Změkčování hranic v klasifikačních stromech

Jakub Dvořák

Seminář strojového učení a modelování

24.5.2012

Obsah

Obsah

Klasifikační stromy

Jedna z metod strojového učení.
Z trénovacích dat generována struktura zakořeněného stromu:

- Ve vnitřních uzlech rozhodovací podmínka.
- V listech **score** pro jednotlivé třídy.

Nejznámější metody: CART, C4.5, C5.0.

Nevýhoda

Výstup nespojitý — po částech konstantní. To často neodpovídá skutečnosti.

Generování stromů

1. Růst stromu
2. Prořezávání

Klasifikační strom

Rozhodovací podmínka

Typicky ve vnitřním uzlu testuje hodnotu jednoho atributu předloženého vzoru.
U kontinuálních atributů test mívá podobu $x_A \leq c$, kde

- x_A označuje hodnotu atributu A,
- c je split-hodnota.

Score v listech

Pro K tříd, N vzorů v daném listu, z nichž N_i je třídy i:

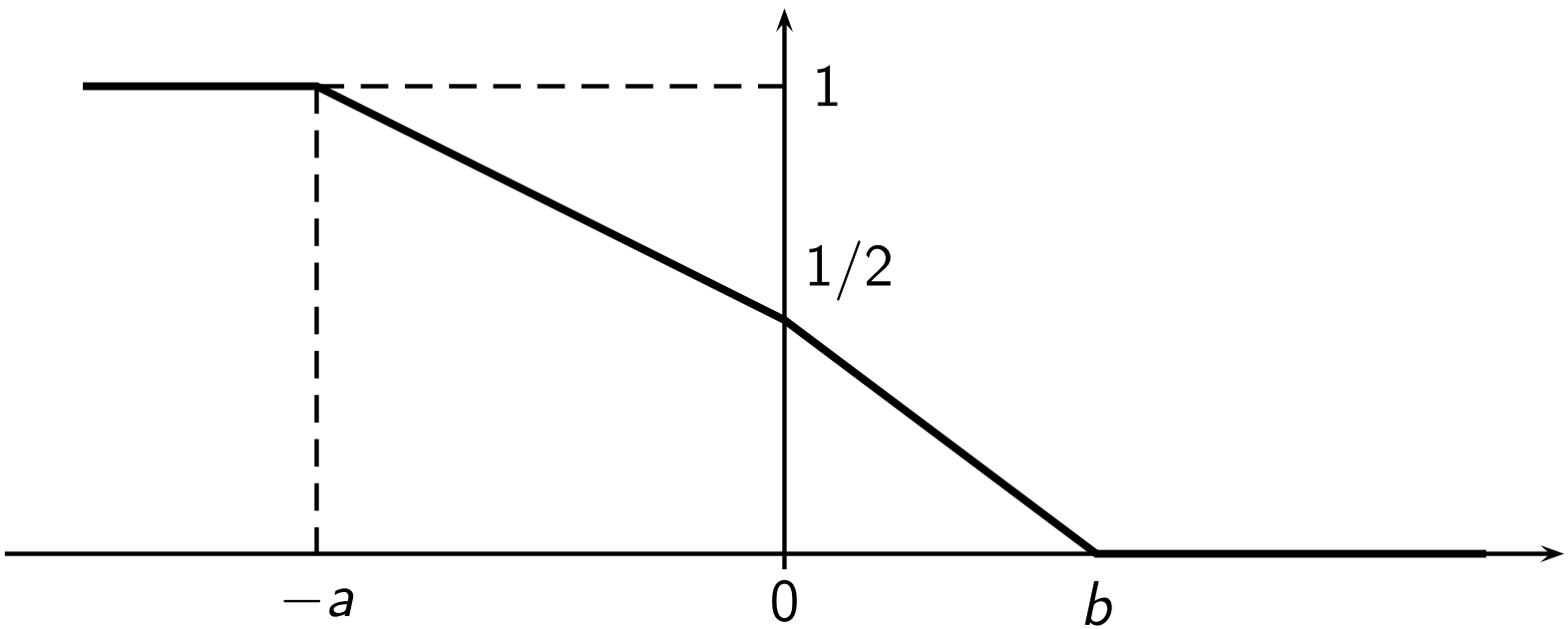
- Relativní frekvence třídy: N_i/N
- Laplaceova korekce: $(N_i + 1)/(N + K)$
- Jiné možnosti (m-smoothing).

Obsah

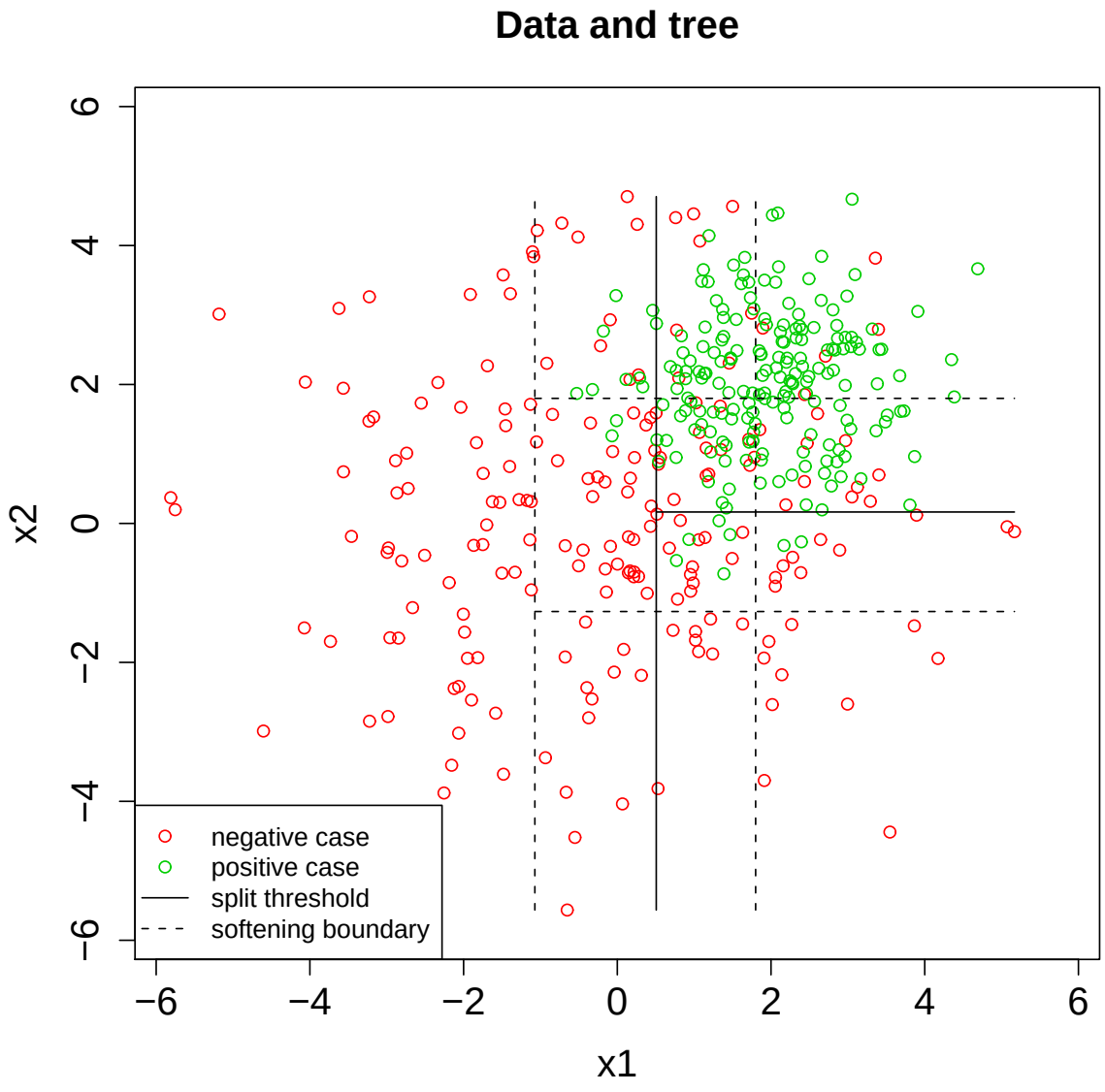
Změkčování hran

V uzlech, kde je rozhodovací podmínka $x_A \leq c$.
Je-li hodnota x_A blízko c, potom pokračovat do obou větví.
Výsledky z podstromů zkombinovat váženým průměrem.
Váhy závisí na $x_A - c$.

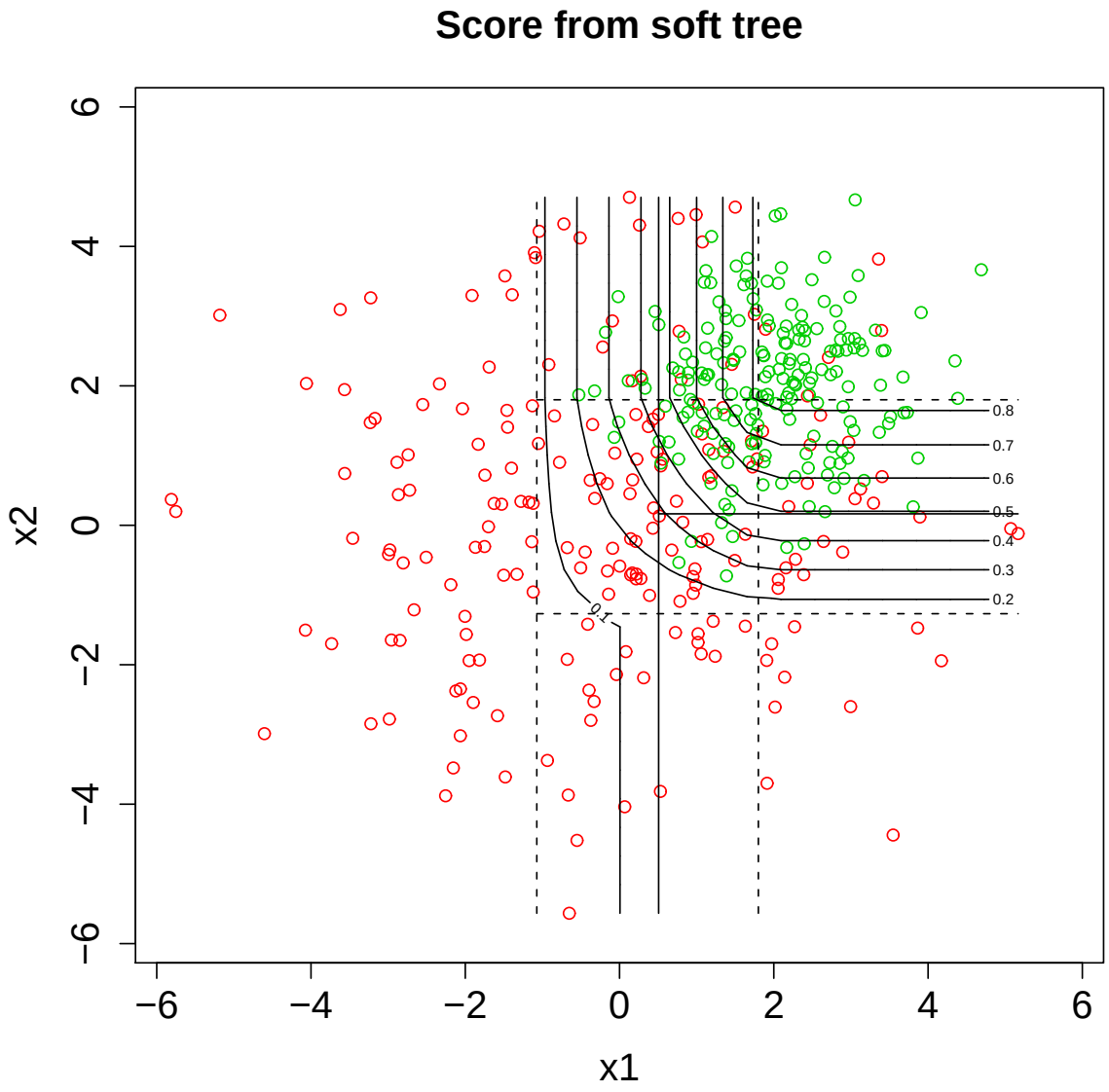
Změkčující křivka



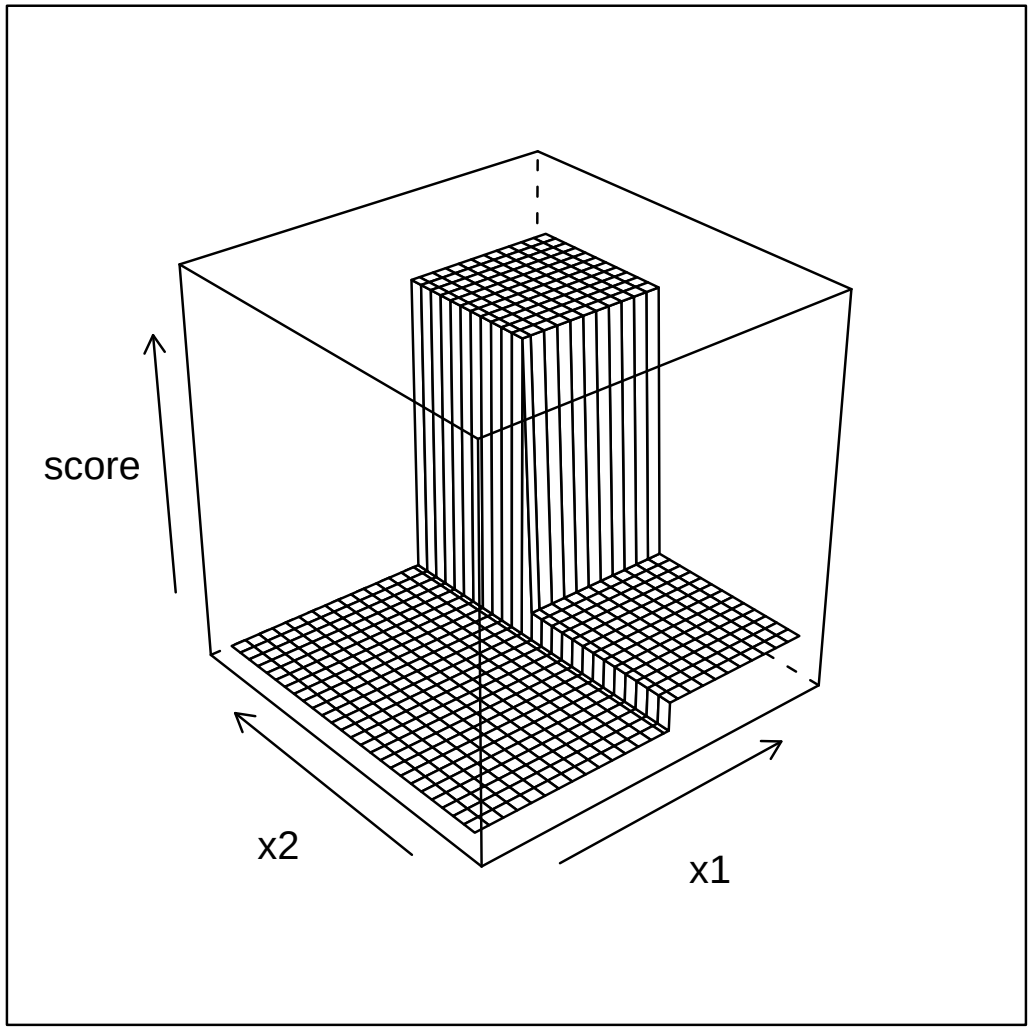
Příklad


Data and tree
x2
x1
6
4
2
0
−2
−4
−6
−6
−4
−2
0
2
4
6
negative case
positive case
split threshold
softening boundary

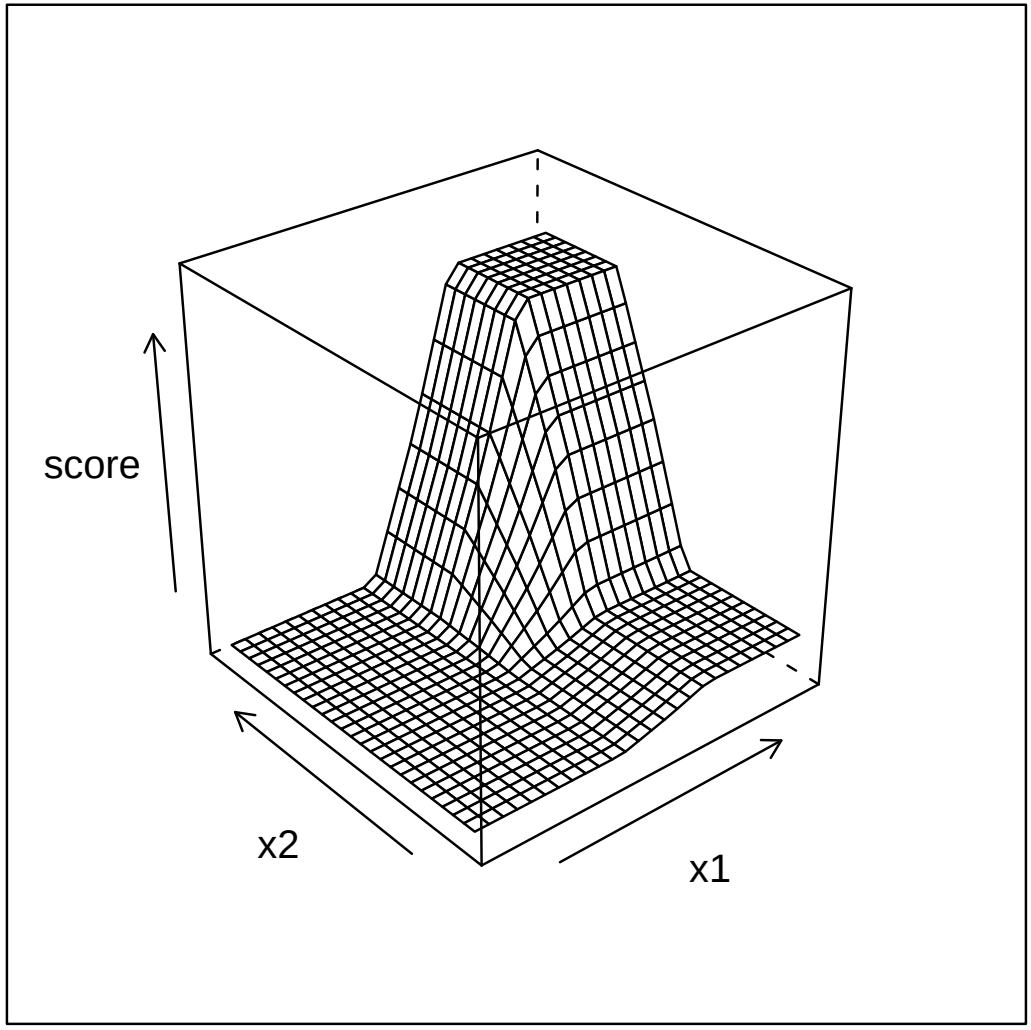
Příklad



Příklad



Příklad



Obsah

Ranking

Pro klasifikaci do dvou tříd (pozitivní, negativní).
Seřazení množiny vzorů od nejspíše negativních k nejspíše pozitivním.
Obvykle uspořádání hodnot score.
Není nutné, aby score bylo dobrým odhadem pravděpodobnosti — je to tedy lehčí problém.

ROC křivka

ROC "Receiver Operating Characteristic".
Mějme klasifikátor do dvou tříd poskytující score pro každý vzor.
Lze libovolně zvolit práh a vzory se score nad tímto prahem klasifikovat jako pozitivní, ostatní jako negativní.
Pro množinu vzorů, u nichž známe skutečnou třídu, lze pro každou hodnotu prahu určit

$$\text{True positive rate} = \frac{\#\text{ správně klasifikovaných pozitivních vzorů}}{\#\text{ skutečně pozitivních vzorů}}$$

$$\text{False positive rate} = \frac{\#\text{ chybně klasifikovaných negativních vzorů}}{\#\text{ skutečně negativních vzorů}}$$

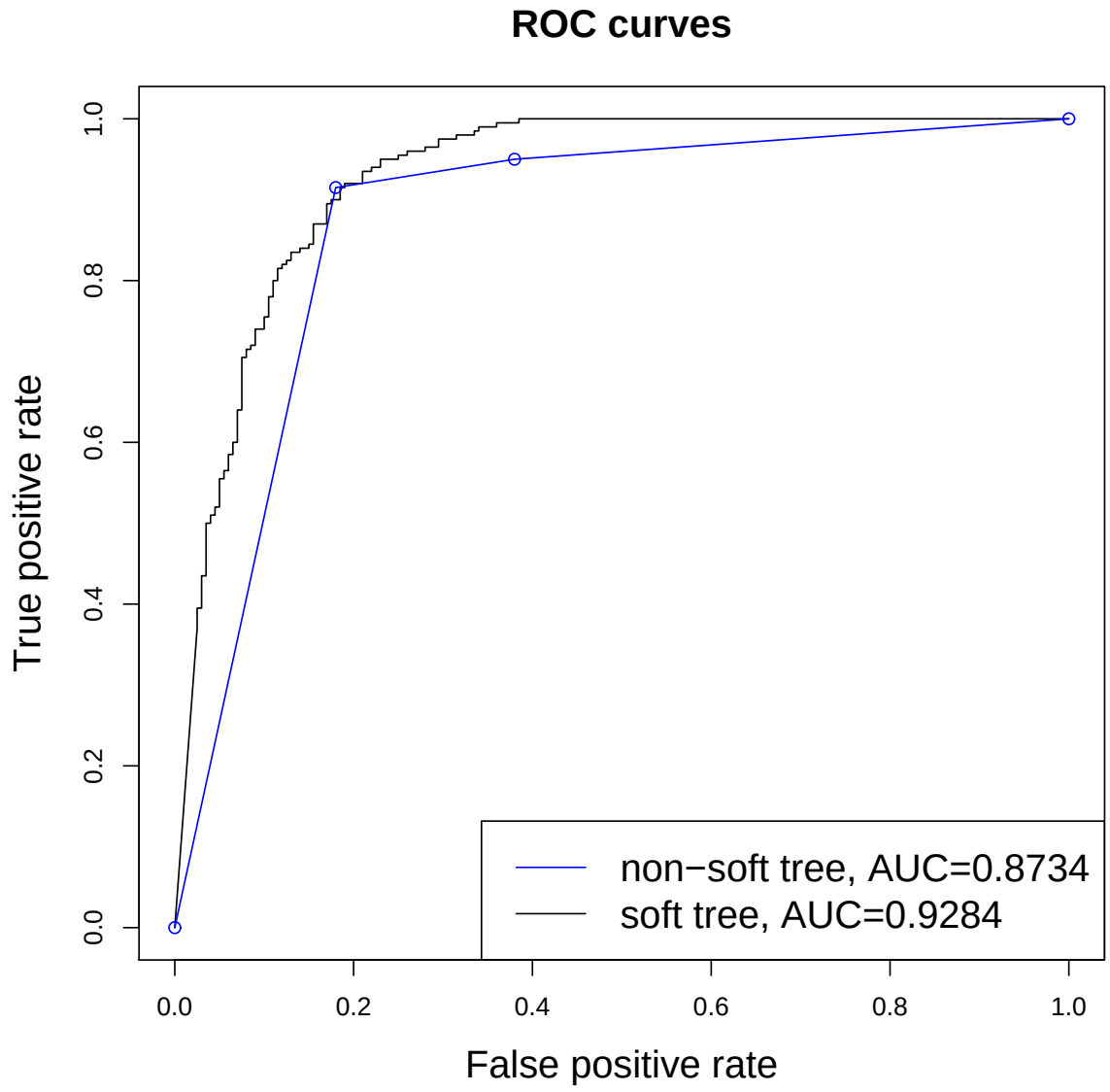
ROC křivka graficky zachycuje vztah True positive rate a False positive rate pro různé hodnoty prahu.

AUC

AUC "Area Under (ROC) Curve" — plocha pod ROC křivkou.
Jedno číslo vyjadřující kvalitu klasifikátoru poskytujícího score pro dvě třídy.
Hodnota od 0 do 1, čím větší, tím lepší.

Příklad

ROC křivky k příkladu jednoduchého změkčeného stromu:



Obsah

Metody změkčování

Určení parametrů a, b změkčující křivky.
Parametry se mohou v různých vnitřních uzlech lišit, tedy každému vnitřnímu uzlu s podmínkou na kontinuální atribut určujeme jeho vlastní parametry.
Postprocessing — po dokončení konstrukce celého stromu:

1. Růst stromu
2. Prořezávání
3. Změkčování

Dva typy metod:

- Bez použití optimalizace
- Pomocí optimalizace Nelder-Mead

Metody změkčování

Změkčování bez optimalizace

Parametry změkčující křivky se určují na základě trénovacích vzorů, které padnou před změkčením do příslušného uzlu stromu.
V jednotlivých uzlech jsou určovány nezávisle.
Při změkčování není brán ohled na to, jaký je vliv změkčení na výstup klasifikátoru.

Změkčování pomocí optimalizace

Parametry všech změkčujících křivek ve stromě se hledají v rámci jednoho optimalizačního procesu.
Cílová funkce je vypočtena na základě výstupu změkčeného klasifikátoru na trénovací množině.

Změkčování bez optimalizace

Nechť ve vnitřním uzlu v_j je rozhodovací podmínka $x_{A_j} \leq c_j$.
Označme C_j množinu trénovacích vzorů, které padnou do uzlu v_j.
Definujme:

- $l_j = \min_{\mathbf{x} \in C_j} x_{A_j}$
- $u_j = \max_{\mathbf{x} \in C_j} x_{A_j}$

Potom pro zvolené q nastavme parametry změkčující křivky:

- $a_j = 2^{-q}\,(c_j - l_j)$
- $b_j = 2^{-q}\,(u_j - c_j)$.

Metodu, která takto nastaví parametry ve všech změkčovaných uzlech stromu, označujeme DR(q).
Používáme $q = 0, 1, 2, 3, 4$.

Změkčování bez optimalizace

Do kategorie změkčování bez optimalizace patří tak změkčování ("probabilistic splits") v C4.5 a C5.0.

Změkčování v C4.5

Odhad standardní odchylky chyby

$$\tilde{\sigma} = \sqrt{\frac{(e + 1/2)(n - e - 1/2)}{n}}$$

kde n je $|C_j|$, e je počet vzorů z C_j, které jsou klasifikovány chybně při použití prahu $1/2$.
Potom hranice změkčení jsou nastaveny tak, že kdyby c_j bylo posunuto na tuto hranici, potom by se počet chybně klasifikovaných vzorů z C_j zvýšil o $\tilde{\sigma}$.

Změkčování v C5.0 se liší od C4.5, ale nebylo (pokud je mi známo) publikováno vysvětlení, jak je provedeno.

Optimalizace parametrů změkčení

Když s je počet změkčovaných uzlů ve stromu, potom počet parametrů změkčení je $2s$.
Nechť $v_1, \ldots, v_s$ jsou všechny změkčované uzly ve stromu, potom vektor parametrů uspořádáme: $a_1, \ldots, a_s, b_1, \ldots, b_s$.
Optimalizace probíhá v prostoru $\mathbb{R}^{2s}$, každý parametr změkčení je získán transformací jednoho optimalizovaného parametru.
Je použit škálovací vektor $\mathbf{z} = (z_1, \ldots z_{2s})$, který je určen jako výsledek změkčení DR(1).
Pro vektor $(p_1, \ldots, p_{2s})$ z prostoru parametrů optimalizace definujeme odpovídající parametry změkčení:

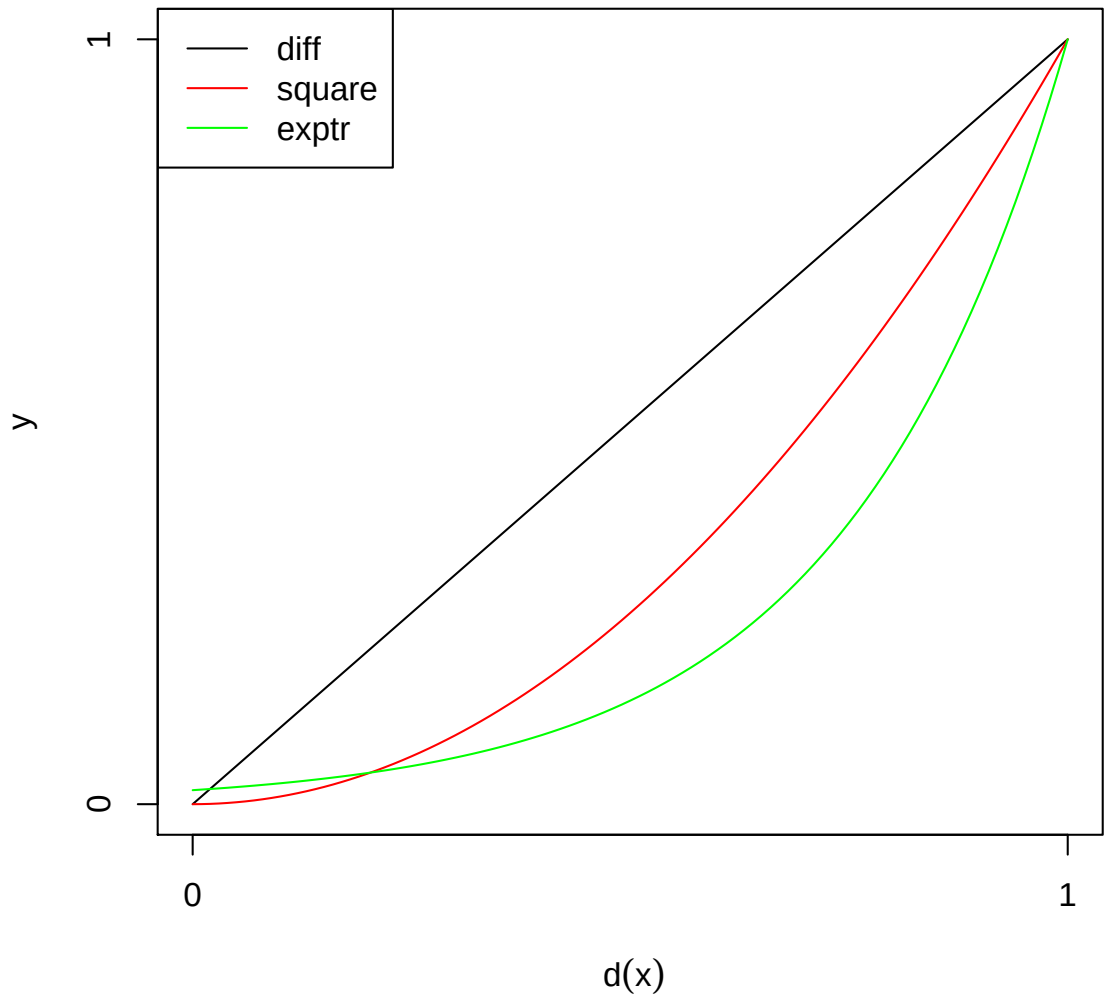
$$\left.\begin{array}{rcl} a_j & = & z_j p_j^2 \\ b_j & = & z_{s+j} p_{s+j}^2 \end{array}\right\} \text{pro } j = 1, \ldots, s$$

Cílové funkce pro optimalizaci

Trénovací množinu označme C.
Označme $r(\mathbf{x})$ score, které klasifikátor vrací, je-li předložen vzor $\mathbf{x}$.
Definujme $\tilde{r}(\mathbf{x}) = \begin{cases} 1 & \text{když } \mathbf{x} \text{ je pozitivní} \\ 0 & \text{když } \mathbf{x} \text{ je neagativní} \end{cases}$
Nechť $d(\mathbf{x})$ označuje $|r(\mathbf{x}) - \tilde{r}(\mathbf{x})|$

diff	$\frac{1}{\lvert C\rvert} \sum_{\mathbf{x}\in C} d(\mathbf{x})$
square	$\frac{1}{\lvert C\rvert} \sum_{\mathbf{x}\in C} d(\mathbf{x})^2$
exptr	$\frac{1}{\lvert C\rvert} \sum_{\mathbf{x}\in C} \exp\left(4\left(d(\mathbf{x}) - 1\right)\right)$
AUC	AUC vypočtená na trénovací množině.

Cílové funkce pro optimalizaci



Optimalizační metoda

Iterační optimalizace: Simplexový algoritmus pro nelineární minimalizaci — Nelder-Mead.
Během iterací se automaticky adaptuje na lokální tvar cílové funkce.
Tato metoda nepoužívá gradient, pouze funkční hodnoty cílové funkce.
Metoda s touto vlastností byla zvolena proto, že AUC je po částech konstantní, takže gradient nelze použít.
Ukončení iteračního procesu: dosažení počtu iterací $200s$.
Iniciální hodnota pro iterační optimalizaci: $(1, \ldots, 1)$, což odpovídá změkčení pomocí DR(1).

Obsah

Experiment

4 rozdělení dat pocházející z UCI Machine Learning repository.
Z každého rozdělení vygenerováno 10 trénovacích množin a testovací množina.
Z každé trénovací množiny vygenerováno aspoň 10 stromů různých velikostí metodami CART a C5.0.
Na stromech použity metody změkčování:

- DR(q), $q = 0, 1, 2, 3, 4$
- C4.5
- C5.0 — pouze stromy generované C5.0 metodou
- optimalizace diff
- optimalizace square
- optimalizace exptr
- optimalizace AUC

Kvalita výsledných klasifikátorů porovnávána podle AUC vypočtené na základě testovací množiny.

Datové množiny

Magic Telescope

Pocházejí z počítačové simulace fyzikálního procesu.
Popisují pozorování částic pomocí atmosférického Cherenkovova teleskopu.
Klasifikace rozlišuje pozorované záření způsobené gama paprsky (pozitivní) od záření vyvolaného částicemi v horních vrstvách atmosféry (negativní).
Asi 65 % pozitivních případů.
10 numerických atributů.
K dispozici je 19020 vzorů, rozděleny 10-krát nezávisle na 12680 trénovacích a 6340 testovacích.

Datové množiny

Waveform

Uměle generovaná data.
3 třídy stejné pravděpodobnosti.
21 numerických atributů.
Používáme úlohy rozlišení jedné třídy od sjednocení ostatních.
Ze tří takto vzniklých rozdělení jsou dvě ekvivalentní až na uspořádání parametrů, proto používáme pouze dvě (označujeme Waveform A, Waveform B).
Vygenerováno 10 trénovacích množin velikosti 12000 a jedna testovací množina velikosti 12000.

Datové množiny

MiniBooNE Particle Identification

Sjednocení skutečných pozitivních případů s Monte-Carlo generovanými negativními případy.
Klasifikace rozlišuje elektronová neutrina (pozitivní případy) od mionových neutrin (negativní případy).
V množině je asi 28 % pozitivních případů.
50 reálných atributů.
K dispozici je 129596 případů, byly rozděleny na 10 trénovacích množin velikosti 12000 a testovací množinu s 9596 případy.

Iniciální stromy

Metoda CART

Použit R-project a package `tree`.
Jako míra impurity použita deviance.
Trénovací množina rozdělena v poměru 2:1 na množinu pro růst a validační množinu pro prořezávání.
Růst stromu pomocí dokud není v každém listu pouze jedna třída.
Pomocí validační množiny získána sekvence stromů postupným prořezáváním (cost complexity pruning).
V sekvenci prořezaných stromů nalezen strom s nejnižší celkovou deviancí, stromy větší než tento nejsou použity.
Jako nejmenší použit strom se 3 vnitřními uzly.
Dále vybrány stromy v pravidelných rozestupech, aby jich bylo celkem 10.

Iniciální stromy

Metoda C5.0

Použita originální aplikace c5.0 od Rulequest Research.
Prořezávání řízeno parametrem “confidence level”.
Zvoleno 20 různých hodnot parametru, ale někdy prořezávání vedlo k identickým stromům.
Pro experiment použity všechny unikátní stromy, pro každou trénovací množinu jich bylo aspoň 10, zpravidla o něco více.

Obsah

Iniciální stromy

Počty splitů ve stromech

	CART-trees				C5.0-trees			
	počet stromů	min	med	max	počet stromů	min	med	max
Magic	100	3	27.5	75	122	13	71.5	209
Waveform A	100	3	17.5	42	136	19	103	294
Waveform B	100	3	19	45	144	25	107	290
MiniBooNE	100	3	17	46	134	19	90.5	217

Stromy z C5.0 jsou podstatně větší, než stromy z CART.

Iniciální stromy

100AUC nezměkčených stromů

	CART-trees			C5.0-trees		
	min	med	max	min	med	max
Magic	79.59	87.62	89.45	82.59	86.70	88.88
Waveform A	82.85	89.91	91.21	85.85	88.73	89.88
Waveform B	85.52	92.20	93.29	86.81	90.70	92.18
MiniBooNE	84.22	93.42	94.25	86.82	90.09	91.93

Stromy z C5.0 mají nižší AUC, než stromy z CART.

Změkčování bez optimalizace

- Změkčování DR(0) na datech Magic vede spíše ke zhoršení AUC.
- Na stromech z C5.0 na všech datových rozděleních je změkčení DR(2) a C4.5 lepší, než změkčování implementované v C5.0.
- Na stromech z C5.0 je zlepšení získané změkčením větší, než na stromech CART.
- Nejčastěji je nejlepší změkčení DR(1), velice často také DR(2).
- Nelze říci že by některá z metod byla nejlepší univerzálně.

Kombinace s Laplaceovou korekcí

- Na nezměkčených stromech použití Laplaceovy korekce v některých případech vede ke zvýšení AUC, především na stromech z C5.0.
- V kombinaci se změkčením se rozdíl zmenšuje.
- Při změkčení metodami, které samy vedou k většímu zvýšení AUC, tzn. DR(1) a DR(2), Laplaceova korekce už k žádnému dalšímu zlepšení nevede.
- Laplaceova korekce nezpůsobuje žádné podstatné zhoršení.

Změkčování pomocí optimalizace

- Optimalizace diff ve většině případů vedla ke zhoršení vůči iniciálnímu stavu.
- Optimalizace square asi v polovině případů vedla ke zhoršení.
- Optimalizace exptr a AUC většinou klasifikátor zlepšuje.
- Optimalizace AUC dává nejlepší výsledky.
- Porovnání metod na jednotlivých stromech ukazuje, že uspořádání podle kvality je velmi silné.

$$\text{diff} < \text{square} < \text{exptr} < \text{AUC}$$

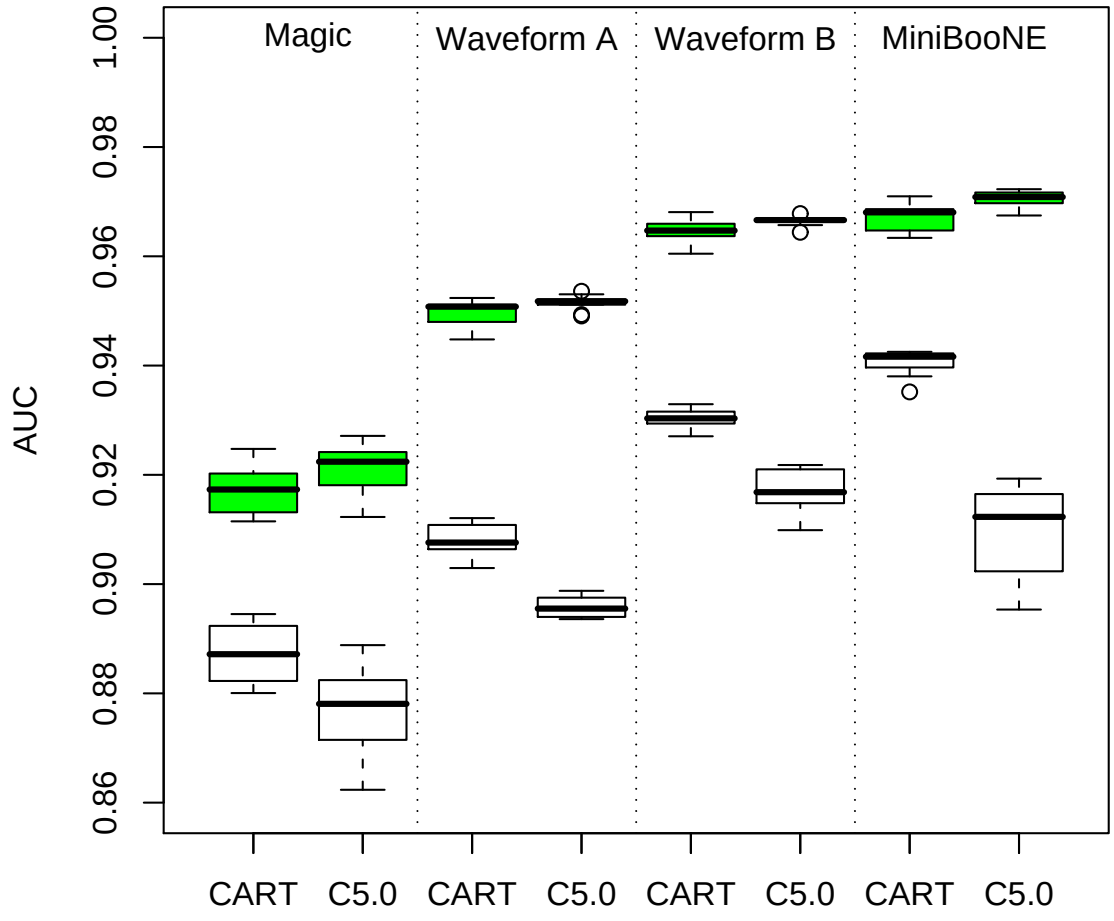
Porovnání nejlepších stromů

Pro každou trénovací množinu vybereme podle AUC na testovací množině:

- nejlepší nezměkčený strom
- nejlepší strom změkčený pomocí optimalizace AUC

zvlášť pro stromy generované CART a generované C5.0.
Porovnáme AUC na testovací množině.

Porovnání nejlepších stromů



Zelená barva označuje změkčené stromy.

Redukce velikosti stromu

Jak použít změkčení pro redukci velikosti stromu bez ztráty kvality?
Vezmeme-li v sekvenci prořezávaných stromů nejlepší nezměkčený strom, lze najít menší změkčený strom, který není horší?
Porovnáváme AUC na testovací množině.
U stromů generovaných C5.0 jsou v našem experimentu všechny změkčené stromy lepší, než nejlepší nezměkčený.
U stromů generovaných CART je nalezený změkčený strom výrazně menší.

Redukce velikosti CART stromu (počet splitů)

Magic										
nezměkčený	45	49	61	64	43	63	38	52	75	44
změkčený	11	17	10	14	13	10	10	8	18	10
Waveform A										
nezměkčený	35	28	42	42	27	25	32	28	42	28
změkčený	7	9	7	8	5	6	7	8	7	5
Waveform B										
nezměkčený	32	26	32	31	45	37	38	44	31	39
změkčený	6	5	7	6	7	7	6	7	8	6
MiniBooNE										
nezměkčený	31	33	23	27	27	34	25	29	39	46
změkčený	13	5	6	5	6	6	5	5	7	7

Obsah

Závěr

- Některé postupy navržené pro minimalizaci error-rate klasifikátoru nemusí být vhodné pro ranking.
- Použití Laplaceovy korekce společně s dostatečně účinným změkčením nepřináší další zlepšení.
- Laplaceova korekce při změkčení nevede ani ke zhoršení.
- Při optimalizaci metodou Nelder-Mead je dosaženo nejlepšího výsledku při použití AUC jako cílové funkce, ačkoliv jde o funkci nespojitou.
- Ačkoliv AUC stromů generovaných C5.0 byla nižší než u stromů z CART, změkčení optimalizací AUC kvalitu vyrovnalo.
- Změkčování může být využito pro významnou redukci velikosti stromu, aniž by se zhoršila kvalita měřená AUC.