběžka s osou neodhalí.



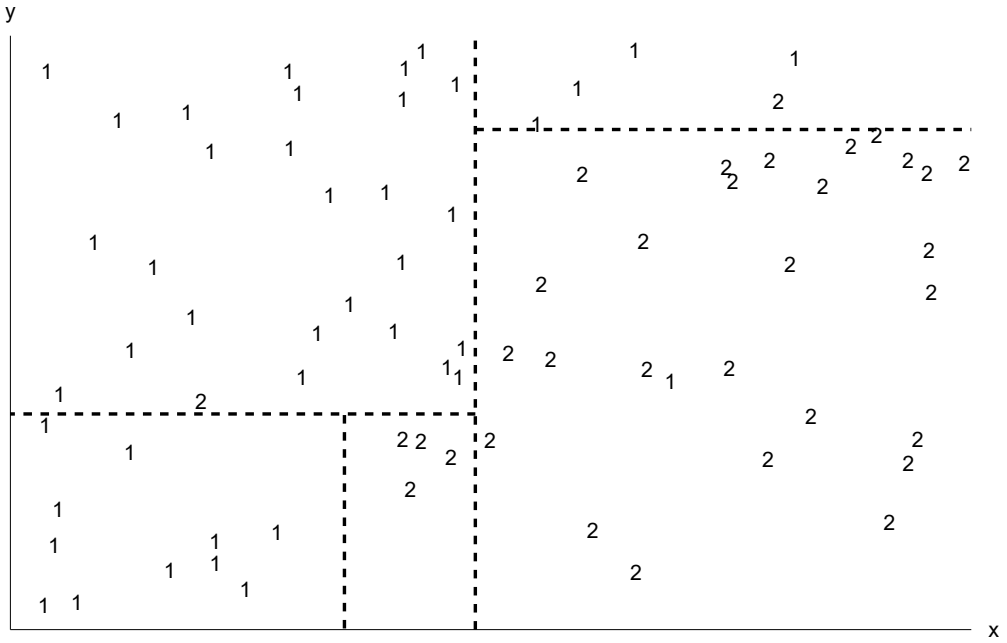
(a) jednoduché dělení (b) Lineární dělení dvou prediktorů

Obrázek 1.4: Lineární dělení

Na obrázku 1.4(a) vidíme, jak jednoduché dělení neoddělilo případy třídy 1 od případů třídy 2 tak dobře, jak to udělalo lineární dělení dvou prediktorů na obrázku 1.4(b). Proto by byly potřeba alespoň 4 jednoduchá dělení, aby se vyrovnaly lineárnímu dělení, jak je vidět na obrázku 1.5.

Toto byl pouze ilustrační příklad, v praxi lze samozřejmě použít více než 2 prediktory. Bez újmy na obecnosti předpokládejme, že z M prediktorů je jich M_1 numerických. Nechť je dán uzel t, potom uvažujme množinu koeficientů

$$\boldsymbol{a} = (a_1, \ldots, a_{M_1})$$



Obrázek 1.5: Více jednoduchých dělení

takovou, aby platilo

$$||\boldsymbol{a}||^2 = \sum_{m=1}^{M_1} a_m^2 = 1.$$

Lineární dělení více prediktorů má potom tvar

$$\sum_{m=1}^{M_1} a_m x_m \leq c,\ c \in \mathbb{R}.$$

Dále by bylo možné pomocí určitého algoritmu odstraňovat postupně prediktory s menší důležitostí tak, aby v daném dělení zůstaly pouze ty prediktory, které jsou dostatečně efektivní. Existuje i metoda, která zahrnuje také kategorické proměnné. Ačkoliv tyto metody zmenšují počet potřebných dělení v klasifikačním stromě, jsou výpočetně neefektivní a navíc špatně interpretovatelné. Jak bude možné vidět ve třetí kapitole, pokročilejší a sofistikovanější algoritmy pro hledání klasifikačních stromů využívají pouze jednoduchá dělení.

1.3.4 Náklady chybné klasifikace

Dosud se předpokládalo, že náklady chybné klasifikace jsou pro všechny třídy stejné. To znamenalo, že chybné klasifikování případu z každé třídy má stejnou váhu. To není vhodné například v úvodním příkladu, kdy bychom chtěli dát větší váhu chybné klasifikace těm případům, kdy žena má komplikovaný porod, ale je klasifikována do třídy nekomplikovaných porodů. Pro takový požadavek je možné zavést pojem náklad chybné klasifikace následovně:

Definice 11 (Náklad chybné klasifikace)**.** *$C(i|j)$ je náklad chybné klasifikace objektu třídy j jako objektu třídy i, pokud pro něj platí*

1. *$C(i|j) \geq 0$ pro každé $i \neq j$*

2. $C(i|j) = 0$ *pro každé* $i = j$.

Za předpokladu nejednotkových nákladů chybné klasifikace se používá k dělení uzlů například modifikovaný Giniho index definovaný jako

$$i_{GI}(t) = \sum_{i,j} C(i|j)p(i|t)p(j|t). \tag{1.15}$$

Zde je možné sčítat i přes $i = j$, jelikož $C(i|j) = 0$ pro $i = j$. Nyní ovšem vyvstává jeden problém, a to ten, že index v (1.15) nezohledňuje rozdíly nákladů chybné klasifikace mezi dvěma třídami. Totiž, pokud pro fixní i a j platí $C(i|j) \neq C(j|i)$, na takový rozdíl nebude brán zřetel a náklad chybné klasifikace těchto dvou tříd se sečte a zprůměruje mezi obě třídy. To je možné vyřešit docela složitým dodefinováním sekundárních forem apriorních pravděpodobností, jak lze vidět v knize Breiman (1993, str. 114). Tím se zde nebudeme zabývat, ale zjednodušeně řečeno, je snaha převést nejednotkové náklady chybné klasifikace $C(i|j)$ co nejblíže jednotkovým při vhodné změně apriorních pravděpodobností $\pi(j)$.

Nyní se podíváme na volbu třídy $j(t)$ pro koncový uzel za podmínky, že jsou použité nenulové náklady. Definice 11 říká, že když je náhodný případ, který padl do uzlu t, klasifikován do třídy i, pak je odhadovaná míra chybné klasifikace rovna

$$\sum_{j \in C} C(i|j)p(j|t). \tag{1.16}$$

Rozumný postup je přiřadit koncovému uzlu t takovou třídu $j(t)$, která minimalizuje (1.16). Potom je při daném uzlu t resubstituční odhad $r(t)$ míry chybné klasifikace definován jako

$$r(t) = \min_{i \in C} \sum_{j \in C} C(i|j)p(j|t). \tag{1.17}$$

Všimněme si, že (1.17) s jednotkovými náklady chybné klasifikace, tedy

$$C(i|j) = 1,\ i \neq j,$$

odpovídá resubstitučnímu odhadu (1.14) a platí

$$r(t) = \min_{i \in C} \sum_{j \in C} C(i|j)p(j|t) = 1 - \max_{j \in C} p(j|t). \tag{1.18}$$

1.4 Tvorba klasifikačního stromu

Již bylo popsáno, jak je klasifikační strom strukturovaný, jak se hledá v každém uzlu nejlepší dělení a jak se koncovým uzlům přiřadí vhodná třída. Nyní je potřeba vysvětlit, jak se takový strom zkonstruuje. Myšlenka bude taková, že se začne s jedním počátečním uzlem. Ten se bude dělit pomocí nejlepších dělení do dceřiných uzlů tak dlouho, dokud v každém uzlu nebude jen pár případů. Poté se pomocí metody ořezávání s minimální nákladovou složitostí bude strom zmenšovat tak, že vznikne posloupnost ideálních stromů, ze kterého bude nakonec vybrán jeden nejlepší strom. Pro takový algoritmus je nejdříve potřeba dodefinovat resubstituční odhad míry chybné klasifikace stromu, který bude vycházet z definice resubstitučního odhadu míry chybné klasifikace uzlu definovaném v (1.17).

Definice 12 (Resubstituční odhad míry chybné klasifikace stromu)**.** *Nechť je dán strom T a resubstituční odhad míry chybné klasifikace uzlu $r(t)$ tvaru*

$$r(t) = \min_{i \in C} \sum_{j \in C} C(i|j)p(j|t).$$

Položme $R(t) = p(t)r(t)$, kde $p(t)$ je pravděpodobnost spadnutí případu do uzlu t. Potom je resubstituční odhad míry chybné klasifikace stromu definován jako

$$R(T) = \sum_{t \in \tilde{T}} R(t),$$

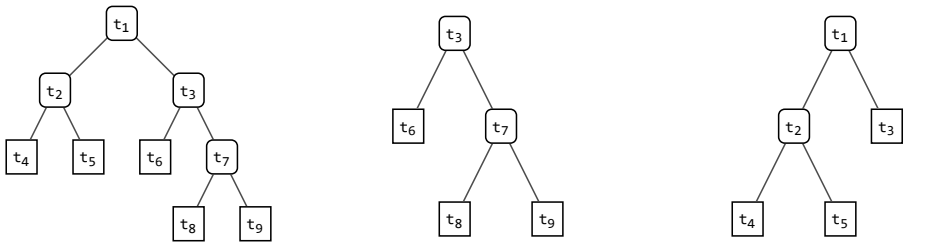
kde $\tilde{T}$ je množina koncových uzlů stromu T.

Tato definice poskytuje nástroj, jak odhadnout $R^*(T)$. $R(T)$ a $R(t)$ budou nyní označovány jako míra chybné klasifikace stromu T a míra chybné klasifikace uzlu t.

1.4.1 Ořezávání s minimální nákladovou složitostí

Klíčové pro tuto metodu je nejdříve definovat *ořezání* větve.

Definice 13 (Ořezání větve)**.** *Ořezání větve T_t ze stromu T spočívá v odstranění všech dceřiných uzlů uzlu t ze stromu T. Takto ořezaný strom je označený $T - T_t$.*



(a) Strom T (b) Větev T_{t_3} (c) Strom $T - T_{t_3}$

Obrázek 1.6: Ořezávání

Pokud je strom T' ořezán ze stromu T, pak se T' označuje jako ořezaný podstrom T a značí se $T' \prec T$. Na obrázku 1.6 lze vidět příklad stromu T a jeho konečná podoba po ořezání větve T_t.

Prvním krokem v metodě ořezávání je nechat vyrůst velký strom T_{MAX}, který má v každém koncovém uzlu pouze jeden případ. Smyslem této metody je nyní postupně ořezat postupně větve tak, aby vznikla posloupnost $T_{MAX}, T_1, T_2, \ldots, \{t_1\}$, kde $\{t_1\}$ je počáteční uzel stromu T_{MAX}.

Definice 14. *Pro libovolný podstrom $T \prec T_{MAX}$ definujeme jeho složitost jako $|\tilde{T}|$ – počet koncových uzlů v T. Nechť $\alpha \geq 0$, $\alpha \in \mathbb{R}$, je parametr složitosti. Pak je míra nákladové složitosti $R_\alpha(T)$ stromu T definována jako*

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|. \tag{1.19}$$

V rovnosti (1.19) vyjadřuje $R(T)$ míru chybné klasifikace stromu a $\alpha|\tilde{T}|$ penalizaci za složitost. To znamená, že větší strom s více koncovými uzly bude penalizován více než malý strom s nškolika málo koncovými uzly. To je důležité, jelikož příliš velké stromy v sobě nesou velké riziko malé generalizace.

Nechť je zkonstruovaný strom T_{MAX} tak, aby v každém koncovém uzlu měl pouze jeden případ. Nyní pro každou hodnotu $\alpha \geq 0$ najdeme následně podstrom $T(\alpha) \prec T_{MAX}$ tak, aby minimalizoval $R_\alpha(T)$:

$$R_\alpha(T(\alpha)) = \min_{T \preccurlyeq T_{MAX}} R_\alpha(T).$$

Pro malé hodnoty α bude penalizace za velký počet koncových uzlů zanedbatelná a vybraný strom bude velký. Na druhou stranu, pro velké α bude penalizace dominantní a vybraný strom bude mít jen pár koncových uzlů. Pro nějakou hodnotu α_0 bude pro všechny $\alpha \geq \alpha_0$ vybraným stromem dokonce pouze počáteční uzel $\{t_1\}$. Takto je tedy definována posloupnost stromů T_{MAX}, T_1, $T_2, \ldots, \{t_1\}$. Penalizací za složitost ovšem může být i jiná funkce, například $\alpha|\tilde{T}|^\nu$ pro $\nu \in (0,1)$, jak lze vidět v článku C. Scotta (Scott, 2005).

Myšlenka spočívá v tom, že na určitém intervalu $\alpha \in [0,a)$ volí ořezávání s minimální nákladovou složitostí strom T_1. Poté dojde ke skoku v bodě a a v následujícím intervalu $\alpha' = [a,b)$ algoritmus vybere strom T_2. Tento proces bude pokračovat tak dlouho, dokud nebude dosaženo koncového uzlu $\{t_1\}$. Proto bude posloupnost konečná. Zde jsou ovšem dva problémy. Prvním je existence podstromu $T \preccurlyeq T_{MAX}$ pro každou hodnotu α a jeho jednoznačnost. Druhým problémem je otázka: Je v posloupnosti stromů T_{MAX}, T_1, $T_2, \ldots, \{t_1\}$ každý strom podstromem předchozího stromu, tj. platí $T_{MAX} \succ T_1 \succ T_2 \succ \ldots \succ \{t_1\}$?

Definice 15. *Nejmenší minimalizující podstrom $T(\alpha)$ pro parametr složitosti α splňuje:*

1. $R_\alpha(T(\alpha)) = \min_{T \preccurlyeq T_{MAX}} R_\alpha(T)$.
2. *Jestliže $R_\alpha(T) = R_\alpha(T(\alpha))$, pak $T(\alpha) \preccurlyeq T$.*

Tato definice říká, že pokud má více různých stromů stejnou hodnotu $R_\alpha(T)$, je vždy vybrán menší strom. Posledním problémem je tedy existence. Nicméně:

Věta 2. *Pro každou hodnotu $\alpha \geq 0$ existuje nejmenší minimalizující podstrom podle definice 15.*

Důkaz. Viz (Breiman, 1993, Theorem 10.9). □

Existence $T(\alpha)$ je tedy zajištěna. Zbývá ukázat, jak takovou posloupnost skutečně najít a zaručit, aby splňovala $T_{MAX} \succ T_1 \succ T_2 \succ \ldots \succ \{t_1\}$. Výhodnější je nyní začít s T_1 místo T_{MAX}. Zvolíme tedy T_1 jako $T_1 = T(0)$. T_1 je proto nejmenší podstrom T_{MAX} splňující podmínku

$$R(T_1) = R(T_{MAX}).$$

K tomu ovšem bude potřeba ukázat, že pro každý nekoncový uzel t platí

$$R(t) \geq R(t_L) + R(t_R).$$

Věta 3. *Pro libovolné dělení $s \in S$ uzlu $t \in T$ platí*

$$R(t) \geq R(t_L) + R(t_R). \tag{1.20}$$

Rovnost nastává právě tehdy, když přiřazené třídy uzlů t_L, t_R a t jsou totožné. Tedy platí

$$j(t) = j(t_L) = j(t_R). \tag{1.21}$$

Důkaz. Bez újmy na obecnosti předpokládejme, že uzlu t je přiřazena třída $j(t)$. Potom platí:

$$\begin{aligned} R(t) = r(t)p(t) &= \sum_{j \in C} C(j(t)|j)p(j|t)p(t) = \\ &= \sum_{j \in C} C(j(t)|j)p(j,\, t) = \\ &= \sum_{j \in C} C(j(t)|j)[p(j,\, t_L) + p(j,\, t_R)] = \\ &= \sum_{j \in C} C(j(t)|j)p(j,\, t_L) + \sum_{j \in C} C(j(t)|j)p(j,\, t_R) = \\ &= R(t_L) + R(t_R). \end{aligned}$$

Poslední rovnost platí právě tehdy, když uzly t_R a t_R mají přiřazenou třídu $j(t)$. Tím jsme dokázali rovnost ve vztahu (1.21). Třída $j(t)$ byla ale každému uzlu přiřazena tak, aby byl minimalizována míra chybné klasifikace

$$\sum_{j \in C} C(i|j)p(j|t). \tag{1.22}$$

Proto tedy platí

$$\begin{aligned} R(t) - R(t_L) - R(t_R) &= \\ &= r(t)p(t) - r(t)p(t_L) - r(t)p(t_R) = \\ &= \sum_{j \in C} C(j(t)|j)p(j,\, t_L) + \sum_{j \in C} C(j(t)|j)p(j,\, t_R) - \\ &\quad - \min_{i \in C} \sum_{j \in C} C(i|j)p(j,\, t) - \min_{i \in C} \sum_{j \in C} C(i|j)p(j,\, t). \end{aligned} \tag{1.23}$$

Kdykoliv je dceřiným uzlům t_L, t_R přiřazena stejná třída jako uzlu t, je zaručeno, že se celková míra chybné klasifikace nezvětší. Může ovšem existovat jiná třída, která bude minimalizovat (1.22) a celkový náklad se zmenší. Pravá strana rovnosti v (1.23) je tedy vždy nezáporná a nerovnost (1.20) je dokázána.

□

Tímto postupem vznikne T_1, jehož každý koncový uzel je čistý a žádné další čisté koncové uzly již ořezáváním nelze vytvořit. Máme tedy strom T_1 a další krok spočívá v tom najít další strom T_2. Nejdříve ale potřebujeme definovat míru chybné klasifikace větve stromu T_1. Tedy pro libovolnou větev $T_{1,t}$ z T_1 definujeme její míru chybné klasifikace $R(T_{1,t})$ jako

$$R(T_{1,t}) = \sum_{t' \in \tilde{T}_{1,t}} R(t').$$

Tedy $R(T_{1,t})$ je součet měr chybné klasifikace všech koncových uzlů větve $T_{1,t}$ a platí

Věta 4. *Pro libovolný nekoncový uzel t stromu T_1 platí*

$$R(t) > R(T_{1,t}).$$

Důkaz. Důkaz této věty je jednoduchý, stačí si pouze uvědomit, že koncové uzly větve jsou dceřiné uzly uzlu t, a aplikovat Větu 3. Ostrá nerovnost bude zaručena, protože strom T_1 má všechny uzly čisté a žádné jiné již ořezáním nelze vytvořit. □

Nechť pro libovolný uzel $t \in T_1$ označuje $\{t\}$ podvětev větve $T_{1,t}$ obsahující pouze samotný uzel t. Potom

$$R_\alpha(\{t\}) = R(t) + \alpha|\tilde{\{t\}}| = R(t) + \alpha$$

vyjadřuje míru nákladové složitosti uzlu t jako podvětve větve $T_{1,t}$ a

$$R_\alpha(\tilde{T}_{1,t}) = R(T_{1,t}) + \alpha|\tilde{T}_{1,t}|$$

je míra nákladové složitosti větve $T_{1,t}$. Jelikož podle věty 4 platí, že $R(t) > R(T_{1,t})$, potom pro nějaké dostatečně velké $\alpha \geq 0$ musí platit

$$R_\alpha(\{t\}) > R_\alpha(\tilde{T}_{1,t}). \tag{1.24}$$

Tedy lze najít hodnotu α takovou, že se obě strany v nerovnosti (1.24) rovnají. Obě míry nákladové složitosti budou stejně velké, ale podvětev $\{t\}$ má méně koncových uzlů než větev $T_{1,t}$, a proto bude preferovaná. K tomu, abychom našli takovou hodnotu α, stačí uvažovat vztah (1.24) vzhledem k α, odkud dostáváme

$$\alpha > \frac{R(t) - R(T_{1,t})}{|\tilde{T}_{1,t}| - 1}. \tag{1.25}$$

Pravá strana nerovnosti (1.25) bude ovšem vždy kladná podle věty 4. Proto definujeme funkci $g_1(t)$, $t \in T_1$ následovně

$$g_1(t) = \begin{cases} \frac{R(t)-R(T_{1,t})}{|\tilde{T}_{1,t}|-1} & \text{pro } t \notin \tilde{T}_1, \\ +\infty & \text{pro } t \in \tilde{T}_1. \end{cases}$$

Následně definujeme nejslabší uzel $\bar{t_1} \in T_1$ jako uzel který splňuje rovnost

$$g_1(\bar{t_1}) = \min_{t \in T_1} g_1(t)$$

a položme $\alpha_2 = g_1(\bar{t_1})$. Uzel $\bar{t}_1$ je nejslabší v tom smyslu, že jak roste α od nuly, tento uzel jako první ze všech splňuje

$$R_\alpha(\{t\}) = R(t) - \alpha(T_{1,t}).$$

Potom je tedy uzel $\bar{t}_1$ upřednostněn před jeho větví $T_{1,\bar{t}_1}$. Proto je tato větev ořezána a definujeme nový strom $T_2 \prec T_1$ splňující

$$T_2 = T_1 - T_{1,\bar{t}_1}.$$

Nechť je tedy dán strom T_2, potom se strom T_3 získá podobným způsobem jako T_2. Definujeme funkci $g_3(t)$:

$$g_2(t) = \begin{cases} \frac{R(t)-R(T_{2,t})}{|\tilde{T}_{2,t}|-1} & \text{pro } t \notin \tilde{T}_2, \\ +\infty & \text{pro } t \in \tilde{T}_2. \end{cases}$$

Následně dostáváme nejslabší uzel $\bar{t_2} \in T_2$, který splňuje

$$g_2(\bar{t_2}) = \min_{t \in T_2} g_2(t) = \alpha_3$$

a dostáváme strom T_3 splňující

$$T_3 = T_2 - T_{2,\bar{t_2}}.$$

Takto pokračujeme do doby, než je vybrán jako ideální strom pouze počáteční uzel $\{t_1\}$. Dostáváme tedy sestupnou posloupnost do sebe vnořených stromů

$$T_1 \succ T_2 \succ \ldots \succ \{t_1\}.$$

Pokud by náhodou v nějaké fázi bylo více nejslabších uzlů, jsou ořezány všechny větve náležející těmto uzlům. Proto je tedy $\{\alpha_k\}$ rostoucí posloupnost pro $k \geq 1$, kdy $\alpha_1 = 0$, a pro $\alpha_k \leq \alpha < \alpha_{k+1}$ platí

$$T(\alpha) = T(\alpha_k) = T_k.$$

Celý proces tedy funguje tak, že je na začátku dán strom T_1. V tomto stromě je nalezen nejslabší uzel $\bar{t}_1$, jeho větev je ořezána a vzniká strom T_2 právě když α dosáhne hodnoty α_2. Poté je nalezen nejslabší uzel stromu T_2, vhodná větev je ořezána a vzniká strom T_3 když α dosáhne α_3. Tento proces běží do té doby, než je celý strom ořezaný a zbývá pouze počáteční uzel $\{t_1\}$. Takto vznikla posloupnost ideálním podstromů, ze které nyní bude vybrán jeden nejlepší strom pomocí takzvaných spravedlivých odhadů.

1.4.2 Výběr nejlepšího stromu

Nyní je potřeba z kandidátů na nejlepší strom, tedy z posloupnosti

$$T_1 \succ T_2 \succ \ldots \succ \{t_1\}$$

vybrat jeden podle nějakého kritéria. K tomu využijeme odhady míry chybné klasifikace $\hat{R}(T_k)$, které nám řeknou, který strom je nejpřesnější. Potom zvolíme z této posloupnosti takový nejlepší strom T_{k_0}, pro který bude platit

$$\hat{R}(T_{k_0}) = \min_k \hat{R}(T_k).$$

K takovému odhadu bude potřeba trénovací výběr $\mathcal{L}$ a následující pravděpodobnostní model.

Nechť je $\mathcal{L}$ náhodný výběr N nezávislých případů vybraných z pravděpodobnostního rozdělení $P[A,\ j]$ definovaném na prostoru $\mathcal{X} \times C$. Potom ($\boldsymbol{X}$,Y) je náhodný nový výběr z $P[A,\ j]$ nezávislý na $\mathcal{L}$ a míra chybné lasifikace klasifikátoru d je $R^*(d) = P[d(\boldsymbol{X}) \neq Y]$. Dále $C(i|j)$ je náklad chybné klasifikace

případu z třídy i jako případ z třídy j. Následně můžeme definovat $Q^*(i|j)$ jako pravděpodobnost, že případ z j-té třídy je klasifikovaný do třídy i a platí

$$Q^*(i|j) = P(d(\boldsymbol{X}) = i|Y = j).$$

Potom má odhad míry chybné klasifikace všech případů z j-té třídy tvar

$$R^*(j) = \sum_{i=1}^{J} C(i|j)Q^*(i|j)$$

a odhad míry chybné klasifikace klasifikátoru d:

$$R^*(d) = \sum_{j=1}^{J} R^*(j)\pi(j),$$

kde $\pi(j) = P[Y = j]$. Využijeme tento model a nejdeme odpovídající odhady na základě reálných dat.

Odhad míry chybné klasifikace pomocí testovacího výběru

Nechť je dán trénovací výběr $\mathcal{L}$ obsahující N případů. Z $\mathcal{L}$ je náhodně vybráno N_2 případů. $\mathcal{L}$ je tedy rozdělen na nový trénovací výběr $\mathcal{L}_1$ s $N_1 = N - N_2$ případy a nový testovací výběr $\mathcal{L}_2$ s N_2 případy. Na základě $\mathcal{L}_1$ je pomocí dříve popsaného algoritmu sestavena posloupnost stromů

$$T_1 \succ T_2 \succ \ldots \succ \{t_1\} \tag{1.26}$$

Tato posloupnost je zřejmě nezávislá na výběru $\mathcal{L}_2$ a proto bude tento výběr použitý k odhadu míry chybné klasifikace a zvolení nejlepšího stromu z (1.26). Definujme $N_j^{(2)}$ jako počet všech případů z $\mathcal{L}_2$, jejichž třída je j. Teď pro každý strom z (1.26) spočítáme $N_{i,j}^{(2)}$ jako počet případů z $\mathcal{L}_2$ třídy j, které byly klasifikovány do třídy i. Potom máme pro každý strom odhad $Q^*(i|j)$, který splňuje rovnost

$$Q^{TV}(i|j) = \frac{N_{i,j}^{(2)}}{N_j^{(2)}}.$$

Může se ovšem stát, že v $\mathcal{L}_2$ není zastoupena nějaká třída $j \in C$ žádným případem. Pokud tedy $N_j^{(2)} = 0$, dodefinujeme $Q^{TV}(i|j) = 0$. Následně můžeme pro každý strom spočítat odhad míry chybné klasifikace všech případů z j-té třídy jako

$$R^{TV}(j) = \sum_{i=1}^{J} C(i|j)Q^{TV}(i|j)$$

a odhad míry chybné klasifikace celého stromu T_k jako

$$R^{TV}(T_k) = \sum_{j=1}^{J} R^{TV}(j)\pi(j). \tag{1.27}$$

Pokud $\pi(j) = N_j^{(2)}/N_2$, potom (1.27) můžeme upravit následovně:

$$\begin{aligned} R^{TV}(T_k) &= \sum_{j=1}^{J} R^{TV}(j)\pi(j) = \\ &= \sum_{j=1}^{J} \left[\sum_{i=1}^{J} C(i|j) Q^{TV}(i|j) \right] \frac{N_j^{(2)}}{N_2} = \\ &= \sum_{i,j} C(i|j) \frac{N_{i,j}^{(2)}}{N_j^{(2)}} \frac{N_j^{(2)}}{N_2} = \\ &= \frac{1}{N_2} \sum_{i,j} C(i|j) N_{i,j}^{(2)}. \end{aligned} \tag{1.28}$$

Výraz (1.28) říká, že sečteme náklady chybné klasifikace každého případu z $\mathcal{L}_2$ a poté tento součet zprůměrujeme. Následně vybereme nejlepší strom z posloupnosti (1.26) takový, který minimalizuje (1.27). Tedy nejlepší strom T_{k_0} volíme tak, aby splňoval

$$R^{TV}(T_{k_0}) = \min_k R^{TV}(T_k).$$

Následné vylepšení této metody spočívá v tom, že nebudeme $\mathcal{L}_2$ volit úplně náhodně, ale zvolíme proporce tříd f_j podle N_j pro všechna $j = 1,\ldots,J$, podle kterých se následně vybere $\mathcal{L}_2$ tak, aby měl tento výběr odpovídající proporce tříd. Tento postup zaručí, že budou třídy rozdělené podle zastoupení v $\mathcal{L}$ a odhad tak bude přesnější. Odhad míry chybné klasifikace pomocí testovacího výběru se používá u větších trénovacích výběrů, kde není problém takový výběr uměle zmenšit a část použít na testování.

Odhad míry chybné klasifikace metodou křížové validace

Budiž $\mathcal{L}$ trénovací výběr. Náhodně rozdělíme $\mathcal{L}$ na V podobně velkých podmnožin $\mathcal{L}_1,\ldots,\mathcal{L}_V$. Nejčastěji se volí $V = 10$. Nyní se pro každé $v = 1,2,\ldots,V$ sestrojí posloupnost ideálních stromů za použití trénovacího výběru $\mathcal{L} - \mathcal{L}_v$. Zároveň se sestrojí posloupnost ideálních stromů za pomoci celého trénovacího výběru $\mathcal{L}$. Tedy $T(\alpha)$ a $T^{(v)}(\alpha)$ jsou ideální stromy s minimální nákladovou složitostí pro libovolnou hodnotu parametru α. Je potřeba si uvědomit, že $T(\alpha)$ jsou stromy sestaveny z celého trénovacího výběru a pro libovolné $v = 1,\ldots,V$ jsou stromy $T^{(v)}(\alpha)$ sestaveny z trénovacího výběru $\mathcal{L} - \mathcal{L}_v$, a proto jsou nezávislé na $\mathcal{L}_v$. Nyní pro zafixovanou hodnotu α a pro všechna v, i, j definujme počet případů třídy j z $\mathcal{L}_v$ klasifikovaných do třídy i stromem $T^v(\alpha)$ jako $N_{i,j}^{(v)}$. Položme

$$N_{i,j} = \sum_{v=1}^{V} N_{i,j}^{(v)},$$

kdy každý případ z $\mathcal{L}$ byl k testování použitý právě jednou, jelikož se vyskytl právě v jednom testovacím výběru. Dále definujme

$$Q^{KV}(i|j) = \frac{N_{i,j}}{N_j}$$

a vypočítejme $R^{KV}(j)$ jako

$$R^{KV}(j) = \sum_{i=1}^{J} C(i|j) Q^{KV}(i|j).$$

Potom odhad míry chybné klasifikace metodou křížové validace je

$$R^{KV}(T(\alpha)) = \sum_{j=1}^{J} R^{KV}(j)\pi(j) = \frac{1}{N}\sum_{i,j} C(i|j)N_{i,j}, \tag{1.29}$$

kdy poslední rovnost platí jen tehdy, když $\pi(j) = N_j/N$. Připomeňme, že ideální posloupnost stromů zkonstruovaných na trénovacím výběru $\mathcal{L}$ splňuje $T(\alpha) = T_k$ pro $\alpha_k \leq \alpha \leq \alpha_{k+1}$. Položme α'_k jako geometrický průměr

$$\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}.$$

Jelikož α' je v intervalu $[\alpha_k,\alpha_{k+1})$, platí

$$R^{KV}(T_k) = R^{KV}(T(\alpha'_k)), \tag{1.30}$$

kde pravá strana (1.30) je odhad metodou křížové validace stromu $T_k = T(\alpha'_k)$ definovaný v (1.29). Tedy odhad metodou křížové validace stromu T_k se spočítá podle (1.29) za použití $N_{i,j}$, což je součet hodnot $N_{i,j}^{(v)}$ získaných ze stromů $T^{(v)}(\alpha'_k)$ pro každé v. Nakonec je zvolen nejlepší strom T_{k_0} z posloupnosti stromů $T_1 \succ T_2 \succ \ldots \succ \{t_1\}$, zkonstruované na trénovacím výběru $\mathcal{L}$, jako

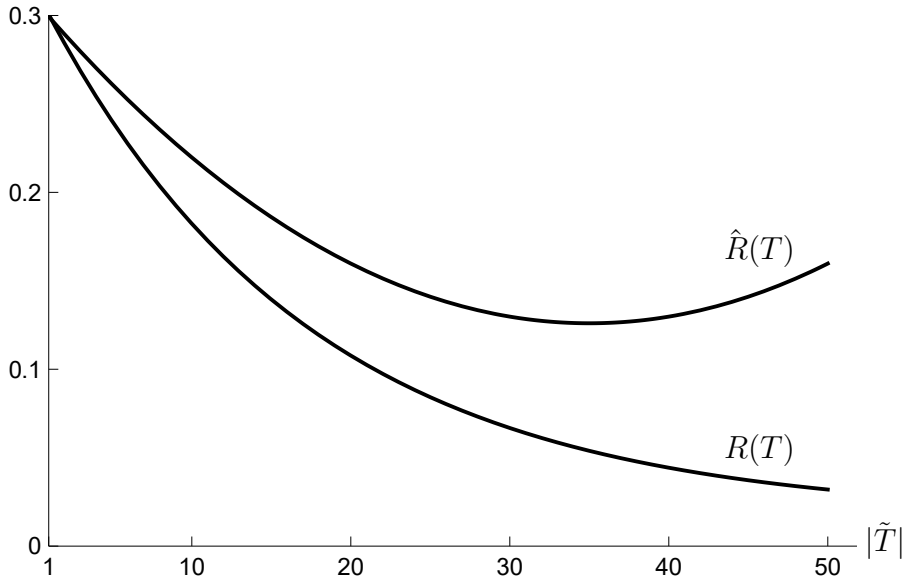
$$R^{KV}(T_{k_0}) = \min_k R^{KV}(T_k).$$

Podobně jako u metody odhadu pomocí testovacího výběru, i zde je možnost volit množiny $\mathcal{L}_1,\ldots,\mathcal{L}_V$ tak, aby mezi nimi byly jednotlivé třídy rovnoměrně zastoupeny. Takový postup dává znatelně přesnější odhady v problémech, kdy nějaké třídy v $\mathcal{L}$ nejsou příliš početné. Odhad metodou křížové validace je vhodnou volbou pro menší trénovací výběry, jelikož nevyžaduje vytvoření testovacího výběru.

1SE Pravidlo

Na obrázku 1.7 je možné vidět typický průběh funkcí resubstitučního odhadu $R(T)$ a spravedlivého odhadu $\hat{R}(T)$ v závislosti na počtu koncových uzlů. Jelikož se nejlepší strom volí na základě minimálního spravedlivého odhadu, zaměřme se nyní na funkci $\hat{R}(T)$. Na grafu je vidět, že funkce je okolo minima přibližně rovnoběžná s horizontální osou. To znamená, že více různě velkých stromů okolo stromu T_{k_0} má podobně velkou hodnotu spravedlivého odhadu. Proto může být například při malém pozměnění trénovacího výběru $\mathcal{L}$ vybrán jako nejlepší strom jiný strom než T_{k_0}.

Z toho důvodu se zaměříme na standardní chybu takových odhadů, abychom byli schopni říct, jak přesný daný odhad je a jestli není v mezích standardní chyby nějaký stejně dobrý, ale menší strom. Zaměříme se tedy na odhad R^{TV}, a jeho výpočet poté aplikujeme i na odhad R^{KV}. Nechť je strom T zkonstruovaný pomocí $\mathcal{L}_1$ a testovaný výběrem $\mathcal{L}_2$. Označme pravděpodobnost, že je případ z $\mathcal{L}_2$ chybně klasifikovaný stromem T, jako p^*. Potom je klasifikace N_2 případů z $\mathcal{L}_2$ binomický problém N_2 nezávislých náhodných veličin s pravděpodobnosti úspěchu p^*. Úspěch ovšem znamená chybnou klasifikaci případu. Odhadneme



Obrázek 1.7: Funkce odhadů

hodnotu p^* jako proporci chybně klasifikovaných případů z $\mathcal{L}_2$ a označíme ji jako p. Potom platí

$$\begin{aligned} \mathrm{E}\,(p) &= p^* \\ \mathrm{Var}(p) &= \frac{p^*(1-p^*)}{N_2} \end{aligned}$$

Následně se standardní chyba odhadu $R^{TV}(T)$ vypočítá jako

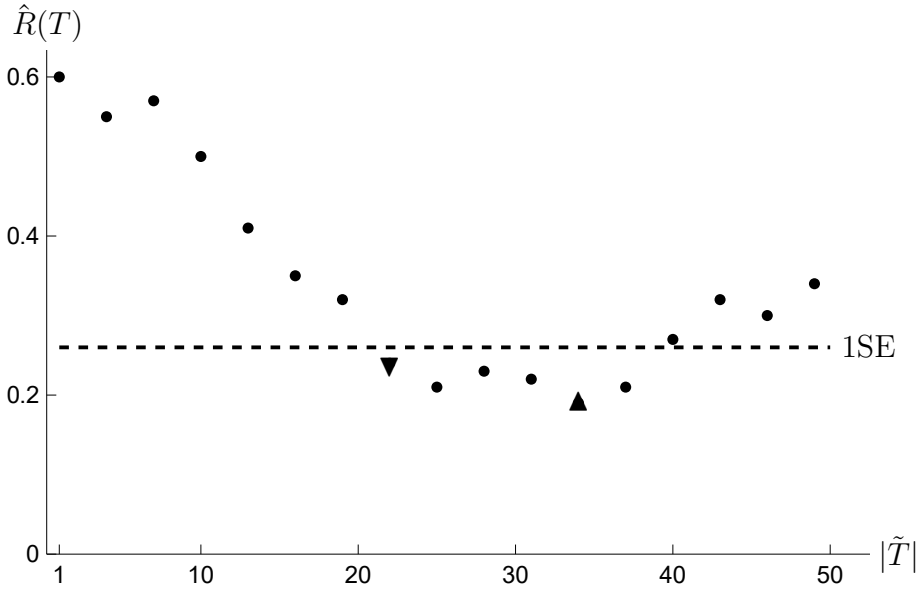
$$\mathrm{SE}(R^{TV}(T)) = \sqrt{\frac{R^{TV}(T)(1-R^{TV}(T))}{N_2}}. \tag{1.31}$$

Nahrazením R^{TV} odhadem R^{KV} v (1.31) dostaneme standardní chybu odhadu míry chybné klasifikace metodou křížové validace.

Nyní bude definováno 1SE pravidlo. Nechť $T_1 \succ T_2 \succ \ldots \succ \{t_1\}$ je posloupnost ideálních stromů a nechť je zvolen jako nejlepší strom T_{k_0} jako $\min_k \hat{R}(T_k)$. Potom je jako nový nejlepší strom podle 1SE pravidla zvolen strom T_{k_1} s nejméně koncovými uzly takový, který splňuje

$$\hat{R}(K_{k_1}) \leq \hat{R}(T_{k_0}) + SE(\hat{R}(T_{k_0})).$$

Obrázek 1.8 graficky vysvětluje 1SE pravidlo. Jako nejlepší strom byl podle nejmenší hodnoty spravedlivého odhadu $\hat{R}(T)$ zvolen strom odpovídající odhadu ▲. Přerušovanou přímkou je znázorněná standardní chyba odhadu ▲ a podle 1SE pravidla byl tedy zvolen nejlepší strom s odhadem ▼. Všimněme si, že spravedlivý odhad má zpravidla pro rostoucí počet koncových uzlů nejdříve sestupnou tendenci, poté se drží v nížině pro odhady kandidátů na nejlepší strom a pro velké množství konečných uzlů odhad opět roste.



Obrázek 1.8: 1SE pravidlo

1.5 Problém velkých dat

Častý problém stromových struktur je jejich složitost a výpočetní náročnost. Uvažujme situaci, kdy je potřeba zkonstruovat klasifikační strom, ale množství případů v trénovacím výběru je příliš velké. Nedávalo by žádný smysl k růstu stromu používat celý výběr $\mathcal{L}$, jelikož by to bylo příliš výpočetně náročné. Potřebujeme tedy naleznout nějakou metodu, jak vybrat z trénovacího výběru rozumně velkou sadu případů tak, aby výsledný strom byl dostatečně efektivní.

Předpokládejme tedy, že máme trénovací výběr $\mathcal{L}$ a je potřeba sestrojit klasifikační strom. Největší problém vzniká v počátečním uzlu a v uzlech jemu blízkých, jelikož se pro nalezení nejlepšího dělení v těchto uzlech používá celý nebo velká část trénovacího výběru. Proto představíme metodu podvzorkování.

Uvažujme dodanou konstantu N_0, která se volí individuálně na základě povahy problému. Pokud se při dělení do jakéhokoliv uzlu t dostane více než N_0 případů, potom je vybráno za jistých pravidel pouze N_0 případů, které jsou pak použity k nalezení nejlepšího dělení. Připomeňme, že $N_j(t)$, $j = 1, \ldots, J$, je počet případů v uzlu t třídy j. Nechť je v uzlu t více případů než N_0, tedy platí

$$\sum_{j=1}^{J} N_j(t) = N(t) > N_0. \tag{1.32}$$

Potom vybere ze všech případů vzorek $N_1', \ldots, N_J'$ případů splňující

1. $N_j' \leq N_j(t)$, $j = 1, \ldots, J$,
2. $\sum_{j=1}^{J} N_j' = N_0$,
3. $\sum_{j=1}^{J} (N_j' - \frac{N_0}{J})^2 = \text{minimum}$.

Třetí podmínka říká, že N_j' mají být zvoleny tak, aby byly co možná nejvíce stejně velké. Pro jednoduchost teď uvažujme, že N_0 je dělitelné J. Potom, pokud pro

každé j platí

$$N_j(t) \geq \frac{N_0}{J},$$

tak volba je jednoduchá. Z $N(t)$ případů v uzlu t pro každou třídu j vybereme náhodně N_0/J případů. Tedy bude platit

$$N'_1 = \ldots = N'_J.$$

Problém ovšem nastává, pokud alespoň pro jednu třídu j platí

$$N_j(t) < \frac{N_0}{J}. \tag{1.33}$$

Za platnosti (1.33) tedy seřaďme třídy následovně

$$N_1(t) \leq N_2(j) \leq \ldots \leq N_J(t).$$

První hodnotu N'_1 zvolme jako

$$N'_j = N_1(t)$$

a ostatní hodnoty $N'_2, \ldots N'_J$ zvolme rekurzivně podle vzorce

$$N'_{j+1} = \min\left(N_{j+1}(t), \frac{N_0 - N'_1 - N'_2 - \ldots - N'_j}{J - j}\right)$$

pro $j = 1, \ldots, J-1$. Tento proces určí N'_1 a následně zprůměruje zbývající potřebný počet případů $(N_0 - N'_1)$ mezi zbylé třídy. Pokud má $N_2(t)$ více případů než je požadovaný průměr, problém je vyřešený. Pokud ovšem $N_2(t)$ tolik případů neobsahuje, potom $N'_2 = N_2(j)$ a zbylý potřebný počet případů $(N_0 - N'_1 - N'_2)$ je opět zprůměrován mezi zbývající třídy. Tento proces pokračuje do doby, dokud nějaké $N_j(t)$ nesplňuje

$$N_j(t) \geq \frac{N_0 - N'_1 - \ldots - N'_{j-1}}{J - j - 1}.$$

Na uzlu t je poté nalezeno nejlepší dělení s^* pouze za použití vybraného vzorku.

Jelikož je ovšem použita metoda podvzorkování, která v uzlech maže rozdíly počtů případů mezi třídami, musíme definovat váhy $w_1, \ldots, w_J$, s jejichž pomocí budeme schopni vyjádřit $p(j|t)$, p_L a p_R závislé pouze na vahách w_j a hodnotách $N'_1, \ldots, N'_J$.

Předpokládejme nyní, že je k nalezení nejlepšího dělení zvoleno Giniho kritérium. To vyžaduje znát pravděpodobnosti příslušení případu do třídy j v uzlu t, tedy $p(1|t), \ldots, p(J|t)$. Rozepišme $p(j|t)$ jako

$$p(j|t) = \frac{p(j,\, t)}{p(t)} = \frac{\pi(j)\frac{N_j(t)}{N_j}}{\sum_{j=1}^{J} \pi(j)\frac{N_j(t)}{N_j}}. \tag{1.34}$$

Připomeňme, že $\pi(j)$ označuje pravděpodobnost příslušení do třídy j. Snažíme se najít váhy w_j, $j = 1, \ldots, J$, takové, abychom byli schopni vyjádřit pravděpodobnost $p(j|t)$ jako vztah vah w_j a hodnot $N'_1, \ldots, N'_J$. Z toho důvodu rozšiřme

čitatel i jmenovatel pravé strany vztahu (1.34) zlomkem N'_j/N'_j. Tedy

$$p(j|t) = \frac{\frac{\pi(j)N_j(t)}{N_j N'_j} N'_j}{\sum_{j=1}^{J} \frac{\pi(j)N_j(t)}{N_j N'_j} N'_j}. \tag{1.35}$$

Stačí tedy definovat w_j jako

$$w_j = \frac{\pi(j)N_j(t)}{N_j N'_j}.$$

Dále označíme pravděpodobnost $p(j|t)$ definovanou pomocí vah w_j a hodnot N'_j jako $p'(j|t)$. Potom platí

$$p'(j|t) = \frac{w_j N'_j}{\sum_{j=1}^{J} w_j N'_j}$$

a zároveň

$$p(j|t) = p'(j|t),$$

což jsme požadovali. Následně je v uzlu t na vybraném vzorku použito Giniho kritérium s pravděpodobnostmi $p'(1|t), \ldots, p'(J|t)$ a je nalezeno nejlepší dělení s^*.

Pokud je k určování nejlepšího dělení vybráno Twoing kritérium, budeme potřebovat definovat hodnoty $p'(j|t_L)$, $p'(j|t_R)$, $p'(t_L)$ a $p'(t_R)$ závislé pouze na vahách w_j a hodnotách $N'_1, \ldots, N'_J$. Položme tedy $N'_{j,L}$ a $N'_{j,R}$ jako počet případů z N'_j, které spadnou do uzlu t_L, respektive do t_R. Potom hodnoty $p'(j|t_L)$ a $p'(j|t_R)$ definujeme následovně

$$p'(j|t_L) = \frac{w_j N'_{j,L}}{\sum_{j=1}^{J} w_j N'_{j,L}}, \qquad p'(j|t_R) = \frac{w_j N'_{j,R}}{\sum_{j=1}^{J} w_j N'_{j,R}}. \tag{1.36}$$

Hodnoty $p'(t_L)$ a $p'(t_R)$ vezmeme podle pravidla podmíněné pravděpodobnosti jako

$$p'(t_L) = \sum_{j=1}^{J} w_j N'_{j,L}, \qquad p'(t_R) = \sum_{j=1}^{J} w_j N'_{j,R}. \tag{1.37}$$

Následně je na vybraném vzorku v uzlu t použito Twoing kritérium s pravděpodobnostmi definovanými v (1.36) a (1.37). S jejich pomocí je nalezeno nejlepší dělení s^*.

Jakmile je na uzlu t nalezeno nejlepší dělení, jsou všechny případy z t daným dělením děleny do dceřiných uzlů t_L a t_R. Jestliže nějaký dceřiný uzel splňuje nerovnost (1.32), opět je použita metoda podvzorkování. Pakliže ovšem uzel nerovnost nesplňuje, tedy neobsahuje více než N_0 případů, je pro nalezení nejlepšího dělení v daném uzlu využito všech příslušných případů.

2. Regresní stromy

Regresní strom se, na rozdíl od klasifikačního, využívá v problémech, kdy je predikovaná proměnná, dále označovaná jako odezva, spojitá. Již tedy nebude potřeba klasifikovat případ do určité třídy, ale úkolem mu bude na základě prediktorů přiřadit odpovídající predikovanou hodnotu. Ačkoliv se konstrukce regresního stromu od klasifikačního tolik neliší, je struktura regresního stromu o mnoho jednodušší. Není potřeba předpokládat náklady chybné klasifikace a řešit problémy s náležitostí do tříd, jako to bylo nutné u klasifikačního stromu. Dříve, než se zaměříme na samotnou konstrukci regresního stromu, je potřeba představit samotnou myšlenku regrese a definovat vhodné pojmy.

2.1 Úvod do regrese

Nejdříve je potřeba specifikovat strukturu dat. Ta je totožná jako u klasifikačního problému s jedinou výjimkou, a tou je, že predikovaná proměnná y je nyní spojitá. To znamená, že nabývá hodnot na intervalu $(-\infty, +\infty)$. Předpokládáme standardní strukturu dat. Tedy $\mathcal{X}$ je M-dimenzionální prostor všech vektorů měření tvaru $\boldsymbol{x} = (x_1, \ldots, x_M)$. Potom je d regresní funkce z prostoru $\mathcal{X}$ do reálných čísel:

$$d : \boldsymbol{x} \to d(\boldsymbol{x}),\ \forall \boldsymbol{x} \in \mathcal{X},\ d(\boldsymbol{x}) \in \mathbb{R}.$$

Smyslem regresní funkce je tedy přiřadit vektoru měření $\boldsymbol{x}$ hodnotu odezvy. Nechť $\mathcal{L}$ je trénovací výběr N složek tvaru

$$\mathcal{L} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_N, y_N)\}.$$

Pomocí výběru $\mathcal{L}$ a metodou nejmenších čtverců se zkonstruuje odhad regresní funkce d, který následně bude predikovat hodnotu odezvy budoucích případů. Přesnost regresního prediktoru se bude odhadovat reziduálním součtem čtverců. To je nejpoužívanější metoda, jednoduše interpretovatelná a její výpočet je jednoduchý. Trénovací výběr $\mathcal{L}$ je opět rozdělen na disjunktní výběry, a to trénovací podvýběr $\mathcal{L}_1$ a testovací výběr $\mathcal{L}_2$. Předpokládejme nyní, že byl odhad d^* funkce d zkonstruován pomocí $\mathcal{L}_1$, a my chceme zjistit jeho přesnost. K tomu nám poslouží střední čtvercová chyba tvaru

$$\frac{1}{N_2} \sum_{n=1}^{N_2} (y'_n - d^*(\boldsymbol{x}'_n))^2,$$

kde

$$\mathcal{L}_2 = \{(\boldsymbol{x}'_1, y'_1), \ldots, (\boldsymbol{x}'_{N_2}, y'_{N_2})\}$$

je testovací výběr. Střední čtvercová chyba tedy sčítá kvadráty odchylek (residuí) odezev případů od jejich opravdových hodnot, a následně tento součet průměruje mezi všech N_2 případů testovacího výběru.

Uvažujme nyní rozdělení na prostoru $\mathcal{X} \times \mathbb{R}$ všech dvojic tvaru $(\boldsymbol{x}, y)$. Nechť je trénovací výběr $\mathcal{L}$ náhodný výběr z tohoto rozdělení a $(\boldsymbol{X}, Y)$ náhodný nový vektor z téhož rozdělení. Předpokládejme, že odhad regresní funkce d^* byl zkonstruován pomocí trénovacího výběru $\mathcal{L}$.

Definice 16 (Střední čtvercová chyba)**.** *Střední čtvercová chyba $R^*(d)$ regresní funkce d je definována jako*

$$R^*(d) = E[(y - d(\boldsymbol{X})^2].$$

Cílem regresního problému za daného trénovacího výběru $\mathcal{L}$ je tedy odhadnout regresní funkci d a následně spočítat střední čtvercovou chybu tohoto odhadu. K takovému odhadu se používají již dříve specifikované postupy. První z nich je resubstituční odhad

$$R(d^*) = \frac{1}{N}\sum_{n=1}^{N}(y_n - d^*(\boldsymbol{x}_n))^2,$$

kdy N je počet případů v $\mathcal{L}$. To znamená, že ze všech dat odhadneme d^* a ta samá data využijeme pro odhad kvality tohoto odhadu. Nyní opět předpokládejme rozdělení $\mathcal{L}$ na $\mathcal{L}_1$ a $\mathcal{L}_2$. Potom je druhou metodou odhad pomocí testovacího výběru, který má tvar

$$R^{TV}(d^*) = \frac{1}{N_2}\sum_{(\boldsymbol{x}_n, y_n)\in\mathcal{L}_2}(y_n - d^*(\boldsymbol{x}_n))^2.$$

Třetí metodou je odhad metodou křížové validace. $\mathcal{L}$ je náhodně rozdělen do V podobně velkých podmnožin a odhad $d^{*(v)}$ je zkonstruován pomocí $\mathcal{L} \setminus \mathcal{L}_v$. Potom má odhad metodou křížové validace odhadu d^* regresní funkce d sestrojené prostřednictvím $\mathcal{L}$ tvar

$$R^{KV}(d^*) = \frac{1}{N}\sum_{v=1}^{V}\sum_{(\boldsymbol{x}_n, y_n)\in\mathcal{L}_v}(y_n - d^{*(v)}(\boldsymbol{x}_n))^2.$$

Zde je ovšem problém, že hodnota střední čtvercové chyby závisí na měřítku hodnot, ve kterých je odezva měřená. To znamená, že jedna regresní funkce pro stejná data má rozdílnou střední čtvercovou chybu, pokud změníme měřítko hodnoty odezvy. Je tedy potřeba střední čtvercovou chybu nějakým způsobem standardizovat. Označme střední hodnotu odezvy $\mathrm{E}[Y]$ jako μ. Potom vypočítáme rozptyl odezvy následovně:

$$R^*(\mu) = \mathrm{Var}[Y] = \mathrm{E}[(Y - \mu)^2].$$

To nám umožní standardizovat $R^*(d)$.

Definice 17 (Relativní střední čtvercová chyba)**.** *Relativní střední čtvercovou chybu regresní funkce d a odezvy Y definujeme vztahem*

$$RE^*(d) = \frac{R^*(d)}{R^*(\mu)}.$$

Tato definice je ovšem pouze teoretická. V reálném problému střední hodnotu odezvy neznáme, a proto je potřeba ji odhadnout z dat. Nechť $\mathcal{L}$ je trénovací výběr obsahující N případů. Nestranným a konsistentním odhadem střední hodnoty odezvy z $\mathcal{L}$ je výběrový průměr

$$\bar{y} = \frac{1}{N}\sum_{n=1}^{N} y_n.$$

Následně můžeme odhadnout rozptyl odezvy z $\mathcal{L}$ jako

$$\mathrm{Var}[\bar{y}] = R(\bar{y}) = \frac{1}{N}\sum_{n=1}^{N}(y_n - \bar{y})^2. \tag{2.1}$$

Pro odhad regresní funkce d zkonstruovaný za použití $\mathcal{L}$ následně můžeme standardizovat resubstituční odhad, odhad trénovacího výběru a odhad metodou křížové validace následovně:

$$RE(d^*) = \frac{R(d^*)}{R(\bar{y})}, \qquad RE^{TV}(d^*) = \frac{R^{TV}(d^*)}{R^{TV}(\bar{y})}, \qquad RE^{KV}(d^*) = \frac{R^{KV}(d^*)}{R(\bar{y})},$$

kdy $R^{TV}(\bar{y})$ je odhad rozptylu odezvy testovacího výběru $\mathcal{L}_2$.

Definujme nyní RSS (residual sum of squares) a TSS (total sum of squares) jako

$$\mathrm{RSS} = \sum_{n=1}^{N}(y_n - d^*(\boldsymbol{x}_n))^2,$$
$$\mathrm{TSS} = \sum_{n=1}^{N}(y_n - \bar{y})^2.$$

RSS vyjadřuje celkový rozptyl všech reziduí za použití odhadu d^* a TSS vyjadřuje celkový rozptyl odezvy v trénovacím výběru. Proto výraz $(\mathrm{TSS} - \mathrm{RSS})$ vyjadřuje celkový rozptyl odezvy, který zůstal nevysvětlený po použití d^*. Dále definujeme koeficient determinace $\mathcal{R}^2$ jako podíl nevysvětleného rozptylu odezvy za použití d^* a platí

$$\mathcal{R}^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}} = 1 - RE(d^*).$$

$\mathcal{R}^2$ zřejmě nabývá hodnot na intervalu [0,1]. Jelikož je smyslem odhadu d^* vysvětlit vztah mezi prediktory a odezvou, tedy vysvětlit rozptyl odezvy co možná nejvíce, potom je nejlepší odhad d^* na $\mathcal{L}$ takový, který maximalizuje míru $\mathcal{R}^2$. To odpovídá tomu volit takový odhad d^*, který minimalizuje resubstituční odhad $R(d^*)$. koeficient determinace $\mathcal{R}^2$ je ve statistice, zejména v regresi hojně používaná a je to jeden z hlavních ukazatelů kvality odhadu regresní funkce. Více o této problematice lze najít v knize Garetha Jamese (James a kol., 2021).

2.2 Konstrukce regresního stromu

Regresní strom má podobnou strukturu jako strom klasifikační. Má počáteční uzel t_1, každý nekoncový uzel t je dělen nejlepším dělením s^* na dva dceřiné uzly, t_L a t_R, a obsahuje koncové uzly. Jediný rozdíl je v tom, že každý koncový uzel t má přidělenou hodnotu $y(t)$, která je následně přiřazená případu, který v daném uzlu skončí. Stejně jako konstrukce klasifikačního stromu, má i konstrukce regresního stromu tři nejdůležitější problémy. Prvním problémem je, jak najít nejlepší dělení v každém nekoncovém uzlu. Druhý problém je přiřazení hodnoty $y(t)$ každému koncovému uzlu na základě nějakého pravidla. Třetím problémem je vytvoření posloupnosti stromů a určit kritérium, které z ní vybere ten nejlepší.

2.2.1 Pravidlo pro přiřazení hodnoty uzlu

Uvažujme již zkonstruovaný regresní strom T jako regresní funkci d_T. Potřebujeme rozhodnout, jakou hodnotu $y(t)$ přiřadit každému koncovému uzlu. Nejlepší regresní funkce d_T bude taková, která minimalizuje resubstituční odhad

$$R(d_T) = \frac{1}{N}\sum_{n=1}^{N}(y_n - d_T(\boldsymbol{x}_n))^2.$$

Předpokládejme nyní, že strom T má $|\tilde{T}|$ koncových uzlů a nechť t je jeden z nich. Potom na základě všech případů, které skončily v uzlu t při konstrukci T, potřebujeme danému uzlu přiřadit hodnotu $y(t)$ takovou, aby celkový resubstituční odhad stromu T byl co nejmenší. Proto je v každém koncovém uzlu t potřeba zvolit $y(t)$ tak, aby byla minimalizována jeho střední čtvercová chyba koncového uzlu

$$R(t) = \sum_{(\boldsymbol{x}_n, y_n)\in t}(y_n - y(t))^2. \tag{2.2}$$

Hodnota $y(t)$, která minimalizuje (2.2) je výběrový průměr odezvy případů padnoucích do uzlu t označovaný jako $\bar{y}(t)$. Tedy

$$y(t) = \bar{y}(t) = \frac{1}{N(t)}\sum_{(\boldsymbol{x}_n, y_n)\in t} y_n.$$

Proto každému koncovému uzlu přiřadíme hodnotu $y(t)$ jako výběrový průměr odezvy jeho případů. Následně definujeme resubstituční odhad $R(T)$, od teď nazývaný resubstituční chyba stromu T, jako součet středních čtvercových chyb jeho koncových uzlů a platí

$$R(T) = \sum_{t\in\tilde{T}} R(t) = \frac{1}{N}\sum_{t\in\tilde{T}}\sum_{(\boldsymbol{x}_n, y_n)\in t}(y_n - \bar{y}(t))^2,$$

jelikož

$$\sum_{t\in\tilde{T}} N(t) = N.$$

Odhad d^* regresní funkce d lze tedy chápat jako po částech konstantní funkci. Tyto „části“ jsou právě koncové uzly, které rozkládají celý prostor $\mathcal{X}$.

2.2.2 Pravidlo pro dělení uzlů

Nechť je dána množina S všech dělení s koncového uzlu t ve stromu T.

Definice 18 (Nejlepší dělení)**.** *Nejlepší dělení $s^* \in S$ koncového uzlu je takové, které nejvíce snižuje resubstituční chybu $R(T)$ stromu T.*

Po vzoru poklesu nečistoty definujeme pokles resubstituční chyby při dělení s v uzlu t jako

$$\Delta R(s,t) = R(t) - R(t_L) - R(t_R).$$

Potom je podle Definice 18 nejlepší dělení s^* na koncovém uzlu t takové, které maximalizuje pokles resubstituční chyby. Proto

$$\Delta R(s^*, t) = \max_{s\in S}\Delta R(s, t).$$

Takové dělení maximalizuje pokles resubstituční chyby a splňuje proto Definici 18.

2.2.3 Konstrukce a výběr nejlepšího stromu

Jakmile byla definována resubstituční chyba uzlu $R(t)$ a stromu $R(T)$, je možné definovat metodu ořezávání s minimální chybovou složitostí pro regresní stromy. Pro uzel t a libovolné dělení s zřejmě platí

$$R(t) \geq R(t_L) + R(t_R).$$

Proto nejprve zkonstruujeme strom T_{MAX}, který má v každém koncovém uzlu méně případů než N_{MIN}. N_{MIN} se většinou volí jako 1,2 nebo 5. poté definujeme míru chybové složitosti $R_\alpha(T)$ jako

$$R_\alpha(T) = R(T) + \alpha|\tilde{T}|.$$

Takto definovaná míra je shodná s mírou nákladové složitosti až na to, že se místo míry chybné klasifikace použije resubstituční chyba. Ořezávání s minimální chybovou složitostí následně proběhne totožně jako ořezávání s minimální nákladovou složitostí a výstupem je posloupnost ideálních stromů

$$T_1 \succ T_2 \succ \ldots \succ \{t_1\},$$

kde

$$T_1 \preccurlyeq T_{MAX}.$$

Stejně jako u metody ořezávání s minimální nákladovou složitostí, i zde je posloupnost $0 = \alpha_1 < \alpha_2 < \ldots$, pro kterou platí, že pro $\alpha_k < \alpha < \alpha_{k+1}$ splňuje T_k rovnost

$$R_\alpha(T_k) = \min_{T \preccurlyeq T_{MAX}} R_\alpha(T).$$

2.2.4 Volba nejlepšího stromu pomocí spravedlivého odhadu

Stejně jako u klasifikačního problému, i zde je potřeba z posloupnosti ideálních stromů vybrat ten nejlepší. Začněme s metodou odhadu pomocí testovacího výběru. Trénovací výběr $\mathcal{L}$ je náhodně rozdělen na disjunktní výběry $\mathcal{L}_1$ a $\mathcal{L}_2$. S pomocí $\mathcal{L}_1$ je nalezena posloupnost ideálních stromů. Nechť je d_{T_k} regresní funkce odpovídající stromu T_k a definujeme odhad pomocí testovacího výběru pro strom T_k jako

$$R^{TV}(T_k) = \frac{1}{N_2} \sum_{(\boldsymbol{x}_n, y_n) \in \mathcal{L}_2} (y_n - d_{T_k}(\boldsymbol{x}))^2.$$

Potom zvolíme z posloupnosti ideálních stromů takový strom T_{k_0}, který splňuje

$$R^{TV}(T_{k_0}) = \min_k R^{TV}(T_k).$$

Další metodou spravedlivého odhadu je odhad metodou křížové validace. Trénovací výběr $\mathcal{L}$ je rozdělen na V stejně velkých disjunktních množin $\mathcal{L}_1, \ldots, \mathcal{L}_V$. Pro každé v, $v = 1, \ldots, V$, představuje strom $T^{(v)}(\alpha)$ ideální strom pro parametr α zkonstruovaný pomocí trénovacího výběru $\mathcal{L} \setminus \mathcal{L}_v$. Pomocí celého výběru $\mathcal{L}$ je

sestavena posloupnost ideálních stromů $T_1 \succ \ldots \succ \{t_1\}$ a definujeme geometrický průměr pro tuto posloupnost jako

$$\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}.$$

Nechť je $d^{(v)}_{T_k^{(v)}}$ regresní funkce odpovídající stromu $T^{(v)}(\alpha'_k)$. Potom má odhad metodou křížové validace stromu T_k podobu

$$R^{KV}(T_k) = \frac{1}{N} \sum_{v=1}^{V} \sum_{(\boldsymbol{x}_n, y_n) \in \mathcal{L}_v} (y_n - d^{(v)}_{T_k^{(v)}}(\boldsymbol{x}))^2.$$

Na základě takového odhadu zvolíme nejlepší strom T_{k_0} z posloupnosti

$$T_1 \succ \ldots \succ \{t_1\}$$

takový, který splňuje

$$R^{KV}(T_{k_0}) = \min_k R^{KV}(T_k).$$

Nyní představíme možnou modifikaci metody křížové validace, která produkuje přesnější odhad. Seřadíme všechny případy z trénovacího výběru $\mathcal{L}$ vzestupně podle hodnoty odezvy y. Následně rozdělíme případy na V podmnožin po N/V případech tak, aby první podmnožina obsahovala N/V případů s nejmenšími hodnotami odezvy, druhá podmnožina obsahovala N/V případů s druhými nejmenšími hodnotami odezvy a tak dále. Potom je do každého testovacího výběru $\mathcal{L}_v$ vybrán jeden případ z každé podmnožiny bez vracení. Toto rozdělení do výběrů $\mathcal{L}_1, \ldots, \mathcal{L}_V$ zaručí, že každý výběr bude mít podobné rozdělení odezvy a výsledný odhad bude přesnější.

Ze stejného důvodu, jako u klasifikačního problému, je výhodné použít 1SE pravidlo. Bez újmy na obecnosti předpokládejme, že k vybrání nejlepšího stromu byl použit odhad pomocí testovacího výběru. Potom jako nový nejlepší strom T_{k_1} zvolíme strom splňující

$$R^{TV}(K_{k_1}) \leq R^{TV}(T_{k_0}) + \mathrm{SE}(R^{TV}(T_{k_0})),$$

kdy $\mathrm{SE}(R^{TV}(T_{k_0}))$ je standardní chyba odhadu $R^{TV}(T_{k_0})$. U regresních stromů je ovšem tato volba velmi důležitá, jelikož posloupnost ideálních stromů je větší než u klasifikačního problému a mnoho stromů má odhad podobně velký tomu minimálnímu. To je zapříčiněno tím, že proces ořezávání s minimální chybovou složitostí ořezává většinou v každém kroku pouze dva koncové uzly. To je dáno tím, že $R(t)$ je ve většině případů větší než nula, a proto proces ořeže pouze nejmenší možnou větev v každém kroku.

2.3 Problém velkých dat

Co se týče problému velkých dat u regresních stromů, je celá metoda výrazně jednodušší než v případě stromů klasifikačních. Nechť $\mathcal{L}$ je trénovací výběr s N případy a je dodána konstanta N_0. Jestliže se do libovolného uzlu t při dělení dostane více případů než N_0, je náhodně z oněch $N(t)$ případů vybráno N_0 případů. S jejich použitím se následně nalezne nejlepší dělení s^*. Po nalezení

takového dělení je všech $N(t)$ případů v uzlu t děleno za použití s^* a pokračuje se dále stromem, stejně jako v případě klasifikačního stromu. Tato metoda je nejvíce efektivní v problémech, kdy N je mnohonásobně větší než N_0. Jak ale zvolit konstantu N_0? Neexistuje žádné univerzální pravidlo. Je tedy na uvážení datového analytika, jak ji zvolí.

3. Bayesovské stromy

V této kapitole se zaměříme na Bayesovské klasifikační a regresní stromy. Bayesovským stromem, či pravidlem, je takové rozhodovací pravidlo, který je na daném pravděpodobnostním rozdělení ze všech nejpřesnější. To je ovšem skoro nemožné dosáhnout, jelikož je potřeba znát celé pravděpodobnostní rozdělení. V reálných případech je ovšem možné takové rozdělení odhadnout. Nyní bude představen pravděpodobnostní model, který počítá s tím, že je takové rozdělení známé. Dále v této kapitole potom budou představeny metody, které za určitých předpokladů odhadují sdruženou hustotu celého stromu a dávají tím prostor se na celý problém dívat z jiného úhlu. Představují úplně odlišné metody pro tvoření klasifikačních a regresních stromů, než se kterými se zatím bylo možné v této práci seznámit.

3.1 Pravděpodobnostní model

Předpokládejme, že d je rozhodovací pravidlo, která přiřadí vektoru $\boldsymbol{x}$ predikovanou proměnnou y. Predikovaná proměnná může být kategorická nebo numerická. Nechť je nyní $\mathcal{X}$ prostor všech možných vektorů $\boldsymbol{x}$ a A prostor všech uvažovaných hodnot proměnné y. Vezmeme-li náhodný vektor $\boldsymbol{X} \in \mathcal{X}$ s pravděpodobností $P(d\boldsymbol{x})$ a jeho reálnou hodnotu proměnné Y, potom pro bayesovské rozhodovací pravidlo d_B platí

$$P(d_B(\boldsymbol{X}) \neq Y) \leq P(d(\boldsymbol{X}) \neq Y).$$

Tedy d_B je ze všech rozhodovacích pravidel nejpřesnější.

Uvažujeme-li hodnotu a jako predikovanou proměnnou, kterou pravidlo vektoru $\boldsymbol{X}$ přiřadilo, tedy

$$a = d(\boldsymbol{X}),$$

potom $L(Y, a)$ je ztráta, která vznikne přiřazením a za podmínky, že opravdová hodnota predikované proměnné je Y. Následně je tedy možné definovat rizikovou funkci jako střední hodnotu ztráty za použití pravidla d, tedy

$$R(d) = \mathrm{E}\ \ L(Y, d(\boldsymbol{X})). \tag{3.1}$$

To byl ovšem obecný tvar, který se liší v závislosti na tom, zda se jedná o klasifikační nebo regresní problém. V případě klasifikačního problému je A konečná množina tříd $\{1, \ldots, J\}$. Proto se volí ztráta jako náklad chybné klasifikace případu třídy y do třídy a:

$$L(y, a) = C(a|y).$$

Potom je $R(d)$ je riziková funkce chybné klasifikace za použití pravidla d.

Pro regresní problém je ztráta měřena jako v metodě nejmenších čtverců, tedy

$$L(y, a) = (y - a)^2.$$

Ztráta $L(y,a)$ je tedy druhá mocnina rozdílu reálné odezvy a její predikované hodnoty. Potom riziko při použití pravidla d je střední hodnota ztráty a platí

$$R(d) = \mathrm{E}\,[(Y - d(\boldsymbol{X})^2].$$

Přepišme nyní (3.1) jako podmíněnou střední hodnotu, tedy

$$R(d) = \mathrm{E}\,\mathrm{E}\,[L(Y,\, d(\boldsymbol{X}))|\boldsymbol{X}]. \tag{3.2}$$

Pravou stranu rovnosti (3.2) lze upravit na integrál podmíněné střední hodnoty podle rozdělení $\boldsymbol{X}$, tedy

$$R(d) = \int \mathrm{E}\,[L(Y,d(\boldsymbol{x}))|\boldsymbol{X} = \boldsymbol{x}]P(d\boldsymbol{x}). \tag{3.3}$$

Bayesovské pravidlo d_B je nejpřesnější ze všech, tedy minimalizuje $R(d)$. Proto pro každé $\boldsymbol{x} \in X$ za platnosti $a = d(\boldsymbol{x})$ platí

$$R(d_B) = \int \min_a \mathrm{E}\,[L(Y,\, a)|\boldsymbol{X} = \boldsymbol{x}]P(d\boldsymbol{x}). \tag{3.4}$$

Následující věta nám dává zajímavý vztah Bayesovského pravidla s jakýmkoliv jiným rozhodovacím pravidlem při regresním problému.

Věta 5. *Nechť d_B je Bayesovské pravidlo a d je pravidlo na prostoru X. Potom platí*

$$R(d) = R(d_B) - E[(d_B(\boldsymbol{X}) - d(\boldsymbol{X}))^2]. \tag{3.5}$$

Důkaz. Nejdříve definujme $\mu(\boldsymbol{x})$ jako

$$\mu(\boldsymbol{x}) = \mathrm{E}\,[Y|\boldsymbol{X} = \boldsymbol{x}],\ \boldsymbol{x} \in X.$$

Potom platí

$$\mathrm{E}\,[Y - \mu(\boldsymbol{x})|\boldsymbol{X} = \boldsymbol{x}] = 0. \tag{3.6}$$

Nyní lze psát

$$\begin{aligned}
&\mathrm{E}\,[L(Y,\, a)|\boldsymbol{X} = \boldsymbol{x}] = \\
&\quad = \mathrm{E}\,[(Y - a)^2|\boldsymbol{X} = \boldsymbol{x}] = \\
&\quad = \mathrm{E}\,[(Y - \mu(\boldsymbol{x}) + \mu(\boldsymbol{x}) - a)^2|\boldsymbol{X} = \boldsymbol{x}] = \\
&\quad = \mathrm{E}\,[(Y - \mu(\boldsymbol{x}))^2 + 2(Y - \mu(\boldsymbol{x}))(\mu(\boldsymbol{x}) - a) + (\mu(\boldsymbol{x} - a))^2|\boldsymbol{X} = \boldsymbol{x}] = \\
&\quad = \mathrm{E}\,[(Y - \mu(\boldsymbol{x}))^2|\boldsymbol{X} = \boldsymbol{x}] + (\mu(\boldsymbol{x}) - a)^2.
\end{aligned} \tag{3.7}$$

V (3.7) jsme využili toho, že podle (3.6) platí

$$\mathrm{E}\,[(Y - \mu(\boldsymbol{x})(\mu(\boldsymbol{x}) - a)|\boldsymbol{X} = \boldsymbol{x}] = E[Y - \mu(\boldsymbol{x})|\boldsymbol{X} = \boldsymbol{x}]\,\mathrm{E}\,[\mu(\boldsymbol{x} - a)|\boldsymbol{X} = \boldsymbol{x}] = 0,$$

a také

$$\mathrm{E}\,[(\mu(\boldsymbol{x}) - a)^2|\boldsymbol{X} = \boldsymbol{x}] = (\mu(\boldsymbol{x}) - a)^2,$$

protože výraz $(\mu(\boldsymbol{x}) - a)^2$ je konstanta.

Jelikož pravidlo d_B a také $\mu(\boldsymbol{x})$ minimalizují $\mathrm{E}\,[L(Y,\, a)|\boldsymbol{X} = \boldsymbol{x}]$ pro každé $\boldsymbol{x} \in X$, tedy

$$\min_a \mathrm{E}\,[L(Y,\, a)|\boldsymbol{X} = \boldsymbol{x}] = \mathrm{E}\,[L(Y,\, \mu(\boldsymbol{x}))|\boldsymbol{X} = \boldsymbol{x}] = \mathrm{E}\,[(Y - \mu(\boldsymbol{x}))^2|\boldsymbol{X} = \boldsymbol{x}],$$

tak platí

$$\mu(\boldsymbol{x}) = d_B(\boldsymbol{x}).$$

Označíme-li výstup pravidla d jako a, tj. položíme-li

$$d(\boldsymbol{x}) = a,$$

potom můžeme vztah (3.7) přepsat následovně:

$$\mathrm{E}\left[L(Y,\, d(\boldsymbol{x}))|\boldsymbol{X} = \boldsymbol{x}\right] = \min_a \mathrm{E}\left[L(Y,\, a)|\boldsymbol{X} = \boldsymbol{x}\right] + (d_B(\boldsymbol{x}) - d(\boldsymbol{x}))^2.$$

Nyní lze celou rovnici vyintegrovat podle $P(d\boldsymbol{x})$ a dostaneme

$$\begin{aligned}\int \mathrm{E}\left[L(Y,\, d(\boldsymbol{x}))|\boldsymbol{X} = \boldsymbol{x}\right]P(d\boldsymbol{x}) = \\ = \int \min_a \mathrm{E}\left[L(Y,\, a)|\boldsymbol{X} = \boldsymbol{x}\right]P(d\boldsymbol{x}) + \int (d_B(\boldsymbol{x}) - d(\boldsymbol{x}))^2 P(d\boldsymbol{x}),\end{aligned}$$

což odpovídá vztahu (3.5) s využitím (3.3) a (3.4), a také vztahu

$$\int (d_B(\boldsymbol{x}) - d(\boldsymbol{x}))^2 P(d\boldsymbol{x}) = \mathrm{E}\left[(d_B(\boldsymbol{X}) - d(\boldsymbol{X}))^2\right].$$

Tím je věta dokázána.

□

Pro klasifikační problém označme pravděpodobnost $P(Y = j|\boldsymbol{X} = \boldsymbol{x})$ jako $P(j|\boldsymbol{x})$ a náklad chybné klasifikace jako $C(i|j)$. Potom Bayesovské pravidlo přiřadí každému vektoru $\boldsymbol{x} \in \mathcal{X}$ takovou třídu j, aby byla minimalizována střední hodnota ztráty pro danou hodnotu $\boldsymbol{x}$, tedy

$$\mathrm{E}\left[L(Y,\, i)|\boldsymbol{X} = \boldsymbol{x}\right] = \sum_j L(j,\, i)P(j|\boldsymbol{x}) = \sum_j C(i|j)P(j|\boldsymbol{x}).$$

Potom riziko Bayesovského pravidla d_B má tvar

$$R(d_B) = \int \min_i \left[\sum_j C(i|j)P(j|\boldsymbol{x})\right] P(d\boldsymbol{x}).$$

Doposud jsme předpokládali, že pravidlo přiřazuje hodnotu každému $\boldsymbol{x}$ bez ohledu na jeho příslušnost k nějakému koncovému uzlu. Nyní se podíváme na Bayesovské pravidlo za předpokladu rozdělení prostoru $\mathcal{X}$ na podmnožiny, jak tomu je v případě stromové struktury v koncových uzlech.

Předpokládejme proto rozdělení množiny $\mathcal{X}$ do $|\tilde{T}|$ disjunktních množin, pro které platí $P(t) > 0$ pro každé $t \in \tilde{T}$. To můžeme chápat tak, že

$$P[\exists\ \boldsymbol{x} \in \mathcal{X} :\ \boldsymbol{x} \in t] = 1$$

a $P(t)$ je pravděpodobnostní rozdělení množiny všech $\boldsymbol{x}$ náležících uzlu t. Potom lze definovat funkci $\mathcal{T}$ z množiny $\mathcal{X}$ do množiny $\tilde{T}$ takovou, že

$$\mathcal{T}(\boldsymbol{x}) = t \ \Leftrightarrow\ \boldsymbol{x} \in t,\, t \in \tilde{T}.$$

Takto definovaná funkce tedy odpovídá dělení stromu do koncových uzlů. Poté v každém koncovém uzlu t definujeme funkci $\omega(t)$, která všem prvkům $\boldsymbol{x} \in t$ přiřadí odpovídající hodnotu predikované proměnné. Potom je tedy možné zapsat rozhodovací pravidlo $d(\boldsymbol{x})$ ve tvaru

$$d(\boldsymbol{x}) = \omega(\mathcal{T}(\boldsymbol{x})).$$

Pro takto zkonstruované pravidlo lze následně definovat riziko předpisem

$$R(d) = \sum_{t \in \tilde{T}} \mathrm{E}\left[L(Y, \omega(t))|\boldsymbol{X} \in t\right] P(t). \tag{3.8}$$

Ze vztahu (3.8) vidíme, že pro $\boldsymbol{X} \in t$ platí

$$d(\boldsymbol{X}) = \omega(t) = a,$$

kde $a = \omega(t)$ minimalizuje riziko za předpokladu, že $\boldsymbol{X}$ padne do t. Pravidlo d tedy přiřadí stejnou hodnotu všem vektorům náležícím do jednoho koncového uzlu tak, jak to bylo v případě regresního stromu.

Nyní můžeme definovat Bayesovské rozhodovací pravidlo jako takové pravidlo, který v každém koncovém uzlu minimalizuje riziko zavedené v (3.8). Jinými slovy,

$$d_B(\boldsymbol{x}) = \mathcal{T}(\omega(\boldsymbol{x})) \tag{3.9}$$

platí právě tehdy, když v každém uzlu $t \in \tilde{T}$ hodnota $a = \omega(t)$ minimalizuje

$$\mathrm{E}\left[L(Y, a)|\boldsymbol{X} \in t\right]. \tag{3.10}$$

Pokud $a = \omega(t)$ minimalizuje (3.10), lze definovat riziko pro uzel t jako

$$r(t) = \mathrm{E}\left[L(Y, \omega(t))|\boldsymbol{X} \in t\right] = \min_a \mathrm{E}\left[L(Y, a)|\boldsymbol{X} \in t\right],$$

a následně

$$R(t) = r(t)P(t)$$

jako očekávané riziko koncového uzlu t. Potom pro Bayesovské pravidlo d_B s množinou koncových uzlů $\tilde{T}$ platí

$$R(d_B) = \sum_{t \in \tilde{T}} P(t)r(t) = \sum_{t \in \tilde{T}} R(t).$$

Pro regresní problém je zřejmé, že v každém koncovém uzlu t minimalizuje riziko $r(t)$ uzlu t střední hodnota odezvy $\mathrm{E}\left[Y|\boldsymbol{X} \in t\right]$. Proto pro Bayesovské pravidlo máme

$$\mu(t) = \omega(t) = \mathrm{E}\left[Y|\boldsymbol{X} \in t\right].$$

Podobně můžeme definovat i rozptyl v koncovém uzlu t jako

$$\sigma^2(t) = \mathrm{Var}(t) = \mathrm{Var}[Y|\boldsymbol{X} \in t] = \mathrm{E}\left[(Y - \mu(t))^2|\boldsymbol{X} \in t\right].$$

Všimněme si, že platí

$$\sigma^2(t) = \mathrm{E}\left[(Y - \mu(t))^2|\boldsymbol{X} \in t\right] = \mathrm{E}\left[L(Y, \mu(t))|\boldsymbol{X} \in t\right] = r(t).$$

Potom tedy

$$R(d_B) = \sum_{t \in \tilde{T}} P(t)\sigma^2(t) = \sum_{t \in \tilde{T}} P(t)\,\mathrm{E}\left[(Y - \mu(t))^2|\boldsymbol{X} \in t\right].$$

Pro klasifikační problém opět využijeme pravděpodobnosti $P(j|t)$. Vezměme tvar Bayesovského pravidla tak, jak bylo definované v (3.9). Pro každé $t \in \tilde{T}$ nabývá funkce $\omega(t)$ hodnotu některé třídy $i \in C$ tak, aby tato třída i minimalizovala výraz

$$\sum_j C(i|j)P(j|t).$$

Poté definujeme riziko pro uzel t jako

$$r(t) = \min_i \sum_j C(i|)P(j|t),$$

takže dostaneme riziko chybné klasifikace Bayesovského pravidla $R(d_B)$ jako

$$R(d_B) = \sum_{t \in \tilde{T}} r(t)P(t).$$

Následně se podíváme na to, jak pro Bayesovské pravidlo najít dělení uzlu. Je-li dána množina koncových uzlů $\tilde{T}$, zvolme $t \in \tilde{T}$ a uvažujme dělení s do dvou dceřiných uzlů t_L a t_R. Definujme pravděpodobnosti

$$P_L = P(\boldsymbol{X} \in t_L | \boldsymbol{X} \in t) = \frac{P(t_L)}{P(t)}, \tag{3.11}$$

$$P_R = P(\boldsymbol{X} \in t_R | \boldsymbol{X} \in t) = \frac{P(t_R)}{P(t)}, \tag{3.12}$$

kde P_L a P_R jsou pravděpodobnosti příslušení do množin t_L a t_R za podmínky příslušení do množiny t. Zřejmě platí

$$P_L + P_R = 1.$$

Označíme-li množinu koncových uzlů po dodatečném dělení s jako $\tilde{T}'$, potom

$$\tilde{T}' = (\tilde{T} \cup \{t_L, t_R\}) \setminus \{t\}. \tag{3.13}$$

Následně lze definovat pokles rizika jako

$$\Delta R(s, t) = R(\tilde{T}) - R(\tilde{T}') = R(t) - R(t_L) - R(t_R), \tag{3.14}$$

kdy poslední rovnost v (3.14) platí díky (3.13). Výraz (3.14) lze ještě upravit za použití vztahů (3.11) a (3.12) do tvaru

$$\Delta R(s, t) = P(t)[r(t) - r(t_L)P_L - r(t_R)P_R]. \tag{3.15}$$

Následně lze definovat relativní pokles rizika jako

$$\Delta R(s|t) = \frac{\Delta R(s, t)}{P(t)} = r(t) - r(t_L)P_L - r(t_R)P_R. \tag{3.16}$$

V uzlu t je tedy vždy zvoleno takové dělení s, které maximalizuje relativní pokles (3.16). Následující věta ukazuje, že relativní pokles je vždy nezáporný.

Věta 6. *Nechť t je uzel, který je dělen dělením s do dceřiných uzlů t_L a t_R. Potom je relativní pokles očekávané chyby vždy nezáporný, tedy*

$$\Delta R(s|t) \geq 0$$

s rovností právě tehdy, když

$$r(t_L) = E[L(Y, \omega(t))|\boldsymbol{X} \in t_L], \tag{3.17}$$

$$r(t_R) = E[L(Y, \omega(t))|\boldsymbol{X} \in t_R]. \tag{3.18}$$

Důkaz. Jelikož platí

$$P_L + P_R = 1,$$

je možné $r(t)$ rozepsat jako

$$\begin{aligned} r(t) =& (P_L + P_R)\ \mathrm{E}\left[L(Y, \omega(t))|\boldsymbol{X} \in t\right] = \\ =& (P_L + P_R)\ \left(\mathrm{E}\left[L(Y, \omega(t))|\boldsymbol{X} \in t_L\right] + \mathrm{E}\left[L(Y, \omega(t))|\boldsymbol{X} \in t_R\right]\right) = \\ =& \mathrm{E}\left[L(Y, \omega(t))|\boldsymbol{X} \in t_L\right]P_L + \mathrm{E}\left[L(Y, \omega(t))|\boldsymbol{X} \in t_R\right]P_R. \end{aligned}$$

Zároveň platí

$$\begin{aligned} &\Delta R(s|t) = r(t) - P_L r(t_L) - P_R t(t_R) = \\ &= \mathrm{E}\left[L(Y, \omega(t))|\boldsymbol{X} \in t_L\right]P_L + \mathrm{E}\left[L(Y, \omega(t))|\boldsymbol{X} \in t_R\right]P_R - r(t_L)P_L - r(t_R)P_R. \end{aligned} \tag{3.19}$$

Vidíme, že pravá strana rovnosti (3.19) je nulová právě tehdy, když platí (3.17) a (3.18). Tím je věta dokázána.

□

Na závěr si všimněme, že pro regresní problém lze relativní pokles rizika interpretovat jako relativní pokles rozptylu a platí

$$\Delta R(s|t) = \sigma^2(t) - P_L\sigma^2(t_L) - P_R\sigma^2(t_R).$$

Navíc platí, že střední hodnota odezvy v uzlu t je rovná váženému součtu středních hodnot v uzlech t_R a t_R, proto

$$\begin{aligned} \mu(t) =& (P_L + P_R)\ \mathrm{E}\left[Y|\boldsymbol{X} \in t\right] = \\ =& \mathrm{E}\left[Y|\boldsymbol{X} \in t_L\right]P_L + \mathrm{E}\left[Y|\boldsymbol{X} \in t_R\right]P_R = \\ =& \mu(t_L)P_L + \mu(t_R)P_R. \end{aligned}$$

3.2 Bayesovské metody

V prvních dvou kapitolách jsme se seznámili s algoritmem pro hledání klasifikačních a regresních stromů. Takový algoritmus vystavěl binární rozhodovací strom za použití určitého pravidla, které v každém uzlu našlo nejlepší dělení na základě predikčních proměnných a koncovým uzlům byla přidělena predikovaná třída v případě klasifikačního problému, nebo predikovaná hodnota v případě regresního problému. Uzly byly děleny tak dlouho, dokud nebyl sestaven strom T_{MAX}, který byl následně metodou ořezávání s minimální nákladovou složitostí ořezáván do té doby, dokud nezbyl jen počáteční uzel. Tato metoda vytvořila posloupnost ideálních stromů, že které byl podle upřímných odhadů a pravidla 1SE zvolen nejlepší strom. Problém je ovšem v tom, že při každém dělení uzlu bylo zvolené jedno nejlepší dělení, které maximalizovalo pokles nečistoty. Takové dělení se tedy snažilo v každém uzlu zmenšit komplexnost problému a rozdělit případy podle společných vlastností. Jenomže taková volba dělení nemusí často odhalit hlubší závislosti, které nejsou na první pohled zřejmé a proto poté nebudou zohledněny při volbě nejlepšího dělení. V této kapitole bude představena metoda podle práce Davida, Denisona, Mallicka a Smitha (Denison, D.G.T., Mallick, Smith a A.F.M., 1998), využívající apriorní a aposteriorní rozdělení stromu spolu s algoritmem hledání stromu, jejichž produktem je klasifikační či regresní strom.

3.2.1 Struktura modelu

Tato metoda využívá standardní sadu otázek, tak jak byla definována v první kapitole s tím rozdílem, že jsou uvažovány pouze dělení na numerických proměnných. Dělení jedné proměnné x má tedy tvar

$$\{\text{Je } x \leq c?\} \tag{3.20}$$

Každé dělení s_i, $i = 1, \ldots, s_{max}$, uzlu t_i bude mít tři parametry, a to h_i, v_i a r_i. V první kapitole bylo znázorněno, jak dělení uzlů probíhá. Uzel t_i, který je dělen dělením s_i, má levý a pravý dceřiný uzel. Do levého dceřiného uzlu t_{2i} jsou případy s kladnou odpovědí na dělící otázku a do pravého dceřiného uzlu t_{2i+1} případy se zápornou odpovědí. Toto indexování tedy jednoznačně určuje pozici uzlu, a tedy i pozici dělení. Když tedy položíme

$$h_i = i,$$

potom parametr h_i můžeme nazvat parametrem pozice a jednoznačně určuje pozici dělení s_i ve stromu.

Předpokládejme nyní vektor měření

$$\boldsymbol{x} = (x_1, \ldots, x_M)$$

dimenze M, kde každá proměnná x_i, $i = 1, \ldots, M$, je nominální. Jelikož v každém dělení s_i je použit vždy jen jeden prediktor x_m, $m \in \{1, \ldots, M\}$, potom položíme

$$v_i = m.$$

Parametr v_i se nazývá parametrem prediktoru a v každém dělení s_i jednoznačně určuje prediktor použitý k dělení uzlu.

Posledním parametrem dělení s_i je parametr hodnoty r_i. Ten je v případě otázky tvaru (3.20) definován jako

$$r_i = c.$$

Všimněme si, že trojice parametrů $\{h_i, v_i, r_i\}$ jednoznačně určuje pozici i tvar dělení s_i.

Definice 19 (třída modelů)**.** *Třída modelů $\mathcal{M}^k$ je množina všech stromů majících právě k koncových uzlů.*

Zároveň platí, že strom s k koncovými uzly má přesně $(k-1)$ nekoncových uzlů, což lze využít v následující větě.

Věta 7. *Nechť strom T náleží do modelu stromů $\mathcal{M}^k$. Potom je strom T jednoznačně definován vektorem*

$$\boldsymbol{\theta}^k = (h_1, v_1, r_1, \ldots, h_{k-1}, v_{k-1}, r_{k-1}),\ \boldsymbol{\theta}^k \in \Theta^k,$$

až na hodnoty přiřazené v koncových uzlech.

Důkaz. Důkaz plyne z definic jednotlivých parametr a faktu, že každý strom je jednoznačně definovaný svými děleními a hodnotami přiřazenými koncovým uzlům.

□

Θ^k je podprostor prostoru $\mathbb{R}^{3(k-1)}$, jelikož každý parametr dělení nabývá reálné hodnoty na prostoru $\mathbb{R}$.

Nyní předpokládejme, že opravdový model má k koncových uzlů, je neznámý, ale pochází z třídy $\mathcal{M}^k$. Označíme-li vektor dat jako y, potom je možné zkonstruovat apriorní sdružené rozdělení pravděpodobnosti

$$p(k,\, \theta^k,\, y).$$

To lze upravit jako

$$p(k,\, \theta^k,\, y) = p(y|\theta^k,\, k)p(\theta^k,\, k) = p(y|\theta^k,\, k)p(\theta^k|k)p(k),$$

kde $p(y|\theta^k,\, k)$ je věrohodnostní funkce, $p(\theta^k|k)$ je apriorní rozdělení parametrů za podmínky, že strom má k koncových uzlů, a $p(k)$ je pravděpodobnost třídy $\mathcal{M}^k$. Z toho plyne, že všechny stromy s přesně k koncovými uzly jsou z jedné třídy modelů, a proto mají stejnou pravděpodobnost $p(k)$. Nyní se zaměříme na tvar věrohodnostní funkce $p(y|\theta^k,\, k)$.

Nejdříve označme počet případů v koncovém uzlu t_i jako n_i. V regresním problému v každém koncovém uzlu t_i předpokládáme, že proměnná y má n_i-rozměrné normální rozdělení s neznámou střední hodnotou μ_i a neznámým rozptylem σ_i^2. Nechť výraz $|t_i|$ označuje počet případů v koncovém uzlu t_i. Potom volíme konstantu t_{min} jako minimální počet případů v každém koncovém uzlu. Následně má věrohodnostní funkce $p(y|\theta^k,\, k)$ tvar

$$p(y|k,\, \theta^k) \propto \prod_{i=1}^{k} \left(\mathbb{1}(n_i \geq t_{min}) \frac{1}{\sigma_i} \exp\left[-\frac{1}{2\sigma_i^2} \sum_{y_j \in t_i} (y_j - \mu_i)^2 \right] \right). \tag{3.21}$$

Indikátor v (3.21) nám zaručí, že v každém koncovém uzlu bude alespoň t_{min} případů. Pro střední hodnoty $\{\mu_1, \ldots,\, \mu_k\}$ předpokládáme apriorní limitní rovnoměrné rozdělení a pro rozptyly σ_i^2, $(i = 1, \ldots,\, k)$, předpokládáme apriorní Gama rozdělení $\Gamma(10^{-2}{,}10^{-2})$. Střední hodnoty a rozptyly $\{\mu_1,\, \sigma_1^2, \ldots, \mu_k,\, \sigma_k^2\}$ jsou parametry modelu a proto je o ně rozšířen vektor parametrů θ.

Co se týče klasifikačního problému, zde předpokládáme multinomické rozdělení vysvětlované proměnné y. Označíme-li $n_{i,j}$ jako počet případů v koncovém uzlu t_i, jejichž třída je j, $j \in \{1, \ldots,\, J\}$, a příslušnou pravděpodobnost tohoto jevu jako $p_{i,j}$, potom má věrohodnostní funkce $p(y|\theta^k,\, k)$ tvar

$$p(y|k,\, \theta^k) \propto \prod_{i=1}^{k} \left(\mathbb{1}(n_i \geq t_{min}) \prod_{j=1}^{J} (p_{i,j})^{n_{i,j}} \right).$$

Pro jednotlivé pravděpodobnosti $\{p_{i,1}, \ldots, p_{i,J}\}$ v rámci koncového uzlu t_i předpokládáme apriorní Dirichletovo rozdělení $Dir(p_{i,1}, \ldots, p_{i,J}|1, \ldots, 1)$ a pro vektory pravděpodobností $(p_{i,1}, \ldots, p_{i,J})$, $i = 1, \ldots,\, k$, mezi jednotlivými koncovými uzly předpokládáme rovnoměrné rozdělení.

Možnou modifikací regresního problému je předpoklad shodného rozptylu σ^2 pro všechny koncové uzly. Potom má věrohodnostní funkce $p(y|\theta^k,\, k)$ tvar

$$p(y|k,\, \theta^k) \propto \prod_{i=1}^{k} \left(\mathbb{1}(n_i \geq t_{min}) \frac{1}{\sigma} \exp \left[-\frac{1}{2\sigma^2} \sum_{y_j \in t_i} (y_j - \mu_i)^2 \right] \right).$$

Je možné změnit i apriorní pravděpodobnostní rozdělení pro střední hodnoty μ_i,$i = 1, \ldots, k$. Můžeme předpokládat, že takové střední hodnoty pocházejí z normálního rozdělení. Pro případ s rozdílnými rozptyly v každém koncovém uzlu je možné pro tyto rozptyly předpokládat apriorní inverzní Gaussovo rozdělení IG($\nu/2$,$\nu\lambda/2$). Více o tomto přístupu lze najít v článku H. Chipmana, E. George a R. McCullocha (Chipman a kol., 1998).

Pro apriorní rozdělení parametrů $p(\theta^k|k)$ předpokládáme rovnoměrné rozdělení na množině možných hodnot parametru prediktoru v_i, $i = 1, \ldots, J$, tedy na množině všech proměnných $\{x_1, \ldots, x_M\}$. Zároveň předpokládáme rovnoměrné rozdělení na množině hodnot parametru hodnoty r_i. To znamená, že každá proměnná má stejnou pravděpodobnost, že bude použita pro dělení daného uzlu, a každá hodnota takové proměnné má stejnou pravděpodobnost, že bude dělit daný uzel.

Co se týče apriorní pravděpodobnosti modelu $p(k)$, zde předpokládáme Poissonovo rozdělení s parametrem λ. Připomeňme, že p(k) je pravděpodobnost množství koncových uzlů a má tvar

$$p(k) = \frac{\lambda^k}{(e^\lambda - 1)k!},$$

kde $k = 1,2,3, \ldots$

Následný výpočet založený na reálných datech je založen na aposteriorním rozdělení

$$p(k,\, \lambda^k|y) = p(k|y)p(\lambda^k|k,\, y).$$

Hledací algoritmus poté začne pouze s počátečním uzlem a v každém kroku náhodně provede jeden z následujících operací:

- Je náhodně zvolen jeden z koncových uzlů, který je náhodně rozdělen.
- Je náhodně zvolen jeden z koncových uzlů, který je ořezán.
- Je náhodně zvolen jeden z uzlů, pro který je náhodně změněn prediktor pro dělení.
- Je náhodně zvolen jeden z uzlů, pro který je náhodně změněna hodnota prediktoru použitého pro dělení.

Pravděpodobnosti jednotlivých operací se vhodně mění podle počtu koncových uzlů, aby se algoritmus nezasekl například pouze v počátečním uzlu, nebo aby nebyl konstruován moc velký strom. Podrobnosti o tomto algoritmu lze nalézt v článku Denisona (Denison a kol., 1998, str. 368, 369).

4. Praktický příklad

V této kapitole bude představený praktický příklad algoritmu CART pro klasifikační strom. K tomu využijeme knihovnu *rpart*, která implementuje algoritmus CART do prostředí statistického programu *R*.

Jako trénovací výběr $\mathcal{L}$ použijeme dataset $PimaIndiansDiabetes2$ z knihovny *mlbench*. Tato data jsou výsledkem měření Amerického Národního institutu diabetu a onemocnění trávicího traktu a ledvin. Měření probíhalo na populaci indiánek starších 21 let z kmene Pima v Arizoně.

```
> dim(PimaIndiansDiabetes2)
[1] 768   9
> head(PimaIndiansDiabetes2)
  pregnant glucose pressure triceps insulin mass pedigree age diabetes
1        6     148       72      35      NA 33.6    0.627  50      pos
2        1      85       66      29      NA 26.6    0.351  31      neg
3        8     183       64      NA      NA 23.3    0.672  32      pos
4        1      89       66      23      94 28.1    0.167  21      neg
5        0     137       40      35     168 43.1    2.288  33      pos
6        5     116       74      NA      NA 25.6    0.201  30      neg
```

Celkově dataset obsahuje 768 měření celkově devíti proměnných. Osm numerických proměnných:

- *pregnant*... počet těhotenství za svůj život,
- *glucose*... koncentrace glukózy v plazmě (glukózový toleranční test),
- *pressure*... diastolický krevní tlak ($mm\ Hg$),
- *triceps*... tloušťka kožní řasy nad tricepsem,
- *insulin*... sérová koncentrace inzulinu(muU/ml),
- *mass*... index tělesné hmotnosti (BMI),
- *pedigree*... funkce rodokmenového diabetu,
- *age*... věk v letech,

a kategoriální proměnnou *diabetes*, která v tomto problému predikovanou proměnnou. Tato proměnná nabývá hodnot *pos*, tedy pozitivní test na diabetes, a *neg*, tedy negativní test na diabetes. Cílem zkonstruovaného klasifikačního stromu tedy bude v budoucnu na základě měření numerických proměnných rozhodnout, zda daná osoba má diabetes, či nikoliv.

```
> con<-rpart.control(minsplit = 5,cp=0,xval=10,maxcompete = 1,
maxsurrogate = 3)
> T_max<-rpart(diabetes~pregnant+glucose+pressure+triceps+insulin+mass+
pedigree+age,data=PimaIndiansDiabetes2,method = "class", control = con)
```

Nejdříve potřebujeme zkonstruovat strom T_{MAX}. Funkce *rpart*() má mnoho modifikací, a proto jsme pomocí *rpart.control*() určili některá pravidla. Argument *minsplit* specifikuje, kolik nejméně musí být v uzlu případů, aby byl dělen. Argument *cp* nastavuje počáteční hodnotu α jako parametru složitosti. Následně argument *xval* určuje počet množin V pro metodu odhadu křížovou validací, zde jsme zvolili $V = 10$. Argumenty *maxcompete* a *maxsurrogate* určují, kolik dalších nejlepších dělení a kolik náhradních dělení v každém uzlu má být hledáno. Strom T_{MAX} je tedy velký strom, který je nyní potřeba metodou ořezávání s minimální nákladovou složitostí ořezávat a najít posloupnost ideálních stromů

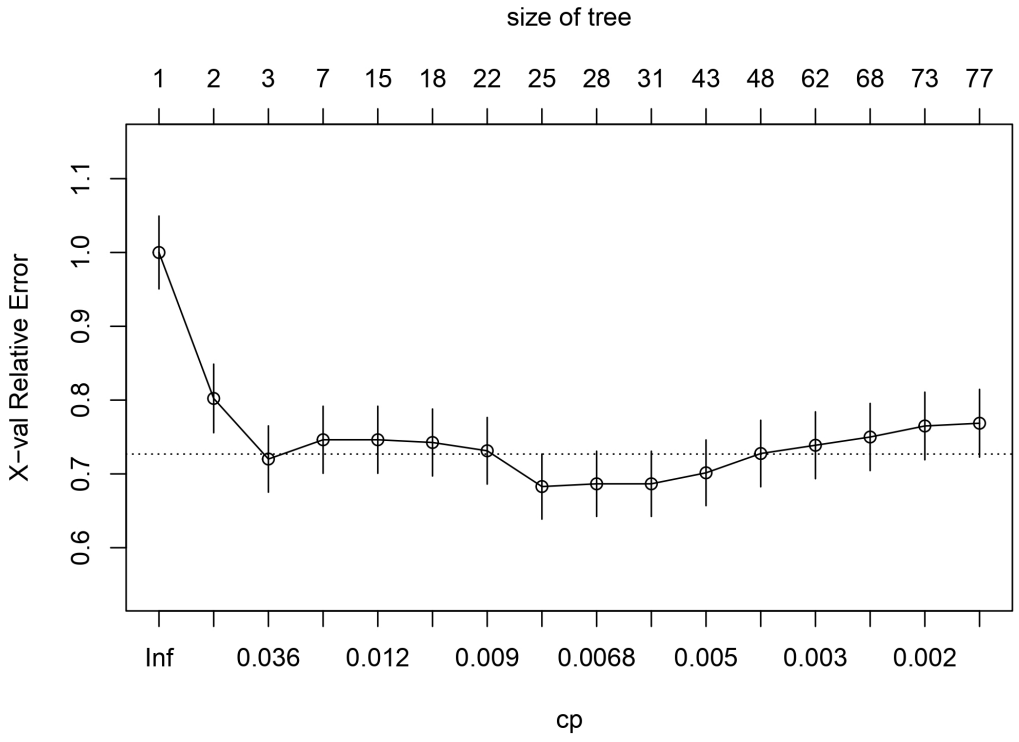
$$T_{MAX} \succ T_1 \succ T_2 \succ \ldots \succ \{t_1\}.$$

Tuto posloupnost je možné vypsat použitím funkce *printcp*().

```
> printcp(T_max)

Classification tree:
rpart(formula = diabetes ~ pregnant + glucose + pressure + triceps +
    insulin + mass + pedigree + age, data = PimaIndiansDiabetes2,
    method = "class", control = con)

Variables actually used in tree construction:
[1] age  glucose  insulin  mass  pedigree  pregnant  pressure  triceps

Root node error: 268/768 = 0.34896

n= 768

           CP nsplit rel error  xerror     xstd
1  0.2425373      0   1.00000 1.00000 0.049288
2  0.1007463      1   0.75746 0.80224 0.046427
3  0.0130597      2   0.65672 0.72015 0.044854
4  0.0118159      6   0.60448 0.74627 0.045381
5  0.0111940     14   0.50373 0.74627 0.045381
6  0.0093284     17   0.47015 0.74254 0.045307
7  0.0087065     21   0.43284 0.73134 0.045083
8  0.0074627     24   0.40672 0.68284 0.044054
9  0.0062189     27   0.38433 0.68657 0.044137
10 0.0055970     30   0.36567 0.68657 0.044137
11 0.0044776     42   0.29478 0.70149 0.044461
12 0.0037313     47   0.27239 0.72761 0.045007
13 0.0024876     61   0.22015 0.73881 0.045233
14 0.0022388     67   0.20522 0.75000 0.045454
15 0.0018657     72   0.19403 0.76493 0.045742
16 0.0000000     76   0.18657 0.76866 0.045813
```

Hodnota *nsplit* vyjadřuje, kolik má daný strom dělení. Vidíme, že pro hodnotu $\alpha = 0.2425373$ již metoda ořezávání s minimální nákladovou složitostí preferuje pouze počáteční uzel t_1. Sloupec *rel error* vyjadřuje hodnotu resubstitučního odhadu vzhledem ke klasifikační chybě kořene (počátečního uzlu), proto se tato hodnota postupně zmenšuje s přibývajícími dělenými uzly. Pro volbu nejlepšího stromu nás ovšem budou zajímat sloupec *xerror*, který vyjadřuje hodnotu odhadu křížovou validací vzhledem ke klasifikační chybě kořene, a *xstd*, tedy standardní chyba takového odhadu. Pomocí funkce *plotcp*() dostaneme graf této tabulky.

```
> plotcp(T_max)
```



Obrázek 4.1: Chyba posloupnosti ideálních stromů dané velikosti

Na výstupu vznikl graf, který můžeme vidět na obrázku 4.1. Přerušovanou čárou je znázorněno 1SE pravidlo, podle kterého zvolíme jako nejlepší strom ten, který má 3 koncové uzly. To odpovídá stromu s dvěma děleními, který je preferovaný pro hodnotu α z intervalu

$$[0{,}0130597;\ 0{,}1007463). \tag{4.1}$$

Stromy se 7, 15, 18 a 22 koncovými uzly mají vyšší hodnotu odhadu metodou křížové validace než náš zvolený strom se třemi koncovými uzly. Proto by jejich volba nedávala žádný smysl. Co se týče uzlů s 25, 28, 31 a 43 koncovými uzly, ty jsou přeci jen přesnější, ale za cenu velkého počtu koncových uzlů. Příliš mnoho koncových uzlů se špatně interpretuje, a proto dáváme vždy přednost menším stromům.

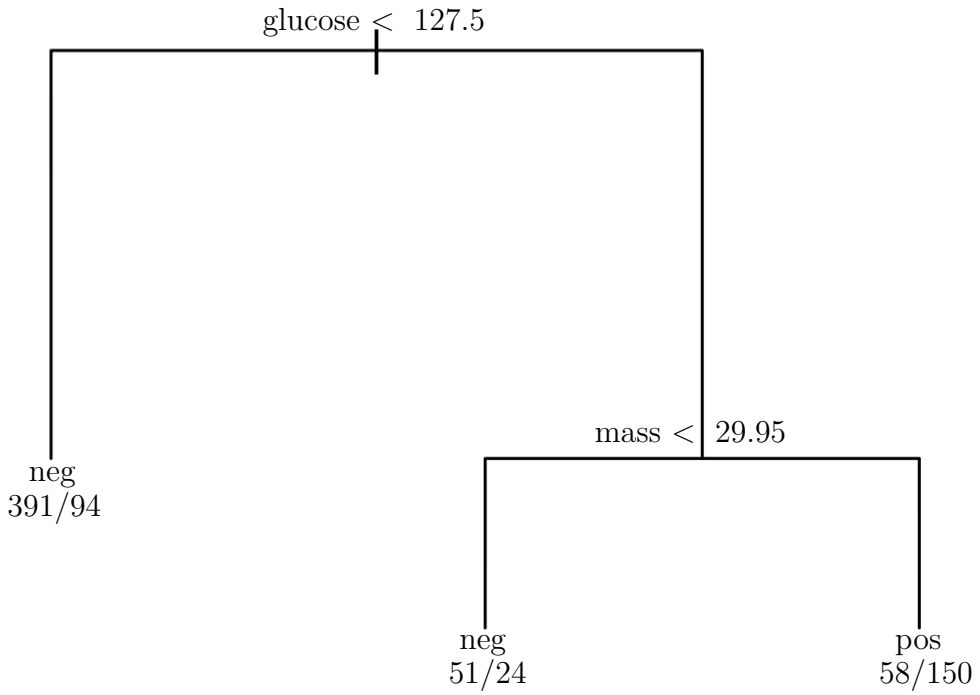
Nyní tedy můžeme ořezat strom T_{MAX} pomocí funkce *prune*(), kdy za parametr *cp* volíme hodnotu α z intervalu (4.1).

```
> T_id<-prune(T_max,cp=0.02)
> plot(T_id)
> text(T_id,use.n = TRUE)
```

Na obrázku 4.2 je znázorněný nejlepší strom, který má tři koncové uzly. V počátečním uzlu bylo vybráno nejlepší pravidlo ve tvaru

$$\{\text{Je } glucose\ < 127{,}5?\}. \tag{4.2}$$

Případy, pro které byla odpověď na tuto otázku kladná, šly do levého dceřiného uzlu, který je zároveň koncový. Tomuto uzlu byla přidělena třída *neg*, což odpovídá negativnímu testu na diabetes. To lze interpretovat tak, že pokud má indiánka z kmene Pima starší 21 let koncentraci glukózy v plazmě menší než 127,5 v tolerančním glukózovém testu, potom je klasifikována tak, že diabetes nemá. Všimněme si ovšem, že v tomto koncovém uzlu skončilo také 94 případů, které měly pozitivní test na diabetes. Tyto případy byly tedy špatně klasifikovány, což se poté projeví na odhadech přesnosti daného stromu.



Obrázek 4.2: Nejlepší klasifikační strom

Případy, pro které byla odpověď na otázku (4.2) záporná, šly do levého dceřiného uzlu, který byl dále dělen otázkou

$$\{\text{Je } mass \ < 29{,}95?\}.$$

Kladná odpověď poslala případy do levého dceřiného uzlu a záporná do pravého. Proto lze vidět, že ženy, které mají výsledek v testu na glukózu v krvi vyšší než 127,5 a zároveň mají index tělesné hmotnosti vyšší než 29,95, jsou klasifikovány na přítomnost diabetu pozitivně. Interpretace je tedy taková, že žena z kmene Pima, která má vysokou hladinu cukru v krvi a zároveň je obézní, je klasifikována pozitivně. Nesmíme ovšem zapomenout i na chybně klasifikované případy v tomto koncovém uzlu.

Celková chybná klasifikace stromu na daném trénovacím výběru je tedy

$$\frac{94 + 24 + 58}{768} \cong 0{,}22917.$$

Po vydělení tohoto čísla klasifikační chybou kořene máme

$$\frac{0{,}22917}{0.34896} \cong 0{,}6567,$$

což odpovídá resubstitučnímu odhadu tohoto stromu, který byl uveden v tabulce ve výstupu z programu *R*.

Závěr

Cílem této práce bylo představit čtenáři klasifikační a regresní stromy, seznámit ho s konstrukcí a modifikacemi těchto modelů. V práci byla předvedena konstrukce stromu, volbu dělení v každém nekoncovém uzlu a následné přiřazení vhodné hodnoty predikované proměnné každému koncovému uzlu. Dále byly diskutovány i odlišné přístupy k tomuto problému a na závěr byl představen praktický příklad, jehož účelem bylo čtenáři představit předem získané znalosti v praxi.

Seznam použité literatury

Breiman, L. (1993). *Classification and regression trees*. Chapman & Hall. ISBN 0412048418.

Chipman, H., George, E. a McCulloch, R. (1998). Bayesian cart model search. *J. Amer. Statist. Assoc.*, **93**, 935–960.

Denison, D.G.T., Mallick, B., Smith a A.F.M. (1998). A bayesian cart algorithm. *Biometrika*, **85**, 363–377.

Hershy, A. (2019). Gini index vs information entropy. *Towards Data Science*. URL `https://towardsdatascience.com/gini-index-vs-information-entropy-7a7e4fed3fcb`.

James, G., Witten, D., Hastie, T. a Tibshirani, R. (2021). *Introduction to Statistical Learning With Applications in R*. Springer. ISBN 9781071614174.

Scott, C. (2005). Tree pruning with subadditive penalties. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, **53**, 4518–4525.