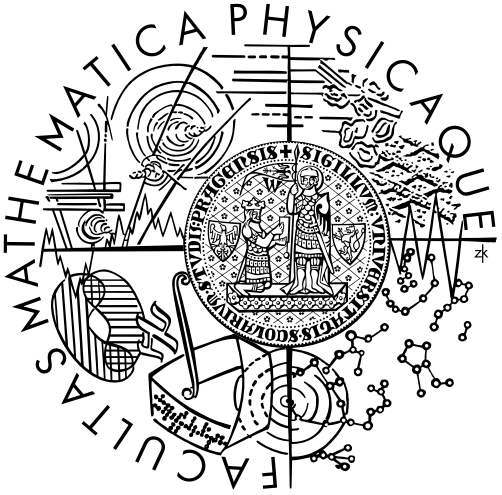
Univerzita Karlova v Praze
Matematicko-fyzikální fakulta

Diplomová práce



Petr Marhoun

Skóringové a klasifikační metody v bankovnictví

Katedra pravděpodobnosti a matematické statistiky

Vedoucí diplomové práce:	RNDr. Petr Franěk, Ph.D.
Studijní program:	Matematika
Studijní obor:	Pravděpodobnost, matematická statistika a ekonometrie
Studijní plán:	Ekonometrie

Poděkování

Rád bych poděkoval RNDr. Petrovi Fraňkovi, Ph.D., za zajímavé téma a za cenné rady, náměty a připomínky.

Prohlašuji, že jsem svou diplomovou práci napsal samostatně a výhradně s použitím citovaných pramenů. Souhlasím se zapůjčováním práce.

V Praze dne 19. dubna 2005 Petr Marhoun

Obsah

Značení

$\boldsymbol{X}$	náhodný vektor modelující obecnou vysvětlující proměnnou (dimenze p)
Y	náhodná veličina modelující obecnou vysvětlovanou proměnnou (hodnoty 0 a 1 - splácený úvěr a default)
$(\boldsymbol{x}, y)$	realizace $(\boldsymbol{X}, Y)$
$\mathcal{L}(\boldsymbol{X}, Y)$	sdružené rozdělení $\boldsymbol{X}$ a Y
$\mathcal{L}(\boldsymbol{X})$	marginální rozdělení $\boldsymbol{X}$
$\mathcal{L}(Y)$	marginální rozdělení Y
$\mathcal{L}(\boldsymbol{X}\|Y)$	podmíněné rozdělení $\boldsymbol{X}$ za předpokladu znalosti Y
$\mathcal{L}(Y\|\boldsymbol{X})$	podmíněné rozdělení Y za předpokladu znalosti $\boldsymbol{X}$
$f(\boldsymbol{x}, y)$	sdružená hustota $\boldsymbol{X}$ a Y
$f(\boldsymbol{x})$	marginální hustota $\boldsymbol{X}$
(π_0, π_1)	marginální hustota Y - apriorní pravděpodobnost
$f(\boldsymbol{x}\|y)$	podmíněná hustota $\boldsymbol{X}$ za předpokladu znalosti Y
$(p_0(\boldsymbol{x}), p_1(\boldsymbol{x}))$	podmíněná hustota Y za předpokladu znalosti $\boldsymbol{X}$ - aposteriorní pravděpodobnost
$p(\boldsymbol{x})$	zjednodušené značení pro $p_1(\boldsymbol{x})$
$(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$	trénovací množina (dimenze $n \times (p+1)$)
$(\boldsymbol{X}_i, Y_i)$	náhodný vektor modelující i-tý z n prvků trénovací množiny (výběr z $\mathcal{L}(\boldsymbol{X}, Y)$, dimenze $p+1$)
$(\boldsymbol{x}_i, y_i)$	realizace $(\boldsymbol{X}_i, Y_i)$ (dimenze $p+1$)
$(\widetilde{\boldsymbol{\mathcal{X}}}, \widetilde{\boldsymbol{\mathcal{Y}}})$	validační množina (dimenze $\widetilde{n} \times (p+1)$)
$(\widetilde{\boldsymbol{X}}_i, \widetilde{Y}_i)$	náhodný vektor modelující i-tý z $\widetilde{n}$ prvků validační množiny (výběr z $\mathcal{L}(\boldsymbol{X}, Y)$, dimenze $p+1$)
$(\widetilde{\boldsymbol{x}}_i, \widetilde{y}_i)$	realizace $(\widetilde{\boldsymbol{X}}_i, \widetilde{Y}_i)$ (dimenze $p+1$)
(c_0, c_1)	vektor ztrát (c_l je ztráta plynoucí z nesprávného zařazení pozorování ve skutečnosti patřícího do třídy l)
(ω_0, ω_1)	rozklad $\mathbb{R}^p$ (platí-li $\boldsymbol{x} \in \omega_l$, pozorování je zařazeno do třídy l)
$\boldsymbol{w}$	váhy
w_0	práh
$w_{in_j}^{h_k}, w_{h_k}^{out}$	indexace vah u neuronových sítí
ρ	ratingová funkce zobrazující $\mathbb{R}^p$ do $\{1, 2, \ldots, R\}$
q_r	předpokládaná odezva pro rating r
$\widetilde{n}, \widetilde{n}_r$	počet pozorování validační množiny celkem a pro rating r
$\widetilde{q}, \widetilde{q}_r$	průměrná odezva validační množiny celkem a pro rating r

Název práce: Skóringové a klasifikační metody v bankovnictví
Autor: Petr Marhoun
Katedra (ústav): Katedra pravděpodobnosti a matematické statistiky
Vedoucí diplomové práce: RNDr. Petr Franěk, Ph.D., Česká národní banka
E-mail vedoucího: petr.franek@cnb.cz

Abstrakt: V bankovnictví stále roste význam klasifikačních a skóringových metod. Tato diplomová práce se zabývá statistickými metodami v praxi používanými pro klasifikaci a tvorbu skóringových modelů. Zaměřuje se tedy na lineární diskriminační analýzu, logistickou regresi a neuronové sítě. Důraz je kladen na předpoklady jednotlivých metod a na hodnocení předpovědní kvality výsledných skóringových modelů. Vlastnosti popisovaných metod jsou demonstrovány na simulační studii. Diskutované postupy jsou implementovány v jazyce R (na přiloženém CD).
Klíčová slova: *klasifikace, skóringové metody, ratingové modely, validace.*

Title: Scoring and Classification Methods in Banking
Author: Petr Marhoun
Department: Department of Probability and Mathematical Statistics
Supervizor: RNDr. Petr Franěk, Ph.D., Czech National Bank
Supervizor's e-mail address: petr.franek@cnb.cz

Abstract: The importance of classification and scoring methods is nowadays increasing in the banking industry. This diploma thesis deals with the statistical methods used in real life for classification and generation of scoring models. It focuses mainly on linear discriminant analysis, logistic regression and neural networks. Emphasis is laid on assumptions of individual methods and on assessment of prediction quality of resulting scoring models. Properties of described methods are demonstrated using simulation study. Discussed principles are implemented in R language (can be seen on attached CD).
Keywords: *classification, scoring methods, rating models, validation.*

1 Motivace

Statisticky podložené klasifikační a skóringové metody se v bankovnictví používají (v USA i v dalších zemích) již po desítky let. Právě v současné době však jejich význam dále roste. V loňském roce totiž byly vydány nové principy výpočtu kapitálové přiměřenosti (tzv. *Basel II*), podle nichž bude výše regulatorního kapitálu počítána právě pomocí těchto metod.

V této části jsou uvedeny některé příklady využití klasifikačních a skóringových metod v oblasti bankovnictví.

Poskytnutí úvěru

Tento příklad je nejen nejsnáze představitelný, ale také nejdůležitější. O většině bankou přijatého rizika se totiž rozhoduje v čase poskytnutí úvěru. Neodhadne-li banka žadatele správně, další aktivity umožní pouze snížení již vzniklé ztráty.

Na procesu poskytnutí úvěru je vhodné ilustrovat některé základní pojmy včetně těch, které se vyskytují v názvu práce. V případě každé žádosti o úvěr banka vyhodnocuje bonitu žadatele. Analyzuje pravděpodobnost, že klient úvěr nesplatí - dojde tedy k tzv. *defaultu*. Banka proto požádá klienta o sdělení některých informací (těmi mohou být např. výše příjmu - 20 tisíc a bydliště - Praha). Na jejich základě spočítá *skóre* (za příjem 20, za Prahu 3, celkem 23 bodů) - proto *skóringové metody.*

Skóre banka porovná s *prahem* a rozhodne o poskytnutí úvěru (práh je 17 bodů, 23 je více než 17, úvěr poskytnut bude). Jinou možností je určit *pravděpodobnost defaultu* a porovnat ho s maximální přípustnou pravděpodobností (23 bodů znamená 4.5 %, což je méně než 6 %, jedná se tedy o přípustné riziko). V obou případech banka tímto rozhodnutím klienta klasifikuje - proto *klasifikační metody.*

Rizikové náklady a tvorba cen

Získá-li klient úvěr, měl by ho nejen splácet, ale také platit úroky. Jejich výše závisí na riziku. Rizikový klient bude platit vysoké úroky, bezrizikový naopak nízké. Banka proto potřebuje odhadovat výši rizika.

Důvodem, proč tomu tak musí být, je konkurenční boj. Pokud by některá banka nabídla všem svým zákazníkům stejný úrok bez ohledu na riziko, které pro ni představují, mohla by se konkurence zachovat stejně. Jen by zvýšila požadavky na bonitu žadatele a snížila úroky. U první banky by pak zůstali pouze rizikovější klienti. To by pro ni znamenalo ztrátu. Pokud by se rozhodla úrokovou sazbu zvýšit (a tím dosáhnout zisku), situace by se mohla opakovat.

Analýza a řízení portfolia

Banka potřebuje vědět, jak se vyvíjí portfolio jejích úvěrů. Mění se? Pokud ano, proč? A je potřeba na situaci nějak reagovat?

Pokud by se např. u hypoték snížila rizikovost některých klientů, mohli by přejít ke konkurenci, která by jim nabídla nižší úrokové sazby. Proto by banka tuto situaci měla předvídat a vybraným zákazníkům snížení nabídnout sama.

Sekuritizace

Pokud již banka portfolio má, může ho dále nabídnout na sekundárním trhu. Jednou z možností je emise cenných papírů podložených úvěry. Pro správné nastavení parametrů těchto cenných papírů je potřeba odhadnout rizikovost jednotlivých pohledávek i celého prodávaného portfolia.

Vymáhání pohledávek

I při správném použití klasifikačních a skóringových metod část klientů úvěr splácet nebude. Banky proto sahají k vymáhání pohledávek, musí však zvolit správný čas a správný způsob. Příliš brzké vymáhání pohledávky může vést ke ztrátě klientů, jejichž platební neschopnost je pouze přechodná. Naopak přílišná prodleva může přinést velkou ztrátu. Odhad budoucího chování je proto u problematických zákazníků ještě důležitější než u ostatních.

Kapitálová přiměřenost a interní ratingy

Důležitou motivací pro používání klasifikačních a skóringových metod je Basel II ([4]), dohoda přijatá v červnu 2004. Jedná se o revizi původní verze přijaté v roce 1988. Ta upravovala minimální kapitálové požadavky pro mezinárodně působící finanční organizace. Stanovila, že banky musí udržovat kapitál ve výši 8 % z tzv. *rizikově vážených aktiv.* Výše těchto aktiv byla určována pevně stanovenými vahami.

Toto revize mění. V budoucnu bude možné namísto pevných vah využívat váhy odvozené od externích ratingů (*standardizovaný přístup*). Banky však budou moci použít i *přístup založený na interních ratinzích,* při němž je výše rizikově vážených aktiv vypočtena pomocí klasifikačních a skóringových metod.

Banky se pro nový přístup nebudou moci rozhodnout samy, využití těchto pokročilých metod bude podléhat souhlasu regulátora. Banky budou muset své metody zdokumentovat, popsat teorii, na které jsou založeny, a ověřit předpoklady. Dále budou muset mít takové množství dat, které umožní dostatečně přesné odhady i validaci ratingových modelů.

A právě těmito úkoly se práce zabývá. V druhé části jsou přesně zformulovány zatím jen volně popsané problémy. Třetí část se zaměřuje na některé metody, které se k jejich řešení využívají, především však na jejich předpoklady. Čtvrtá část ukazuje, jak je možné ratingové modely vytvářet a validovat. V páté části je popsána praktická aplikace teoretických poznatků.

2 Formulace problémů

Předpokládá se, že jsou dána nějaká *pozorování,* pro která jsou známy jak jejich *znaky,* tak i jejich příslušnost k právě jedné z několika *tříd* (v této práci k právě jedné ze dvou tříd). Na jejich základě mají být učiněny závěry o nových pozorováních se známými znaky, ale s neznámou třídou. V této části je zformulováno, co je to pozorování a jaké problémy je možné řešit.

Data

Pozorování se známou i neznámou třídou jsou považována za realizaci $\boldsymbol{x}$ náhodné veličiny $\boldsymbol{X}$. Jde o bod v p-dimenzionálním vektorovém prostoru. To je přirozené u spojitých proměnných (příjem, věk), méně u diskrétních. U ordinálních znaků se jednotlivé faktory přiřazují k několika číslům (vzdělání: 0 - základní, 1 - středoškolské, 2 - vysokoškolské), v jiných případech se znaky kódují pomocí více dimenzí (bydliště: (0, 0) - Praha, (0, 1) - Čechy, (1, 0) - Morava). Některé metody (stromy, loglineární modely) pracují s diskrétními proměnnými přímo, tato práce se však zaměřuje na ty, u kterých to možné není.

Třída je realizací y náhodné veličiny Y. V práci se dále uvažuje, že tato veličina nabývá pouze dvou hodnot: 0 - splácený úvěr a 1 - default.

Pozorování se známou třídou tvoří *trénovací množinu* $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$. Její prvky jsou značeny $(\boldsymbol{X}_i, Y_i)$, jejich realizace pak $(\boldsymbol{x}_i, y_i)$. Předpokládá se, že existuje n takových pozorování dohromady formujících matici dimenze $n \times (p+1)$. Každému pozorování tedy odpovídá jeden řádek.

Na základě trénovací množiny lze sestrojit jeden nebo několik ratingových modelů. Jejich kvalita se hodnotí pomocí *validační množiny* podobné množině trénovací. Značení je odlišené vlnovkou (např. $\boldsymbol{x}_i$ se mění na $\widetilde{\boldsymbol{x}}_i$).

Obě množiny (trénovací a validační) jsou považovány za výběr z $\mathcal{L}(\boldsymbol{X}, Y)$ (sdružené rozdělení $\boldsymbol{X}$ a Y). Jednotlivá pozorování tedy mají být nezávislá. To ve skutečnosti většinou neplatí (např. ekonomická situace ovlivňuje více pozorování podobným způsobem). Některé metody (lineární diskriminační analýza) zase předpokládají výběr z $\mathcal{L}(\boldsymbol{X}|Y)$ (podmíněné rozdělení $\boldsymbol{X}$ za předpokladu znalosti Y) a známé apriorní pravděpodobnosti příslušnosti k jednotlivým třídám, značené π_0 a π_1.

Právě sběrem dat se oblast bankovnictví odlišuje od lékařství, kde se klasifikační a skóringové metody rovněž často používají a mnohé postupy odtud pocházejí. V obou případech je značný nesoulad mezi velikostí jednotlivých tříd (je mnohem více splácených úvěrů než defaultů a mnohem více zdravých lidí než pacientů trpící určitou chorobou). Ale zatímco v lékařství je drahé a bezúčelné shromažďovat data patřící do větší třídy, banky mají údajů o splácených úvěrech dostatek.

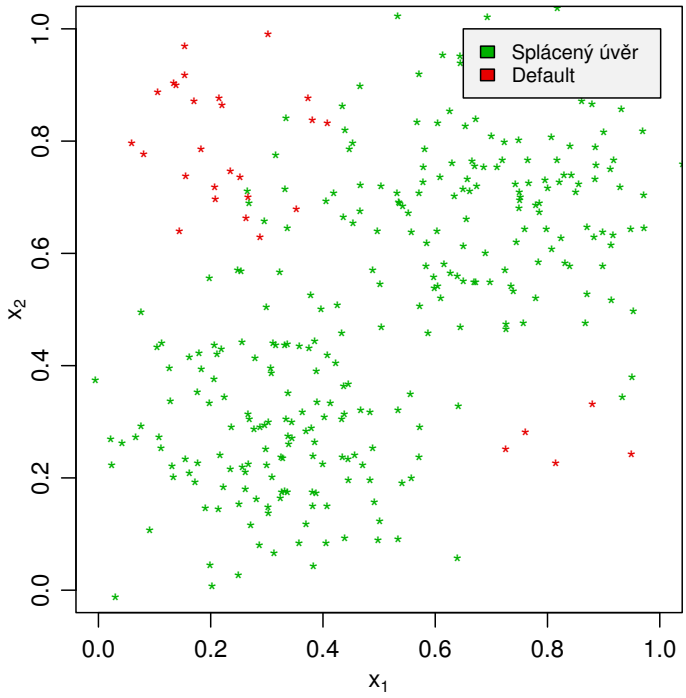
Naopak "špatní" žadatelé úvěr většinou nedostanou, a tak není známo, zda by ho splatili. Tímto problémem, označovaným jako *reject inference*, se zabývá [11], str. 181 - 190. Ideální, ale většinou nerealistickou možností je některým odmítnutým úvěr poskytnout. Jiné varianty (problém ignorovat, považovat všechny odmítnuté za defaulty, modelovat v obou třídách zvlášť, extrapolovat) již tak dobré nejsou ([11], str. 184).

Tato diplomová práce však není o datech, ale o metodách, modelech a validaci. Proto se těmito zajímavými problémy téměř nezabývá. Očekává, že vstupem je trénovací (a případně validační) množina neobsahující nesmyslná pozorování, se znaky, jejichž přítomnost v modelu je potřebná, a se správným podílem obou tříd. Nepopisuje testy schopné odhalit chyby v datech nebo nepotřebné vysvětlující proměnné.

Na jeden aspekt týkající se dat se však práce zaměřuje - na potřebnou velikost trénovací a validační množiny. Nelze ale postupovat přímo - bylo by užitečné vědět, že při použití logistické regrese a 12 vstupech je potřeba např. 343 pozorování, ale takové odpovědi statistika většinou nedává. Tato úloha je ovšem řešitelná i nepřímým způsobem - prostřednictvím vlastností intervalových odhadů některých statistik. Dva různé přístupy jsou zmíněny dále v této části.

Data pro vizualizaci

Skutečná data mívají dimenzi neumožňující snadnou vizualizaci. Pro tento účel jsou tedy v této a v následující části použita data znázorněná na obrázku 1. Metoda generování je popsána v příloze A.1.



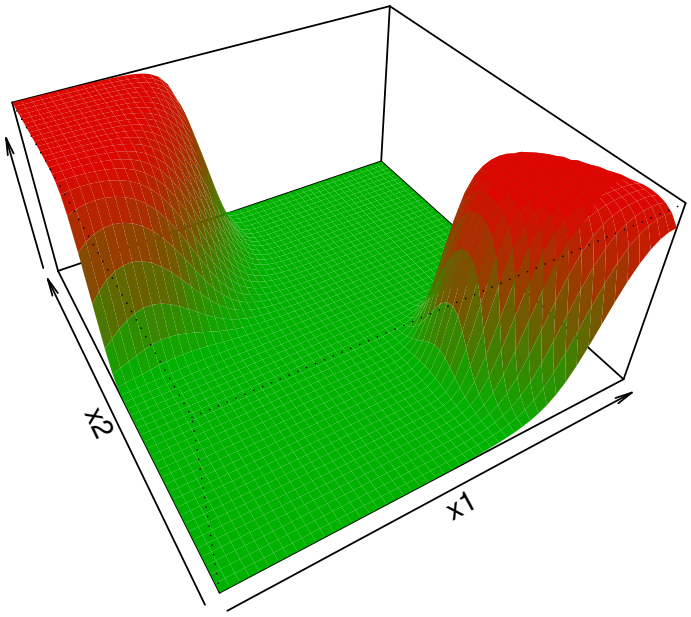
Obrázek 1: Data pro vizualizaci

Čtvrtá část se zabývá postupy graficky hodnotícími ratingové modely bez ohledu na dimenzi dat. Proto jsou v ní pro vizualizaci využita realističtější data popsaná v části 5.

Podmíněná pravděpodobnost

Prvním problémem, který klasifikační a skóringové metody řeší, je konstrukce bodových a intervalových odhadů podmíněné pravděpodobnosti defaultu. Tato pravděpodobnost je značena $p(\boldsymbol{x})$, někdy také $p_1(\boldsymbol{x})$ (pravděpodobnost splácení úvěru je pak $p_0(\boldsymbol{x})$). Mluví se o ní také jako o aposteriorní pravděpodobnosti - vyjadřuje, jaká je pravděpodobnost defaultu s využitím informace, která je o pozorování známa.

Na obrázku 2 je vidět, jak tento odhad může vypadat.



Obrázek 2: Podmíněná pravděpodobnost odhadnutá metodou neuronových sítí

Klasifikace

Druhým problémem je klasifikace. Ta se chápe jako sestrojení pravidla, na jehož základě jsou nová pozorování řazena do jednotlivých tříd. Geometricky to znamená rozdělení $\mathbb{R}^p$ na dvě podmnožiny ω_0 a ω_1. Do třídy l je pozorování zařazeno právě tehdy, je-li prvkem množiny ω_l.

Je-li známá (či odhadnutá) podmíněná pravděpodobnost defaultu a je-li cílem minimalizace ztráty vzniklé chybným zařazením některých pozorování, je klasifikace triviální (což bude nyní ukázáno). Je však ještě potřeba zavést kladná čísla c_0 a c_1 - c_l je ztráta plynoucí z nesprávného zařazení pozorování ve skutečnosti patřícího do třídy l. Dá se očekávat, že c_1 (banka poskytne úvěr, který nebude splácen) je mnohem větší než c_0 (banka neposkytne úvěr a přijde tak o úroky).

Podmíněná střední ztráta (střední ztráta plynoucí z klasifikace pozorování patřících do třídy l) je

$$\int_{\omega_{1-l}} c_l f(\boldsymbol{x}|l)\,\mathrm{d}\boldsymbol{x}$$

($f(\boldsymbol{x}|l)$ je podmíněná hustota $\boldsymbol{X}$ za předpokladu, že Y nabývá hodnoty l), nepodmíněná střední ztráta pak je

$$\pi_0 \int_{\omega_1} c_0 f(\boldsymbol{x}|0)\,\mathrm{d}\boldsymbol{x} + \pi_1 \int_{\omega_0} c_1 f(\boldsymbol{x}|1)\,\mathrm{d}\boldsymbol{x}.$$

Cílem je minimalizace střední ztráty. Do třídy 1 jsou proto zařazena právě ta pozorování, pro která platí

$$c_0 f(\boldsymbol{x}|0)\pi_0 < c_1 f(\boldsymbol{x}|1)\pi_1,$$

po úpravě

$$\frac{f(\boldsymbol{x}|1)\,\pi_1}{f(\boldsymbol{x}|0)\,\pi_0} > \frac{c_0}{c_1}. \tag{1}$$

Použitím Bayesovy věty

$$p_l(\boldsymbol{x}) = \frac{f(\boldsymbol{x}, l)}{f(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|l)\pi_l}{f(\boldsymbol{x})}$$

($f(\boldsymbol{x}, l)$ je sdružená hustota $(\boldsymbol{X}, Y)$, $f(\boldsymbol{x})$ je marginální hustota $\boldsymbol{X}$), lze pravidlo (1) upravit na

$$\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} > \frac{c_0}{c_1}. \tag{2}$$

S využitím toho, že

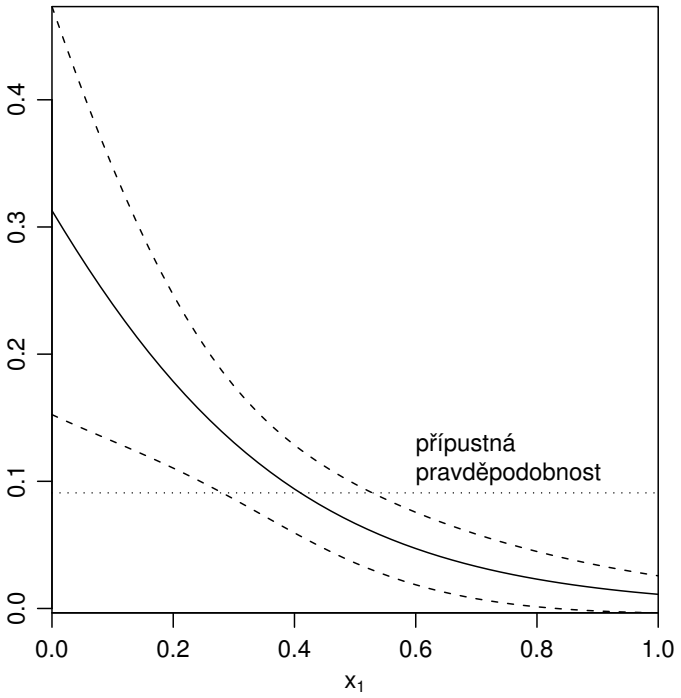
$$p(\boldsymbol{x}) := p_1(\boldsymbol{x}) = 1 - p_0(\boldsymbol{x}),$$

je možné pravidlo (2) přepsat do tvaru

$$p(\boldsymbol{x}) > \frac{c_0}{c_0 + c_1}. \tag{3}$$

Nyní je již vidět souvislost s podmíněnou pravděpodobností - pozorování lze klasifikovat porovnáním odhadu podmíněné pravděpodobnosti $p(\boldsymbol{x})$ s výrazem na pravé straně pravidla (3) (viz obrázek 3). Intervalový odhad určuje statistickou významnost klasifikace. Je-li totiž $c_0/(c_0 + c_1)$ prvkem tohoto intervalu, není možné rozhodnout, zda je klasifikace opodstatněná.

Stane-li se tak v nadměrně mnoha případech, může to znamenat, že trénovací množina je příliš malá. Toto je tedy první z přístupů umožňujících rozhodnout, zda je k dispozici dostatečné množství dat.

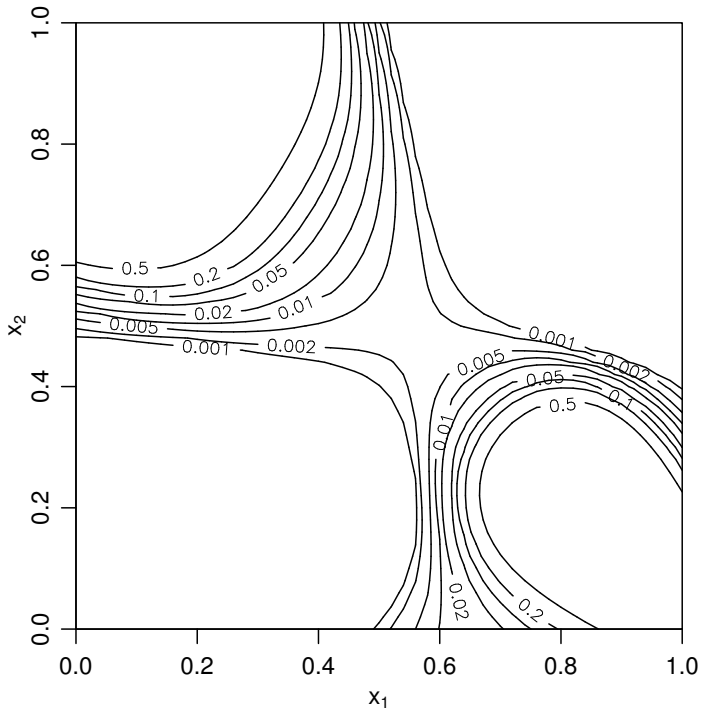


Obrázek 3: Podmíněná pravděpodobnost odhadnutá metodou logistické regrese (pouze jeden vstup, včetně intervalových odhadů)

Ratingové modely

Třetím problémem, přímo vycházejícím z [4], je tvorba ratingových modelů. Hledá se funkce ρ, která bude pozorováním přiřazovat rating - prvek množiny $\{1, 2, \ldots, R\}$. Pozorování s vyšším ratingem by měla být kvalitnější, mít menší pravděpodobnost defaultu. Navíc je nutné, aby ratingová pásma byla dostatečně úzká, nikde by pozorování nemělo být přespříliš. Při známých pravděpodobnostech defaultu může postačovat, aby byly určeny hranice oddělující jednotlivé třídy (viz obrázek 4). Jsou-li však na ratingy kladeny další požadavky, je tato úloha komplexnější.

Kvalitu ratingového modelu by měly zhodnotit statistiky vypočtené na základě validační množiny. A právě intervalové odhady těchto statistik využívá druhý z přístupů umožňujících posuzovat velikost trénovací a validační množiny. Dat je dostatek, jestliže tyto intervaly nejsou příliš široké.



Obrázek 4: Ratingy vytvořené metodou neuronových sítí

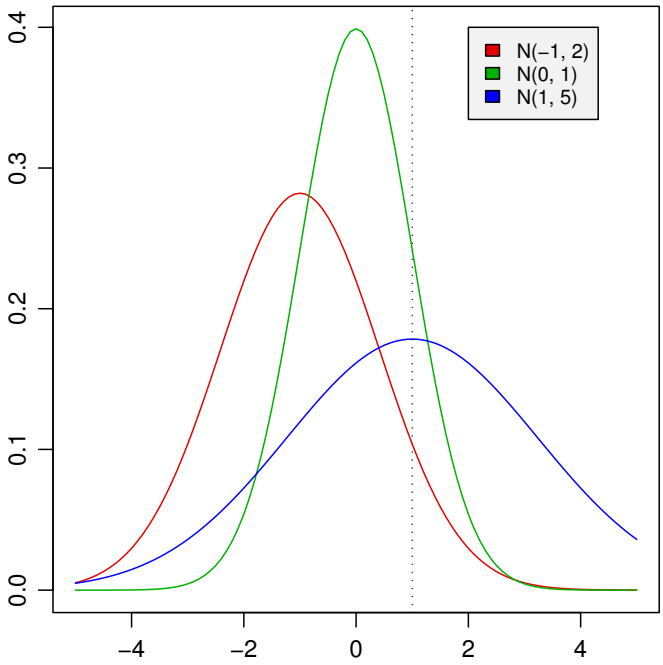
3 Používané metody

K řešení daných problémů (odhadu podmíněné pravděpodobnosti a klasifikaci) lze využít mnoho různých metod. Tato práce se soustředí především na tři z nich - lineární diskriminační analýzu, logistickou regresi a neuronové sítě. Neznamená to, že by musely být (podle některého hlediska) lepší než jiné. Důvodem je jejich rozšířenost v oblasti bankovnictví. V závěru této části jsou však uvedeny i některé z ostatních metod.

Na úvod jsou zmíněny dvě zajímavé vlastnosti. Jedna je pro všechny tři popisované metody společná, druhá je zase charakteristická jen pro lineární diskriminační analýzu a logistickou regresi.

Maximální věrohodnost

Různé metody mají různé předpoklady. Jeden z hlavních je platnost určitého modelu s neznámými parametry. Jednotlivé modely se liší, tři zvolené metody jsou však spojeny kritériem hodnotícím možné hodnoty těchto neznámých parametrů. Za nejvhodnější jsou považovány takové hodnoty, které pro danou trénovací množinu maximalizují sdruženou hustotu - jsou tedy *maximálně věrohodné*. Jak je tomu v jednom jednoduchém případě, ukazuje obrázek 5.



Obrázek 5: Jestliže je pozorována hodnota jedna, pak nejvěrohodnější odhad střední hodnoty vychází z normálního rozdělení s nulovou střední hodnotou a jednotkovým rozptylem (je-li možné volit ze tří zobrazených možností).

Často se neprovádí maximalizace sdružené hustoty, ale operace v některých případech ekvivalentní - minimalizace tzv. *chybové funkce.* Tato souvislost bude nyní ukázána pro případ normálního rozdělení.

Za předpokladu, že náhodná veličina Y má podmíněné normální rozdělení

$$\mathcal{L}(Y_i|\boldsymbol{X}_i) = N(p(\boldsymbol{X}_i), \sigma^2),$$

je sdružená hustota

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - p(\boldsymbol{x}_i))^2}{2\sigma^2}\right).$$

Logaritmus tohoto výrazu má tvar

$$-\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - p(\boldsymbol{x}_i))^2.$$

Hledání maximálně věrohodného odhadu podmíněné střední hodnoty tedy dává v tomto případě stejné výsledky jako minimalizace chybové funkce

$$\sum_{i=1}^{n}(y_i - p(\boldsymbol{x}_i))^2,$$

známé jako *součet čtverců.*

Použití metody maximální věrohodnosti má navíc jeden zajímavý důsledek. Za obecných předpokladů ([2], str. 157 - 159) jsou získané odhady konzistentní a asymptoticky normální. Toho lze využít ke konstrukci intervalových odhadů.

Linearita

Lineární diskriminační analýza a logistická regrese se ve svých předpokladech liší, výsledky však často dávají podobné. Obě metody totiž oddělují ω_0 od ω_1 nadrovinou, do třídy defaultů tedy zařadí taková pozorování, pro která platí

$$\boldsymbol{w}^T\boldsymbol{x} > w_0,$$

kde $\boldsymbol{w}$ jsou *váhy* a w_0 je *práh.* Neuronové sítě mají složitější tvar a umějí separovat defaulty prostřednictvím nelineárních nadploch.

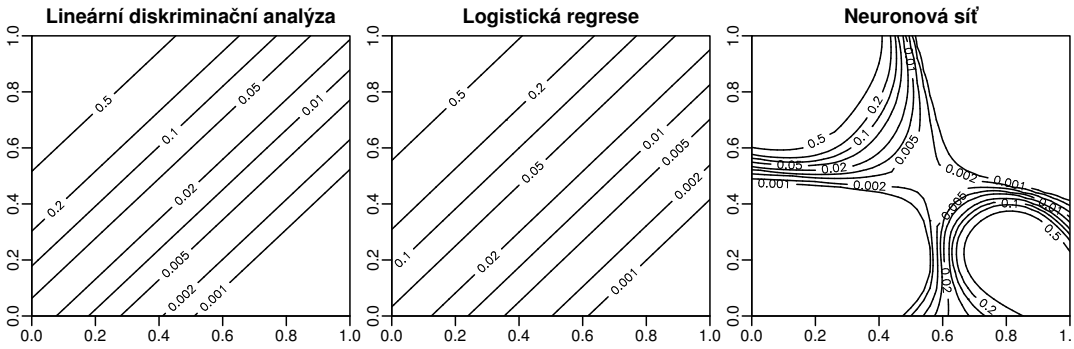
Tato vlastnost se přenáší i na ratingy (viz obrázek 6).

3.1 Lineární diskriminační analýza

V případě lineární diskriminační analýzy se předpokládá, že podmíněné rozdělení vektoru $\boldsymbol{X}$ je normální s varianční maticí nezávisející na skupině, tedy že platí

$$\mathcal{L}(\boldsymbol{X}|Y = l) = N_p(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}), \quad \boldsymbol{\Sigma} > 0, \quad l = 0, 1.$$

Dále se očekává, že jsou známé apriorní pravděpodobnosti π_0 a π_1.



Obrázek 6: Ratingy vytvořené různými metodami.

Váhy

Hustota normálního rozdělení je

$$f(\boldsymbol{x}|l) = (2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{(\boldsymbol{x}-\boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_l)}{2}\right).$$

Výraz na levé straně pravidla (1) má tvar

$$\frac{f(\boldsymbol{x}|1)\,\pi_1}{f(\boldsymbol{x}|0)\,\pi_0} = \frac{(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)/\,2\right)\,\pi_1}{(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)/\,2\right)\,\pi_0}.$$

Logaritmováním, zkrácením a úpravou se dostává

$$\begin{aligned}
\log\frac{f(\boldsymbol{x}|1)\,\pi_1}{f(\boldsymbol{x}|0)\,\pi_0} &= \log\frac{\exp\left(-(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)/\,2\right)\,\pi_1}{\exp\left(-(\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)/\,2\right)\,\pi_0} \\
&= -\frac{(\boldsymbol{x}-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1) - (\boldsymbol{x}-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_0)}{2} + \log\frac{\pi_1}{\pi_0} \\
&= -\frac{-2\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} + \boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 + 2\boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0}{2} + \log\frac{\pi_1}{\pi_0} \\
&= (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0}{2} + \log\frac{\pi_1}{\pi_0}. \qquad (4)
\end{aligned}$$

Bude-li z n pozorování trénovací množiny patřit do l-té třídy právě n_l z nich, lze apriorní pravděpodobnosti odhadnout podílem

$$\widehat{\pi}_l = \frac{n_l}{n},$$

další odhady se získají metodou maximální věrohodnosti ([3], str. 67 - 70, 219)

$$\widehat{\boldsymbol{\mu}}_l = \frac{1}{n_l}\sum_{y_i=l}\boldsymbol{x}_i$$

a

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\left(\sum_{y_i=0}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_0)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_0)^T + \sum_{y_i=1}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_1)(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_1)^T\right).$$

Pro výpočet skóre se odhadnou váhy

$$\widehat{\boldsymbol{w}} = \widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0), \tag{5}$$

odhad prahu je

$$\widehat{w}_0 = \frac{\widehat{\boldsymbol{\mu}}_1^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0^T\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\mu}}_0}{2} - \log\frac{\widehat{\pi}_1}{\widehat{\pi}_0} + \log\frac{c_0}{c_1}$$

(výraz z pravé strany pravidla (1) musí být také zlogaritmován).

Podmíněná pravděpodobnost

Podle (4) platí

$$\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})} = \frac{f(\boldsymbol{x}|1)\,\pi_1}{f(\boldsymbol{x}|0)\,\pi_0} = \exp\left((\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}\boldsymbol{x} - \frac{\boldsymbol{\mu}_1^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_0}{2} + \log\frac{\pi_1}{\pi_0}\right).$$

Tento vzorec umožňuje výpočet konzistentních bodových odhadů podmíněné pravděpodobnosti, analytický vztah pro intervalové odhady však v literatuře popsán není. Je ale možné využít metodu bootstrapu (viz část 4.5).

Předpoklady metody

Mezi předpoklady lineární diskriminační analýzy patří:

- znalost apriorních pravděpodobností,
- nezávislost jednotlivých pozorování,
- normalita dat,
- shoda variančních matic z obou tříd (homoskedasticita).

Specifické jsou především poslední dva body, proto jsou zde uvedeny některé testy schopné tyto předpoklady ověřit. Jsou zmíněny také možné reakce na problémy, i když dále bude ukázáno, že k nejdůležitějšímu výsledku (odhadu vah) normalita nutná není.

Normalita je ověřitelná prostřednictvím šikmosti a špičatosti. Pokud se definují statistiky

$$a_{l,p} = \frac{1}{n_l^2}\sum_{y_i=l}\sum_{y_{i'}=l}\left((\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_l)^T\widehat{\boldsymbol{\Sigma}}_l^{-1}(\boldsymbol{x}_{i'} - \widehat{\boldsymbol{\mu}}_l)\right)^3, \quad l = 0, 1$$

a

$$b_{l,p} = \frac{1}{n_l}\sum_{y_i=l}\left((\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_l)^T\widehat{\boldsymbol{\Sigma}}_l^{-1}(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_l)\right)^2, \quad l = 0, 1$$

($\widehat{\boldsymbol{\Sigma}}_l$ je maximálně věrohodný odhad varianční matice v l-té třídě) a pokud se předpokládá normální model s různými variančními maticemi, tedy $\mathcal{L}(\boldsymbol{X}|Y = l) = N_p(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$, pak asymptoticky platí ([12], str. 173)

$$\frac{n_l}{6} a_{l,p} \sim \chi^2_d, \quad l = 0, 1,$$

kde $d = p(p+1)(p+2)/6$, a

$$\frac{b_{l,p} - p(p+2)}{\sqrt{(8/n_l)p(p+2)}} \sim N(0,1), \quad l = 0, 1.$$

Za předpokladu normality v obou modelech zároveň dále platí (rovněž [12], str. 173)

$$\frac{n_0}{6} a_{0,p} + \frac{n_1}{6} a_{1,p} \sim \chi^2_{2d}$$

a

$$\frac{(n_0/n)b_{0,p} + (n_1/n)b_{1,p} - p(p+2)}{\sqrt{(8/n)p(p+2)}} \sim N(0,1).$$

I když tento předpoklad splněn není, je možné normality dosáhnout zobecněnou Box-Coxovou transformací ([12], str. 178 - 180). Matice $\boldsymbol{\mathcal{X}}$ je nahrazena transformovanou maticí s prvky

$$x_{ij}^{(\lambda_j)} = \begin{cases} (x_{ij}^{\lambda_j} - 1)/\lambda_j, & \lambda_j \neq 0, \\ \log x_{ij}, & \lambda_j = 0. \end{cases}$$

Vektor $\boldsymbol{\lambda}$ je považován za další parametr a je odhadován (spolu s vektory středních hodnot a varianční maticí) metodou maximální věrohodnosti.

Pro test homoskedasticity se očekává, že podmíněné rozdělení je skutečně normální, tedy $\mathcal{L}(\boldsymbol{X}|Y = l) = N_p(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$. Za předpokladu shody variančních matic ($\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_1$) platí ([12], str. 174 - 175)

$$n_0 \log \frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_0|} + n_1 \log \frac{|\widehat{\boldsymbol{\Sigma}}|}{|\widehat{\boldsymbol{\Sigma}}_1|} \sim \chi_{p(p+1)/2}.$$

V případě heteroskedasticity je možné zůstat u modelu s podmíněnými rozděleními $\mathcal{L}(\boldsymbol{X}|Y = l) = N_p(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$ a použít dále popsanou kvadratickou diskriminační analýzu.

Lineární diskriminace založená na maximalizaci separace

Každé lineární diskriminační pravidlo má tvar

$$\boldsymbol{w}^T \boldsymbol{x} > w_0, \quad \boldsymbol{w} \neq \boldsymbol{0},$$

kde $\boldsymbol{w}$ jsou váhy provádějící projekci prostoru $\mathbb{R}^p$ na přímku. Zajímavým přístupem nezaloženým na předpokladu normality je hledání takových vah, které na této přímce maximalizují vzdálenost mezi průměry jednotlivých tříd a zároveň minimalizují vnitrotřídní variabilitu. To znamená, že maximalizují poměr (nezávisející na velikosti $\boldsymbol{w}$)

$$\frac{(\boldsymbol{w}^T \widehat{\boldsymbol{\mu}}_1 - \boldsymbol{w}^T \widehat{\boldsymbol{\mu}}_0)^2}{\boldsymbol{w}^T \widehat{\boldsymbol{\Sigma}} \boldsymbol{w}},$$

po úpravě

$$\frac{\boldsymbol{w}^T(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)^T\boldsymbol{w}}{\boldsymbol{w}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{w}}.$$

Maximum se nalezne, když se tento poměr zderivuje a položí roven nule. Musí platit (matici dimenze 1×1 lze považovat za skalár)

$$\frac{\boldsymbol{w}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{w}\ 2(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)^T\boldsymbol{w} - \boldsymbol{w}^T(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)^T\boldsymbol{w}\ 2\widehat{\boldsymbol{\Sigma}}\boldsymbol{w}}{(\boldsymbol{w}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{w})^2} = \boldsymbol{0},$$

po úpravě

$$\boldsymbol{w}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{w}\ (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)^T\boldsymbol{w} = \boldsymbol{w}^T(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)^T\boldsymbol{w}\ \widehat{\boldsymbol{\Sigma}}\boldsymbol{w}.$$

Protože zajímavý je pouze směr, lze vynechat výrazy $\boldsymbol{w}^T\widehat{\boldsymbol{\Sigma}}\boldsymbol{w}$, $(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)^T\boldsymbol{w}$ a $\boldsymbol{w}^T(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)^T\boldsymbol{w}$. Optimální váhy jsou tedy řešením rovnice

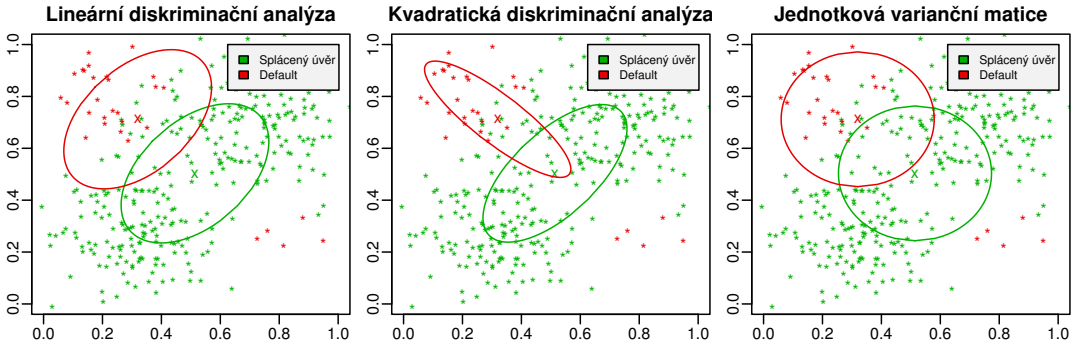
$$\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0 = \widehat{\boldsymbol{\Sigma}}\boldsymbol{w}$$

a shodují se s výrazem (5), odvozeným za předpokladu normality.

Tato metoda nepředpokládá o rozdělení vůbec nic. Je založena na odhadech středních hodnot a (shodné) varianční matice, dává tedy optimální lineární diskriminační pravidlo pro libovolné rozdělení popsané prvními dvěma momenty. Určit práh a podmíněnou pravděpodobnost však bez dalších předpokladů možné není.

Příbuzné metody

Metody příbuzné lineární diskriminační analýze předpokládají jiný tvar varianční matice. Různé varianty ukazuje obrázek 7.



Obrázek 7: Oblasti se stejnou hustotou, různé metody.

V případě kvadratické diskriminační analýzy se předpokládá, že podmíněná rozdělení mají odlišnou varianční matici, tedy

$$\mathcal{L}(\boldsymbol{X}|Y = l) = N_p(\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l), \quad l = 0, 1.$$

Tato metoda je vhodná zvláště v situaci, kdy byla zamítnuta hypotéza o homoskedasticitě. Nevýhodou je nutnost odhadovat více parametrů. Navíc množiny ω_0 a ω_1 nejsou odděleny nadrovinou, ale kvadratickou nadplochou.

Při nedostatku dat může být naopak problém odhadovat parametry (společné) matice $\boldsymbol{\Sigma}$ a je vhodné kovarianční strukturu zanedbat úplně. V tom případě se použije model

$$\mathcal{L}(\boldsymbol{X}|Y=l) = N_p(\boldsymbol{\mu}_l, \sigma^2 \boldsymbol{I}_p), \quad l = 0, 1.$$

Data je však potřeba standardizovat - transformovat je takovým způsobem, aby rozptyl všech složek vektoru $\boldsymbol{X}$ byl přibližně stejný.

Kompromisem mezi jednotlivými metodami může být regularizovaná diskriminační analýza ([12], str. 144 - 152). V první fázi se za varianční matici volí

$$\widehat{\boldsymbol{\Sigma}}_l(\lambda) = \frac{(1-\lambda)(n_l-1)\widehat{\boldsymbol{\Sigma}}_l + \lambda(n-2)\widehat{\boldsymbol{\Sigma}}}{(1-\lambda)(n_l-1) + \lambda(n-2)}, \quad \lambda \in [0,1], \quad l = 0, 1.$$

Extrémní případy tedy znamenají lineární a kvadratickou diskriminační analýzu. V druhé fázi se přidá druhý parametr γ a dostane se

$$\widehat{\boldsymbol{\Sigma}}_l(\lambda, \gamma) = (1-\gamma)\widehat{\boldsymbol{\Sigma}}_l(\lambda) + \gamma \frac{\operatorname{tr} \widehat{\boldsymbol{\Sigma}}_l(\lambda)}{p} \boldsymbol{I}_p, \quad \lambda, \gamma \in [0,1], \quad l = 0, 1.$$

Parametry λ a γ se hledají tak, aby ztráta vzniklá chybnou klasifikací trénovacích dat byla co nejmenší.

3.2 Logistická regrese

V případě logistická regrese se předpokládá, že podmíněná pravděpodobnost $p(\boldsymbol{x})$ závisí na $\boldsymbol{X}$ prostřednictvím logistické funkce. Platí tedy

$$\begin{aligned} p_1(\boldsymbol{x}) = p(\boldsymbol{x}) &= \frac{\exp(\boldsymbol{w}^T\boldsymbol{x})}{1+\exp(\boldsymbol{w}^T\boldsymbol{x})}, \\ p_0(\boldsymbol{x}) = 1 - p(\boldsymbol{x}) &= \frac{1}{1+\exp(\boldsymbol{w}^T\boldsymbol{x})}, \end{aligned} \tag{6}$$

kde $\boldsymbol{w} \in \mathbb{R}^p$ jsou (neznámé) váhy.

Váhy

Skóre $\boldsymbol{w}^T\boldsymbol{x}$ lze použít ke klasifikaci přímo. Platí totiž

$$\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})} = \frac{\dfrac{\exp(\boldsymbol{w}^T\boldsymbol{x})}{1+\exp(\boldsymbol{w}^T\boldsymbol{x})}}{\dfrac{1}{1+\exp(\boldsymbol{w}^T\boldsymbol{x})}} = \exp(\boldsymbol{w}^T\boldsymbol{x}).$$

Pravidlo (2) je tedy v případě logistické regrese ekvivalentní s nerovností

$$\boldsymbol{w}^T\boldsymbol{x} > \log\frac{c_0}{c_1}.$$

Váhy $\boldsymbol{w}$ se odhadují metodou maximální věrohodnosti. Sdružená hustota (závislá na váhách a proto značená $L_n(\boldsymbol{w})$) výběru $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$ má tvar

$$L_n(\boldsymbol{w}) = \prod_{i=1}^{n} f(\boldsymbol{x}_i, y_i|\boldsymbol{w}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{x}_i, \boldsymbol{w}) f(\boldsymbol{x}_i) = \prod_{i=1}^{n} (p(\boldsymbol{x}_i))^{y_i} (1 - p(\boldsymbol{x}_i))^{1-y_i} f(\boldsymbol{x}_i).$$

Její logaritmus (značený $l_n(\boldsymbol{w})$) je

$$\begin{aligned}
l_n(\boldsymbol{w}) &= \log(L_n(\boldsymbol{w})) \\
&= \sum_{i=1}^{n} \Big(y_i \log(p(\boldsymbol{x}_i)) + (1 - y_i) \log(1 - p(\boldsymbol{x}_i)) + \log(f(\boldsymbol{x}_i)) \Big) \\
&= \sum_{i=1}^{n} \Big(y_i \log\frac{p(\boldsymbol{x}_i)}{1 - p(\boldsymbol{x}_i)} + \log(1 - p(\boldsymbol{x}_i)) + \log(f(\boldsymbol{x}_i)) \Big) \\
&= \sum_{i=1}^{n} \Big(y_i \, \boldsymbol{w}^T\boldsymbol{x}_i - \log(1 + \exp(\boldsymbol{w}^T\boldsymbol{x}_i)) + \log(f(\boldsymbol{x}_i)) \Big).
\end{aligned}$$

Po zderivování platí

$$\frac{\partial l_n}{\partial w_j}(\boldsymbol{w}) = \sum_{i=1}^{n} \Big(y_i \, x_{ij} - \frac{\exp(\boldsymbol{w}^T\boldsymbol{x}_i)}{1 + \exp(\boldsymbol{w}^T\boldsymbol{x}_i)} x_{ij} \Big) = \sum_{i=1}^{n} \Big(y_i \, x_{ij} - p(\boldsymbol{x}_i) x_{ij} \Big),$$

vektorově

$$\frac{\partial l_n}{\partial \boldsymbol{w}}(\boldsymbol{w}) = \boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{Y}} - \boldsymbol{\mathcal{X}}^T p(\boldsymbol{\mathcal{X}}).$$

Maximálně věrohodný odhad $\widehat{\boldsymbol{w}}$ tedy splňuje tzv. *věrohodnostní rovnici*

$$\boldsymbol{\mathcal{X}}^T\boldsymbol{\mathcal{Y}} = \boldsymbol{\mathcal{X}}^T \frac{\exp(\boldsymbol{\mathcal{X}}\widehat{\boldsymbol{w}})}{\mathbf{1} + \exp(\boldsymbol{\mathcal{X}}\widehat{\boldsymbol{w}})}. \tag{7}$$

Pro intervalové odhady je dále potřeba druhá derivace

$$\begin{aligned}
\frac{\partial^2 l_n}{\partial w_j \partial w_{j'}}(\boldsymbol{w}) &= -\sum_{i=1}^{n} \frac{\partial}{\partial w_{j'}} \left(\frac{\exp(\boldsymbol{w}^T\boldsymbol{x}_i)}{1 + \exp(\boldsymbol{w}^T\boldsymbol{x}_i)} \right) x_{ij} \\
&= -\sum_{i=1}^{n} \frac{\exp(\boldsymbol{w}^T\boldsymbol{x})}{(1 + \exp(\boldsymbol{w}^T\boldsymbol{x}))^2} x_{ij} x_{ij'} \\
&= -\sum_{i=1}^{n} p(\boldsymbol{x}_i)(1 - p(\boldsymbol{x}_i)) x_{ij} x_{ij'},
\end{aligned}$$

vektorově

$$\frac{\partial^2 l_n}{\partial \boldsymbol{w} \partial \boldsymbol{w}^T}(\boldsymbol{w}) = -\boldsymbol{\mathcal{X}}^T \operatorname{diag}\{p(\boldsymbol{x}_i)(1-p(\boldsymbol{x}_i)),\ i=1,\ldots,n\}\boldsymbol{\mathcal{X}}.$$

Za obecných předpokladů metody maximální věrodnosti lze pomocí druhé derivace logaritmu sdružené hustoty zkonstruovat intervalové odhady vah ([2], str. 118 - 119, 157 - 159). Asymptopicky platí

$$\widehat{\boldsymbol{w}} \sim N(\boldsymbol{w}, \mathrm{E}(\boldsymbol{\mathcal{X}}^T \operatorname{diag}\{p(\boldsymbol{x}_i)(1-p(\boldsymbol{x}_i)),\ i=1,\ldots,n\}\boldsymbol{\mathcal{X}})^{-1}).$$

Po dosazení odhadů podmíněné pravděpodobnosti se dostane

$$\operatorname{var}(\widehat{\boldsymbol{w}}) \doteq (\boldsymbol{\mathcal{X}}^T \operatorname{diag}\{\widehat{p}(\boldsymbol{x}_i)(1-\widehat{p}(\boldsymbol{x}_i)),\ i=1,\ldots,n\}\boldsymbol{\mathcal{X}})^{-1}. \tag{8}$$

Hledání optimálních vah

Váhy se hledají numerickým řešením věrohodnostní rovnice (7). K tomuto účelu se využije Taylorův rozvoj logaritmu sdružené hustoty ($\widehat{\boldsymbol{w}}_t$ značí aproximaci odhadu vah z t-tého kroku)

$$l_n(\widehat{\boldsymbol{w}}) \doteq l_n(\widehat{\boldsymbol{w}}_t) + l_n'(\widehat{\boldsymbol{w}}_t)^T(\widehat{\boldsymbol{w}} - \widehat{\boldsymbol{w}}_t) + \frac{1}{2}(\widehat{\boldsymbol{w}} - \widehat{\boldsymbol{w}}_t)^T l_n''(\widehat{\boldsymbol{w}}_t)(\widehat{\boldsymbol{w}} - \widehat{\boldsymbol{w}}_t).$$

Tento výraz se zderivuje

$$l_n'(\widehat{\boldsymbol{w}}) \doteq l_n'(\widehat{\boldsymbol{w}}_t) + l_n''(\widehat{\boldsymbol{w}}_t)(\widehat{\boldsymbol{w}} - \widehat{\boldsymbol{w}}_t)$$

a pro získání maxima kvadratické funkce (tedy následující aproximace) se hledá jeho kořen

$$\begin{aligned} l_n'(\widehat{\boldsymbol{w}}_t) + l_n''(\widehat{\boldsymbol{w}}_t)(\widehat{\boldsymbol{w}}_{t+1} - \widehat{\boldsymbol{w}}_t) &= 0, \\ l_n''(\widehat{\boldsymbol{w}}_t)(\widehat{\boldsymbol{w}}_{t+1} - \widehat{\boldsymbol{w}}_t) &= -l_n'(\widehat{\boldsymbol{w}}_t), \\ \widehat{\boldsymbol{w}}_{t+1} &= \widehat{\boldsymbol{w}}_t - (l_n''(\widehat{\boldsymbol{w}}_t))^{-1} l_n'(\widehat{\boldsymbol{w}}_t). \end{aligned}$$

Dosadí-li se do tohoto vztahu dříve odvozené výsledky, je možné zformulovat následující algoritmus.

1. Zvol $\widehat{\boldsymbol{w}}_0$ a polož $t := 0$.
2. Vypočti pro $i = 1, \ldots, n$ výraz $\widehat{p}_t(\boldsymbol{x}_i)$

$$\widehat{p}_t(\boldsymbol{x}_i) := \frac{\exp(\widehat{\boldsymbol{w}}_t^T \boldsymbol{x}_i)}{1 + \exp(\widehat{\boldsymbol{w}}_t^T \boldsymbol{x}_i)}.$$

3. Vypočti výraz $\widehat{\boldsymbol{w}}_{t+1}$

$$\widehat{\boldsymbol{w}}_{t+1} := \widehat{\boldsymbol{w}}_t + \left(\boldsymbol{\mathcal{X}}^T \operatorname{diag}\{\widehat{p}_t(\boldsymbol{x}_i)(1-\widehat{p}_t(\boldsymbol{x}_i)),\ i=1,\ldots,n\}\boldsymbol{\mathcal{X}}\right)^{-1} \boldsymbol{\mathcal{X}}^T(\boldsymbol{\mathcal{Y}} - \operatorname{vect}\{\widehat{p}_t(\boldsymbol{x}_i),\ i=1,\ldots,n\}).$$

4. Jsou-li $\widehat{\boldsymbol{w}}_t$ a $\widehat{\boldsymbol{w}}_{t+1}$ dostatečně blízko, skonči. Jinak polož $t := t+1$ a pokračuj krokem 2.

Podmíněná pravděpodobnost

Konzistentní bodové odhady se dostanou přímo z (6). Pro odvození intervalových odhadů je možné použít delta metodu ([7], str. 313 - 315), podle které (přibližně) platí

$$\mathrm{var}(\widehat{p}(\boldsymbol{x})) \doteq \left(\frac{\partial p}{\partial \boldsymbol{w}}(\widehat{p}(\boldsymbol{x}))\right)^T \mathrm{var}(\widehat{\boldsymbol{w}}) \left(\frac{\partial p}{\partial \boldsymbol{w}}(\widehat{p}(\boldsymbol{x}))\right). \tag{9}$$

Vypočte-li se tedy derivace podmíněné pravděpodobnosti

$$\begin{aligned}\frac{\partial p}{\partial w_j}(\boldsymbol{x}, \boldsymbol{w}) &= \frac{\partial}{\partial w_j}\left(\frac{\exp(\boldsymbol{w}^T\boldsymbol{x})}{1+\exp(\boldsymbol{w}^T\boldsymbol{x})}\right) = \frac{\exp(\boldsymbol{w}^T\boldsymbol{x})}{(1+\exp(\boldsymbol{w}^T\boldsymbol{x}))^2}x_j \\ &= p(\boldsymbol{x}, \boldsymbol{w})(1-p(\boldsymbol{x}, \boldsymbol{w}))x_j,\end{aligned}$$

získá se rozptyl odhadu

$$\mathrm{var}(\widehat{p}(\boldsymbol{x})) \doteq (\widehat{p}(\boldsymbol{x}))^2(1-\widehat{p}(\boldsymbol{x}))^2\boldsymbol{x}^T\mathrm{var}(\widehat{\boldsymbol{w}})\,\boldsymbol{x}.$$

Alternativou k delta metodě je metoda bootstrapu (viz část 4.5).

Předpoklady metody

Klíčovým předpokladem metody je, že logaritmus poměru hustot je lineární funkcí pozorování $\boldsymbol{x}$. Platí totiž

$$\log\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} = \boldsymbol{w}^T\boldsymbol{x} \tag{10}$$

(tento zápis je ekvivalentní s (6)). Toto splňuje libovolné *rozdělení exponenciálního typu* s disperzním parametrem nezávislým na třídě. Jeho hustotu lze zapsat ve tvaru

$$p_l(\boldsymbol{x}) = \exp(\boldsymbol{\theta}_l^T\boldsymbol{x} - A(\boldsymbol{\theta}_l) + B(\boldsymbol{x})), \quad l = 0, 1$$

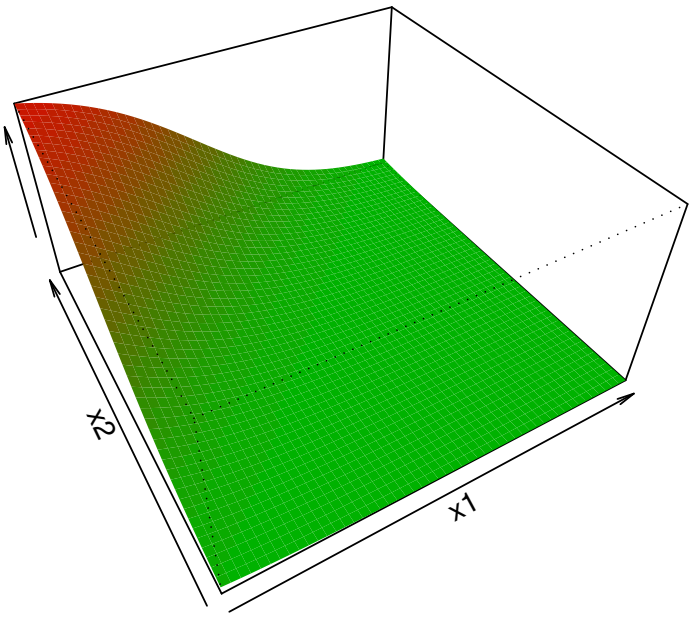
($\boldsymbol{\theta}_l$ je parametr určující konkrétní tvar distribuční funkce - první derivace funkce A v tomto bodě je střední hodnota a druhá derivace je rozptyl). Logaritmus poměru hustot je

$$\log\frac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})} = (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_0)^T\boldsymbol{x} - (A(\boldsymbol{\theta}_1) - A(\boldsymbol{\theta}_0)).$$

Do této rodiny patří mnoho rozdělení včetně normálního (se společnou varianční maticí), alternativního a Poissonova.

Tento předpoklad je velmi obecný (obrázek 8 ukazuje, jak se dá chápat geometricky) a zpravidla nebývá ověřován žádným testem. Vhodnější je zhodnotit kvalitu modelu.

Dále se předpokládá, že výběr je skutečně z $\mathcal{L}(\boldsymbol{X}, Y)$, to však není nijak omezující. Jak je popsáno v [12], str. 259 - 263, při výběru z $\mathcal{L}(\boldsymbol{X}|Y)$ stačí metodu mírně modifikovat.



Obrázek 8: Podmíněná pravděpodobnost odhadnutá metodou logistické regrese

Vztah k lineární diskriminační analýze

Vzorce (4) a (10) ukazují, že pro normální rozdělení dávají obě metody stejné výsledky. V [12], str. 276 - 279, se uvádí, že lineární diskriminační analýza je v případě normality asymptopicky eficientní, logistická regrese je zase robustnější. Zajímavější je však možnost zobecnění metody. Přestože to nebylo explicitně zmíněno, v případě logistické regrese trénovací množina zpravidla obsahuje sloupec jedniček a tím je umožněn odhad tzv. *absolutního členu*, který se stává součástí prahu. Metodu lze podobným způsobem zobecnit přidáním dalších regresorů (interakcí, vyšších mocnin), což u lineární diskriminační analýzy příliš obvyklé není. U ní je zase možné přejít k obecnějšímu tvaru varianční matice.

Příbuzné metody

Logistická regrese transformuje lineární kombinaci vstupů logistickou funkcí. Předpokládá se tedy

$$p(\boldsymbol{x}) = \frac{\exp(\boldsymbol{w}^T\boldsymbol{x})}{1 - \exp(\boldsymbol{w}^T\boldsymbol{x})}.$$

Probitový model ke stejnému účelu využívá distribuční funkci normálního rozdělení (značenou ϕ). Očekává, že platí

$$p(\boldsymbol{x}) = \phi(\boldsymbol{w}^T\boldsymbol{x}).$$

Log-log model předpokládá

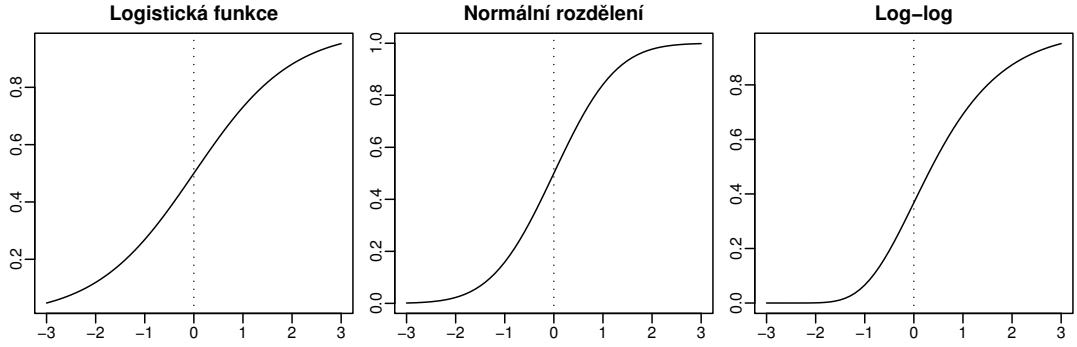
$$p(\boldsymbol{x}) = \exp(-\exp(-\boldsymbol{w}^T\boldsymbol{x})).$$

Logistická regrese je nejlépe interpretovatelná. Porovná-li se poměr hustot pro různé vstupy ($\boldsymbol{e}_j$ je j-tý člen báze prostoru $\mathbb{R}^p$)

$$\frac{\dfrac{p_1(\boldsymbol{x}+\boldsymbol{e}_j)}{p_0(\boldsymbol{x}+\boldsymbol{e}_j)}}{\dfrac{p_1(\boldsymbol{x})}{p_0(\boldsymbol{x})}} = \frac{\exp(\boldsymbol{w}^T(\boldsymbol{x}+\boldsymbol{e}_j))}{\exp(\boldsymbol{w}^T\boldsymbol{x})} = \exp(w_j),$$

zjišťuje se, že exponenciála j-té složky vektoru vah vyjadřuje, kolikrát se zvětší poměr pravděpodobnosti defaultu a pravděpodobnosti splácení úvěru při jednotkové změně j-tého vstupu.

V kontextu zobecněných lineárních modelů (jejichž speciálními případy všechny tyto metody jsou) se navíc ukazuje, že pouze u logistické regrese má věrohodnostní rovnice zaručené jednoznačné řešení. Bez speciálního důvodu je proto nejlepší použít ji. Jak ukazuje obrázek 9, výhodou log-logu může být jeho nesymetričnost.



Obrázek 9: Různé funkce podobné logistické.

3.3 Neuronové sítě

V případě neuronových sítí se předpokládá, že podmíněnou pravděpodobnost lze vyjádřit ve tvaru

$$p(\boldsymbol{x}) = f_2\left(w_{h_0}^{out} + \sum_{k=1}^{K} w_{h_k}^{out}\, f_1\left(w_{in_0}^{h_k} + \sum_{j=1}^{p} w_{in_j}^{h_k} x_j\right)\right), \tag{11}$$

kde f_1 a f_2 značí logistickou funkci

$$f_1(x) = f_2(x) = \frac{\exp(x)}{1+\exp(x)}$$

a $\boldsymbol{w}$ (neznámé) váhy.

Funkce f_1 ztransformuje různé lineární kombinace vstupů x_j a tím vytvoří nové vstupy (tzv. *skryté uzly*) ve skryté vrstvě. Ty se opět zkombinují a prostřednictvím funkce f_2 ztransformují. Tyto tři vrstvy (vstupní, skrytá a výstupní) společně dokáží aproximovat i nelineární zobrazení.

Chybové funkce

Podobně jako u ostatních metod se i v případě neuronových sítí využívá metoda maximální věrohodnosti, respektive často ekvivalentní minimalizace chybové funkce. Jednou z možností je součet čtverců

$$E_1 = \sum_{i=1}^{n} (y_i - p(\boldsymbol{x}_i))^2,$$

který však znamená hledání maximálně věrohodného odhadu pro případ normálního rozdělení (jak je ukázáno v úvodu této části), což pro klasifikaci nemusí být vhodné. Náhodná veličina Y totiž nabývá pouze hodnot 0 a 1, a tak je více opodstatněné vyjít (podobně jako u logistické regrese) z předpokladu

$$\mathcal{L}(Y_i|\boldsymbol{X}_i) = \text{alt}(p(\boldsymbol{X}_i)).$$

Sdružená hustota je

$$\prod_{i=1}^{n} p(\boldsymbol{x}_i)^{y_i}(1 - p(\boldsymbol{x}_i))^{1-y_i}.$$

Po zlogaritmování se dostává logaritmická věrohodnost

$$\sum_{i=1}^{n} \Big(y_i \log(p(\boldsymbol{x}_i)) + (1 - y_i) \log(1 - p(\boldsymbol{x}_i)) \Big)$$

a druhá chybová funkce (nazývaná *entropie*)

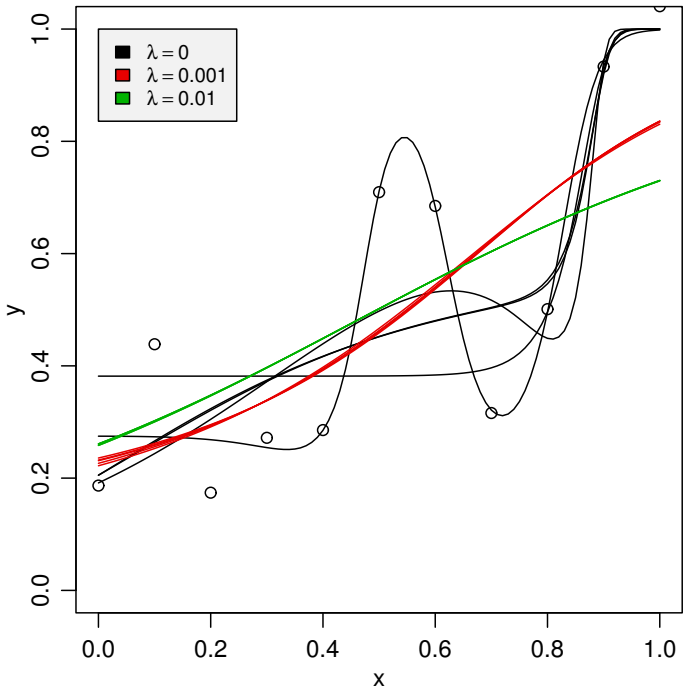
$$E_2 = -\sum_{i=1}^{n} \Big(y_i \log(p(\boldsymbol{x}_i)) + (1 - y_i) \log(1 - p(\boldsymbol{x}_i)) \Big). \tag{12}$$

Výrazné vylepšení metody představuje přidání *regularizačního parametru* λ. Jeho účelem je omezit velikost vah. Logistická funkce totiž dobře rozlišuje pouze hodnoty v blízkosti nuly, v případě příliš velkých parametrů výstup téměř nezávisí na vstupech (viz obrázek 9). Zároveň se takto zamezuje jevu označovanému jako *overfitting* - situaci, kdy se neuronová síť příliš přizpůsobí trénovací množině. To ukazuje obrázek 10. Budou-li vstupy srovnatelně velké (toho lze dosáhnout standardizací), lze tedy zavést další chybové funkce

$$E_3 = E_1 + \lambda \, ||\boldsymbol{w}||^2$$

a

$$E_4 = E_2 + \lambda \, ||\boldsymbol{w}||^2.$$



Obrázek 10: Spojitá vysvětlovaná proměnná y odhadovaná neuronovou sítí - s různými hodnotami parametru λ, vždy pětkrát.

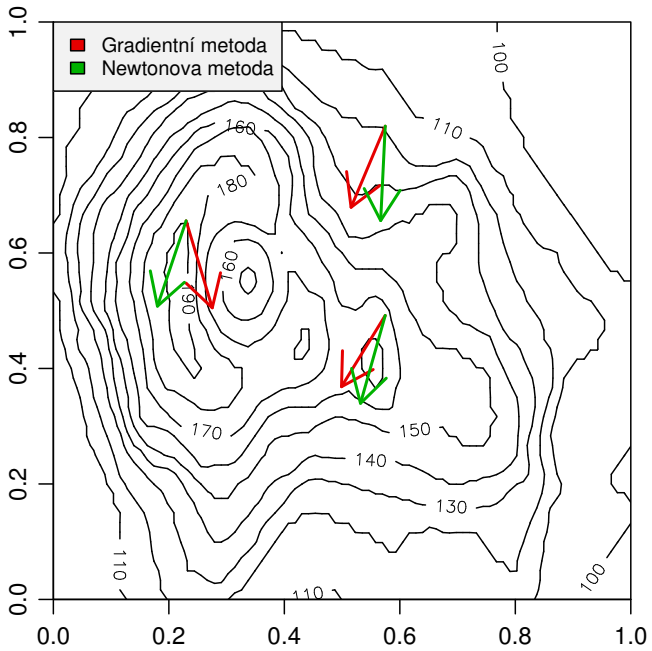
Hledání optimálních vah

Optimální váhy se hledají iterativně. Náhodně se zvolí počáteční řešení, které se v dalších krocích upravuje tak, aby se snižovala hodnota chybové funkce. Jak je uvedeno u obrázku 11, většina minimalizačních algoritmů očekává znalost první, případně i druhé derivace chybové funkce. V této práci byla při implementaci použita funkce [15], `nnet`, a jí využívaný algoritmus BFGS (popsaný v [16], str. 344 - 345, nebo v [5], str. 287 - 289), založený na Newtonově metodě, vyžadující však znalost pouze první derivace. Jednou z příčin úspěšnosti neuronových sítí je právě jednoduchý způsob výpočtu derivace chybové funkce a ten bude nyní popsán.

Nejdříve je zobecněno značení. Vstupy každé vrstvy jsou lineární kombinací výstupů vrstvy předchozí, výstupy dané vrstvy se získají transformací vstupů. Tedy skutečné vstupy se *identickou funkcí* ztransformují na výstupy vstupní vrstvy. Jejich lineární kombinace jsou pak vstupy vrstvy skryté a logistickou transformací se získají výstupy této vrstvy. Podobně se postup opakuje u vrstvy výstupní. Vstupy jsou značeny z_j, výstupy $\widetilde{z}_j$, transformační funkce g_j, váhy w_i^j. Tedy

$$\widetilde{z}_j = g_j(z_j), \quad z_j = \sum_{i \to j} w_i^j \widetilde{z}_i.$$

Symbol $i \to j$ znamená sčítání přes všechny vhodné kombinace.



Obrázek 11: Gradientní metoda postupuje "do kopce" - ve směru první derivace. Newtonova metoda jde "přímo na vrchol" - do středu elipsoidu vytvořeného pomocí prvních dvou derivací.

Hledají se derivace podle jednotlivých vah ($\partial E/\partial w_i^j$). Ty lze formálně rozepsat

$$\frac{\partial E}{\partial w_i^j} = \frac{\partial E}{\partial z_j}\frac{\partial z_j}{\partial w_i^j} = \frac{\partial E}{\partial z_j}\widetilde{z}_i.$$

Konkrétněji (pro jednotlivé váhy ze vzorce (11))

$$\frac{\partial E}{\partial w_{h_k}^{out}} = \frac{\partial E}{\partial z_{out}}\widetilde{z}_{h_k}$$

a

$$\frac{\partial E}{\partial w_{in_j}^{h_k}} = \frac{\partial E}{\partial z_{h_k}}\widetilde{z}_{in_j} = \frac{\partial E}{\partial z_{out}}\frac{\partial z_{out}}{\partial \widetilde{z}_{h_k}}\frac{\partial \widetilde{z}_{h_k}}{\partial z_{h_k}}\widetilde{z}_{in_j} = \frac{\partial E}{\partial z_{out}}w_{h_k}^{out}\frac{\partial g_{h_k}}{\partial z_{h_k}}\widetilde{z}_{in_j}. \tag{13}$$

Díky (13) se tento algoritmus nazývá *back-propagation*. Je-li totiž známa derivace chybové funkce podle vstupu výstupní vrstvy, určí se z ní derivace podle vstupů vrstvy skryté. Navíc je ještě nutné nalézt derivace transformačních funkcí. Platí

$$\begin{aligned}\frac{\partial}{\partial x}g(x) &= \frac{\partial}{\partial x}\left(\frac{\exp(x)}{1+\exp(x)}\right) = \frac{\exp(x)(1+\exp(x)) - \exp(x)\exp(x)}{(1+\exp(x))^2} \\ &= \frac{\exp(x)}{1+\exp(x)}\,\frac{1}{1+\exp(x)} = g(x)(1-g(x)).\end{aligned} \tag{14}$$

Poslední věc, kterou je třeba odvodit, je derivace podle vstupu výstupní vrstvy. Zde již záleží na konkrétní chybové funkci. Pro první platí

$$\begin{aligned}\frac{\partial E_1}{\partial z_{out}} &= \frac{\partial}{\partial z_{out}}\left(\sum_{i=1}^{n}(y_i - p(\boldsymbol{x}_i))^2\right) = -2\sum_{i=1}^{n}(y_i - p(\boldsymbol{x}_i))\frac{\partial}{\partial z_{out}}p(\boldsymbol{x}_i) \\ &= 2\sum_{i=1}^{n} p(\boldsymbol{x}_i)(1 - p(\boldsymbol{x}_i)(p(\boldsymbol{x}_i) - y_i) \qquad (15)\end{aligned}$$

(poslední úprava vychází z (14)), pro druhou pak

$$\begin{aligned}\frac{\partial E_2}{\partial z_{out}} &= \frac{\partial}{\partial z_{out}}\left(-\sum_{i=1}^{n}\Big(y_i \log(p(\boldsymbol{x}_i)) + (1 - y_i)\log(1 - p(\boldsymbol{x}_i))\Big)\right) \\ &= -\sum_{i=1}^{n}\Big(\frac{y_i}{p(\boldsymbol{x}_i)}\frac{\partial}{\partial z_{out}}p(\boldsymbol{x}_i) - \frac{1 - y_i}{1 - p(\boldsymbol{x}_i)}\frac{\partial}{\partial z_{out}}p(\boldsymbol{x}_i)\Big) \\ &= \sum_{i=1}^{n}\Big(-\frac{y_i}{p(\boldsymbol{x}_i)}p(\boldsymbol{x}_i)(1 - p(\boldsymbol{x}_i)) + \frac{1 - y_i}{1 - p(\boldsymbol{x}_i)}p(\boldsymbol{x}_i)(1 - p(\boldsymbol{x}_i))\Big) \\ &= \sum_{i=1}^{n}\Big(-y_i + y_i p(\boldsymbol{x}_i) + p(\boldsymbol{x}_i) - y_i p(\boldsymbol{x}_i)\Big) = \sum_{i=1}^{n}(p(\boldsymbol{x}_i) - y_i). \qquad (16)\end{aligned}$$

Protože $w_{in_0}^{h_k}$ a $w_{h_0}^{out}$ ze vzorce (11) lze považovat za váhy vstupu 1 (jak je to obvyklé u regrese), jsou již odvozeny všechny vztahy potřebné k výpočtu $\partial E_1/\partial w_i^j$ a $\partial E_2/\partial w_i^j$. U chybových funkcí E_3 a E_4 jsou pouze přidány násobky druhých mocnin všech parametrů, to v případě derivace znamená přičtení výrazu $2\lambda w_i^j$.

Nyní už je možné zformulovat algoritmus pro hledání optimálních parametrů.

1. Náhodně zvol počáteční hodnoty vah.

2. Na základě vstupů a aktuálních vah spočítej vstupy a výstupy všech vrstev, porovnáním spočtených a zadaných výstupů dále urči hodnotu chybové funkce.

3. Je-li chyba dostatečně "malá" (od posledního kroku se téměř nesnížila), skonči. Jinak spočítej hodnoty derivace chybové funkce podle aktuálních vah a jejich použitím v nějaké optimalizační metodě urči nové váhy. Pokračuj bodem 2.

Další problémy při konstrukci neuronové sítě

Dříve než se začnou hledat optimální váhy, je nutné určit konkrétní hodnoty dvou parametrů - počet skrytých uzlů K a regularizační parametr λ. Jednou z možností je využít tzv. *bayesovský přístup* ([5], str. 385 - 439), který představuje alternativu k metodě maximální věrohodnosti a umožňuje srovnávat různé modely na základě pouze trénovací množiny. V této práci byla při implementaci použita metoda bootstrapu (viz část 4.5).

Ať už se modely porovnávají jakkoliv, je nutné vědět, jaké hodnoty vyzkoušet. Podle [16], str. 163 - 164, by se λ mělo pohybovat mezi 0.001 - 0.1 (jsou-li všechny vstupy znormovány do intervalu $[0, 1]$). Podobné doporučení pro počet skrytých uzlů se v [16] ani v [5] nenachází, ale ve většině příkladů jich bylo mezi 2 a 10.

Další zajímavou variantou je úprava algoritmu hledajícího optimální váhy. Uvedená *dávková verze* využívala v každé iteraci všechny prvky trénovací množiny. Místo toho je možné použít v každém kroku pouze jedno pozorování (*online verze*) - v (15) a (16) se tedy nebude sčítat.

Podmíněná pravděpodobnost

Bodový odhad se získá přímým dosazením nalezených vah do vzorce (11). Pro intervalové odhady se podle [6] dají použít tři různé postupy - delta metoda, metoda bootstrapu (viz část 4.5) a bayesovský přístup.

Delta metody vychází, podobně jako v případě logistické regrese, ze vzorce (9). Stejným způsobem se vypočte i rozptyl vah - použije se inverze druhé derivace chybové funkce. Ta se určí myšlenkově podobným způsobem jako derivace první (technicky ale mnohem náročněji - viz [16], 151 - 153). Vzorce se však téměř nemusí upravovat při výpočtu derivace podmíněné pravděpodobnosti - jen místo derivace chybové funkce podle vstupu výstupní vrstvy se dosadí $p(\boldsymbol{x})(1 - p(\boldsymbol{x}))$ (derivace podmíněné pravděpodobnosti podle vstupu výstupní vrstvy).

Bayesovské intervaly jsou založeny na přístupu odlišném od metody maximální věrohodnosti. Bayesovský přístup nevychází pouze z trénovací množiny, ale využívá ji společně s předpokládanou distribucí apriorních pravděpodobností k odvození distribuce aposteriorních pravděpodobností. Tím se automaticky získávají odhady nejen bodové, ale i intervalové. Podle [6] se však tyto distribuce musí aproximovat a odvozené konfidenční intervaly jsou proto příliš nepřesné.

Předpoklady metody

Jak bude nyní ukázáno, neuronové sítě (alespoň teoreticky) nepředpokládají téměř nic. Neuronové sítě s konečným počtem skrytých uzlů totiž mají schopnost stejnoměrně aproximovat (na kompaktní množině) libovolnou spojitou funkci. Určitě se ale nejedná o formální důkaz, ten lze nalézt např. v [16], str. 173 - 176.

Vychází se z věty známé z teorie Fourierových řad, která říká, že každou spojitou funkci f, zobrazující interval $[0, \pi]$ do $\mathbb{R}$, lze stejnoměrně aproximovat trigonometrickým polynomem

$$T(x) = \sum_{k=0}^{K} a_k \cos(kx), \quad a_k \in \mathbb{R}, K \in \mathbb{N},$$

tak, že pro libovolné kladné (předem zvolené) ε platí

$$|T(x) - f(x)| < \varepsilon, \quad \forall x \in [0, \pi].$$

1. Každou spojitou funkci $f : [0, \pi] \to \mathbb{R}$ je možné stejnoměrně aproximovat trigonometrickým polynomem obsahujícím konečně mnoho členů.

2. Každou spojitou funkci $f : [0, \pi]^p \to \mathbb{R}$ lze postupně rozepsat

$$\begin{aligned} f(x_1, x_2, \ldots, x_p) &= \sum_{k_1=1}^{K_1} f_1(x_2, x_3, \ldots, x_p) \cos(k_1 x_1) \\ &= \sum_{k_1=1}^{K_1} \sum_{k_2=1}^{K_2} f_2(x_3, x_4, \ldots, x_p) \cos(k_1 x_1) \cos(k_2 x_2) \\ &= \cdots \end{aligned}$$

   na součet konečného počtu součinů ve tvaru

$$\prod_{j=1}^{p} \cos(k_j x_j).$$

3. Každý z těchto součinů může být opakovaným použitím vzorce

$$\cos(x) \cos(y) = \frac{\cos(x + y) + \cos(x - y)}{2}$$

   rozepsán na součet konečného počtu členů ve tvaru

$$\cos\left(\sum_{j=1}^{p} \widetilde{k}_j x_j \right).$$

4. Funkci $\cos(x)$ lze stejnoměrně aproximovat schodovitou funkcí.

5. Schodovitou funkci je možné stejnoměrně aproximovat součtem konečného počtu logistických funkcí.

6. Každou spojitou funkci $f : M \to \mathbb{R}$, kde M je kompaktní podmnožina $\mathbb{R}^p$, lze spojitě rozšířit na nějaký (mnohorozměrný) interval. Funkci f je dále možné rozložit na funkci provádějící lineární transformaci tohoto intervalu na interval $[0, \pi]^p$ a na jinou funkci $\widetilde{f}$ tam definovanou. Použitím předchozích kroků pak lze funkci $\widetilde{f}$ a tedy i funkci f stejnoměrně aproximovat součtem konečného počtu logistických funkcí.

7. Logistická funkce ve výstupní vrstvě zobrazuje $\mathbb{R}$ na interval $(0, 1)$. To znamená, že libovolnou funkci ve tvaru $f : M \to (0, 1)$, kde M je kompaktní podmnožina $\mathbb{R}^p$, lze stejnoměrně aproximovat neuronovou sítí.

Tento postup nabízí teoretické zdůvodnění funkčnosti neuronových sítí. Neříká však, jakého množství skrytých uzlů je potřeba, ani jaké vlastnosti má neuronová síť s nižším počtem skrytých uzlů.

Vztah k logistické regresi

Neuronové sítě lze považovat za přirozené rozšíření logistické regrese. Nebo naopak - logistická regrese je speciálním případem neuronových sítí. Jak je uvedeno v dalším odstavci, neuronové sítě nemusí mít skrytou vrstvu zrovna jednu. Logistická regrese je neuronovou sítí bez skryté vrstvy. Ale právě opakované využití logistické funkce dává neuronovým sítím možnost aproximovat i taková zobrazení, pro která lineární metody vhodné nejsou.

Příbuzné metody

Dosavadní výklad vycházel z určitého zjednodušení. Zabýval se pouze jednou z mnoha možných *architektur* neuronových sítí. Za metody příbuzné lze tedy pokládat takové varianty, které využívají jiný počet skrytých vrstev, spojení přeskakující vrstvy (vstupem výstupní vrstvy je pak lineární kombinace obsahující navíc skutečné vstupy - jednotlivé znaky x_j), nebo jinou transformační funkci.

Je-li touto funkcí ve výstupní vrstvě identita, pak neuronová síť aproximuje libovolný výstup (což se nehodí ke klasifikaci, ale jinak je tato architektura stejně tak důležitá jako ta, kterou se práce zabývá). Ve skryté vrstvě může být zase jako transformační funkce použit *hyperbolický tangens*

$$f(x) = \tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)},$$

který podle [5], str. 127, nezlepšuje aproximační schopnosti metody, ale často zrychluje konvergenci. Spíše historickou možností je prahová funkce

$$f(x) = I_{\{x>0\}},$$

jejíž nevýhodou je nespojitá derivace.

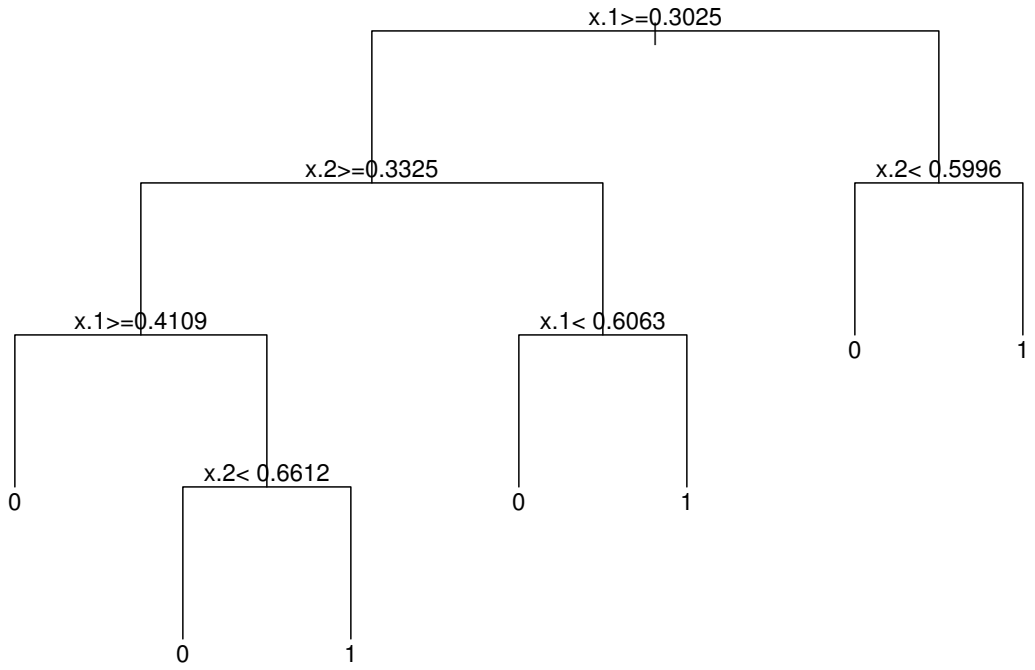
3.4 Jiné metody

Nyní bude stručně zmíněno několik dalších metod využitelných pro klasifikaci.

Stromy

Jak už napovídá název, tato metoda vytváří klasifikační pravidla mající atraktivní podobu stromu (viz obrázek 12). Podle hodnoty některé proměnné se prostor všech možných pozorování dělí na dvě části. Opakováním tohoto postupu vznikne velké množství podmnožin - listů vzniklého stromu. Pro každý list se odhadne podmíněná pravděpodobnost a v souladu s pravidlem (3) se klasifikuje.

Výhodou je přirozený přístup k diskrétním proměnným. Při konstrukci stromu nejsou převáděny na spojité, ale jsou použity přímo.



Obrázek 12: Klasifikace pomocí jednoduchého stromu

Kritériem pro hodnocení kvality stromu je (vážený) *index nečistoty.* Ten lze (pro jednotlivé listy) určit jako ztrátu vzniklou chybnou klasifikací, tedy

$$\min\{c_0\widehat{\pi}_0, c_1\widehat{\pi}_1\},$$

kde $\widehat{\pi}_l$ jsou odhady vypočtené z prvků trénovací množiny, které spadají pod daný list.

Vlastní konstrukce stromu může probíhat tím způsobem, že se listy přidávají tak dlouho, dokud index nečistoty klesá dostatečně rychle. Jinou variantou je nejdříve vytvořit obrovský strom a ten "prořezat" - odstranit ty části, které index nečistoty dostatečně nesnižují.

Další podrobnosti lze nalézt v [9], str. 61 - 78.

Jádrová metoda a metoda nejbližších sousedů

Jak jádrová metoda, tak i metoda nejbližších sousedů vytváří "lokální modely", při odhadech pravděpodobností pro určitý bod $\mathbb{R}^p$ kladoucí důraz na takové prvky trénovací množiny, které jsou v určitém smyslu blízko.

V případě jádrové metody se odhadují podmíněné hustoty $f(\boldsymbol{x}|y)$ a z těch se (podle Bayesovy věty) vypočtou aposteriorní pravděpodobnosti. K odhadu se využívá jádrová funkce (kernel function) K, na jejíž hodnotu mají vliv především blízká pozorování (z trénovací množiny). Odhad podmíněných hustot je

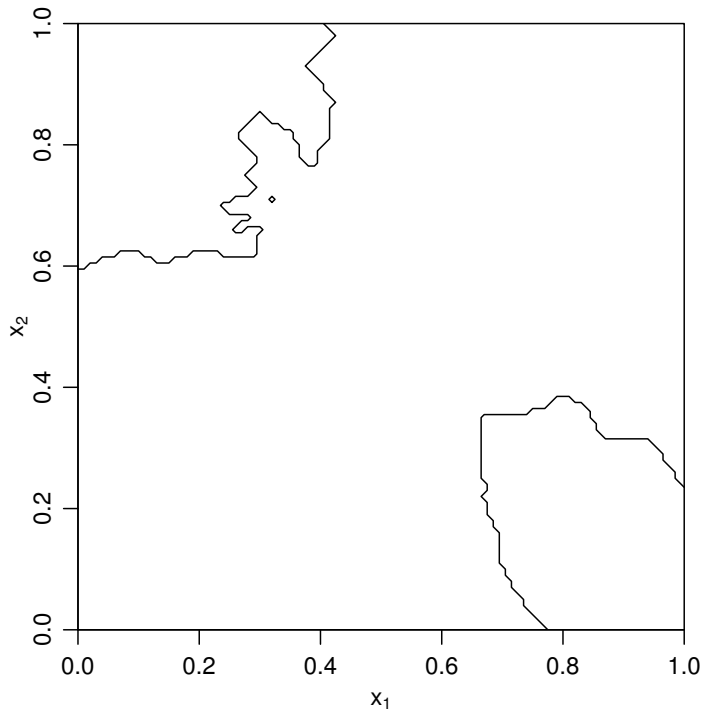
$$\widehat{f}(\boldsymbol{x}|y=l) = \frac{1}{n_l}\sum_{y_i=l} K(\boldsymbol{x}, \boldsymbol{x}_i), \quad l = 0, 1,$$

kde n_l je počet pozorování v l-té třídě. Častou volbou je Gaussovské jádro s odhady

$$\widehat{f}(\boldsymbol{x} \,|\, y = l) = \frac{1}{n_l (2\pi)^{p/2} |\boldsymbol{H}|^{1/2}} \sum_{y_i = l} \exp\left(-\frac{(\boldsymbol{x} - \boldsymbol{x}_i)\boldsymbol{H}^{-1}(\boldsymbol{x} - \boldsymbol{x}_i)}{2} \right), \quad l = 0, 1$$

(je však ještě třeba určit matici $\boldsymbol{H}$).

Pomocí metody nejbližších sousedů se aposteriorní pravděpodobnosti odhadují přímo, a to jako průměrná odezva několika nejbližších pozorování. Nejdříve se však musí stanovit, kolik těchto sousedů má být a jakým způsobem se měří vzdálenost. Ukázka klasifikace provedené touto metodou je na obrázku 13.



Obrázek 13: Klasifikace metodou (tří) nejbližších sousedů, s předpokladem rovných ztrát z chybné klasifikace ($c_0 = c_1$)

Další podrobnosti lze nalézt v [9], str. 79 - 93.

Loglineární modely

Loglineární modely vycházejí z předpokladu, který je při odhadování pravděpodobnosti defaultu velmi zajímavý. Je to předpoklad Poissonova rozdělení počtu defaultů - tedy takového rozdělení, které vznikne jako počet úspěchů při velkém množství alternativních experimentů s malou pravděpodobností úspěchu.

Kvůli tomuto předpokladu metoda očekává diskrétní proměnné. Jinak by těžko pro jeden prvek množiny všech možných pozorování mohl nastat více než jeden default. Proto je nutné rozdělit spojité proměnné na intervaly a tím je přeměnit na proměnné ordinální.

Odhadovaná pravděpodobnost (počtu defaultů) má tvar

$$p(y_i = k) = \frac{\lambda_i^k}{k!} \exp(-\lambda_i), \quad k = 0, 1, \ldots$$

(neindexuje se přes jednotlivé prvky trénovací množiny, ale přes skupiny prvků se stejným vstupem), kde

$$\lambda_i = \exp(\boldsymbol{w}^T \boldsymbol{x}_i).$$

Loglineární modely patří mezi zobecněné lineární modely. Hledání optimálních vah tedy probíhá podobně jako v případě logistické regrese.

Další podrobnosti lze nalézt v [1].

4 Ratingové modely a jejich validace

Tato část se zabývá tvorbou a validací ratingových modelů. Tvorba ratingového modelu spočívá v odvození ratingové funkce ρ (funkce přiřazující pozorováním rating, zavedená v části 2) a stanovení předpokládaných pravděpodobností defaultu pro jednotlivé ratingové třídy. Důležitou součástí ratingové funkce je odhadování pravděpodobnosti defaultu, je proto rozumné ověřit účinnost diskriminace pomocí některého testu specifického pro zvolenou klasifikační a skóringovou metodu.

Dále je vhodné výsledný ratingový model zkalibrovat (tedy ověřit, zda jsou předpokládané pravděpodobnosti defaultu pro jednotlivé ratingové třídy v souladu se skutečností) a zhodnotit (dvěma různými přístupy - jeden je založený na diskriminační schopnosti modelu, druhý na informaci poskytované modelem). K oběma účelům se využívá validační množina.

V této části je také zmíněna metoda bootstrapu, pomocí které lze jednoduchým způsobem konstruovat intervalové odhady a která umožňuje hodnotit model bez použití validační množiny.

4.1 Testy specifické pro jednotlivé metody

Platí-li předpoklady jednotlivých metod, lze využít některé testy hodnotící účinnost diskriminace.

Lineární diskriminační analýza

U lineární diskriminační analýzy mají váhy tvar

$$\boldsymbol{w} = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0),$$

pro účinnost diskriminace je tedy nutné, aby se odhady středních hodnot významně lišily. Pro testování této skutečnosti lze zavést statistiku udávající vzdálenost středních hodnot v prostoru určeném variabilitou dat

$$D_p^2 = (\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0)^T \widehat{\boldsymbol{\Sigma}}^{-1}(\widehat{\boldsymbol{\mu}}_1 - \widehat{\boldsymbol{\mu}}_0).$$

Za předpokladu, že $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_1$, platí

$$\frac{(n-p-1)\, n_0\, n_1}{p\, n^2} D_p^2 \sim F(p, n-p-1).$$

Logistická regrese

U logistické regrese by trénovací množina měla obsahovat sloupec samých jedniček. Při hodnocení účinnosti diskriminace se srovnává sdružená hustota výběru $(\boldsymbol{\mathcal{X}}, \boldsymbol{\mathcal{Y}})$ se sdruženou hustotou výběru obsahující pouze sloupec samých jedniček. Není-li diskriminace účinná,

pak je model vytvořený pouze na základě druhého výběru (tzv. *nulový model*) stejně dobrý jako model hodnocený.

Je-li $L_n(\widehat{\boldsymbol{w}})$ sdružená hustota hodnoceného modelu a je-li $L_n(\widehat{w}_0)$ sdružená hustota nulového modelu (v obou případech pro maximálně věrohodný odhad vah), pak asymptoticky platí

$$2\log\frac{L_n(\widehat{\boldsymbol{w}})}{L_n(\widehat{w}_0)} \sim \chi^2_{p-1}$$

(p je počet sloupců trénovací množiny obsahující sloupec samých jedniček).

Neuronové sítě

Literatura se o žádném testu nezmiňuje. Důvodem bude nejspíše přílišná obecnost předpokladů metody.

4.2 Tvorba ratingového modelu

Ratingová funkce ρ zobrazuje $\mathbb{R}^p$ (množinu všech možných pozorování) do $\{1, 2, \ldots, R\}$ (množiny ratingů). Předpokládá se, že s vyšším ratingem pravděpodobnost defaultu klesá (závěry by samozřejmě platily i v opačném případě).

Složkou ratingové funkce je klasifikační a skóringová metoda, která jednotlivým pozorováním přiřazuje pravděpodobnost defaultu. Zároveň může být stanoveno i skóre - tyto dvě veličiny jsou pak provázány monotónním zobrazením. Ratingovou funkci tedy lze chápat třemi různými způsoby - jako pravidlo rozdělující pravděpodobnosti ($[0, 1]$), skóre ($\mathbb{R}$) nebo možná pozorování ($\mathbb{R}^p$) na R disjunktních podmnožin. Zařazení do ratingové třídy se uskuteční tak, že se odhadne pravděpodobnost defaultu (nebo skóre) a porovnáním s hranicemi jednotlivých pásem se určí rating.

Počet ratingů lze volit libovolně. V případě regulatorních modelů musí existovat (podle [4], str. 84) alespoň sedm ratingů pro splácené úvěry a alespoň jeden pro defaulty. O dalších podrobnostech již banky rozhodují samy. Podle [14], str. 84, většinou mapují výstupy všech modelů na jednu stupnici. Jinou variantou je vytvořit pásma tak, aby všechny ratingové třídy byly zastoupeny v trénovací množině přibližně rovnoměrně.

Dále je potřeba určit předpokládané pravděpodobnosti defaultu pro jednotlivé ratingy, značené q_r. Je možné za ně dosadit průměrnou odezvu

$$q_r = \frac{1}{n_r}\sum_{\rho(\boldsymbol{x}_i)=r} y_i, \quad r = 1, 2, \ldots, R$$

nebo průměrnou podmíněnou pravděpodobnost

$$q_r = \frac{1}{n_r}\sum_{\rho(\boldsymbol{x}_i)=r} p(\boldsymbol{x}_i), \quad r = 1, 2, \ldots, R$$

(v obou případech je n_r počet pozorování, kterým byl přiřazen r-tý rating).

Kalibrace

Jedním z účelů validační množiny je odhalit, zda předpokládané pravděpodobnosti defaultu odpovídají skutečnosti. Není-li tomu tak, je třeba na situaci reagovat. A to buď úpravou ratingové funkce nebo změnou předpokládaných pravděpodobností defaultu.

Shodu předpokládané (určené množinou trénovací) a pozorované (určené množinou validační) pravděpodobnosti defaultu lze ověřit pomocí statistických testů. Zde jsou popsány dva z nich. Oba předpokládají nekorelovanost defaultů (ve validační množině). V [14], str. 119 - 124, se uvádí, že tomu tak většinou není, a zmiňuje se upravený test, který korelaci v úvahu bere. Nekorelovanost nepředpokládá ani *bootstrapový test* (použitý v této práci při implementaci).

Je-li ratingová funkce aplikována na $\widetilde{n}$-prvkovou validační množinu, získá se R tříd po $\widetilde{n}_r$ prvcích. Průměrná odezva (pozorovaná pravděpodobnost defaultu) je

$$\widetilde{q}_r = \frac{1}{\widetilde{n}_r} \sum_{\rho(\widetilde{\boldsymbol{x}}_i)=r} \widetilde{y}_i, \quad r = 1, 2, \ldots, R,$$

podobně je $\widetilde{q}$ průměrná odezva celé validační množiny. Pro oba testy se předpokládá, že odezvy y pozorování s ratingem r jsou náhodným výběrem z alternativního rozdělení se střední hodnotou q_r. Jejich průměr $\widetilde{q}_r$ má tedy střední hodnotu q_r a rozptyl $q_r(1-q_r)/\widetilde{n}_r$.

Podle centrální limitní věty asymptoticky platí

$$\widetilde{q}_r \sim N\left(q_r, \frac{q_r(1-q_r)}{\widetilde{n}_r}\right), \quad r = 1, 2, \ldots, R,$$

což lze využít v *testu založeném na normalitě.* Podmínkou je dostatečné množství dat. Zvláště u vyšších ratingových tříd (s málo defaulty) může být tento test zavádějící.

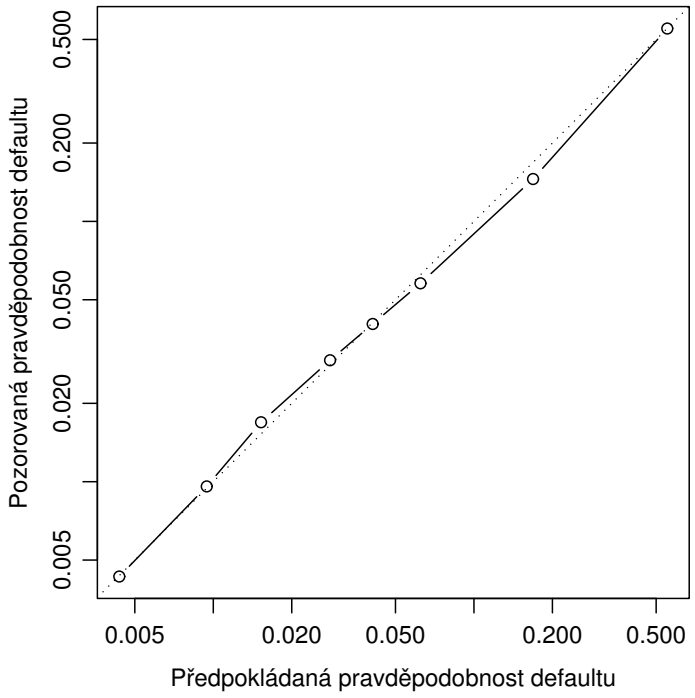
Dnes už však není žádným problémem použít binomické rozdělení přímo - ke konstrukci *binomického testu.* Např. platí-li

$$\sum_{k=0}^{\widetilde{n}_r\widetilde{q}_r} \binom{\widetilde{n}_r}{k} q_r^k (1-q_r)^{\widetilde{n}_r - k} > Q,$$

je předpokládaná pravděpodobnost defaultu ve třídě r podhodnocena na hladině $1-Q$.

Diagram spolehlivosti

Diagram spolehlivosti je grafickou metodou umožňující rychle zkontrolovat, zda předpovědi pravděpodobnosti defaultu platí (viz obrázek 14). Na vodorovné ose je pravděpodobnost předpokládaná, na svislé pak pozorovaná. V ideálním případě by všechny ratingy měly být na diagonále.



Obrázek 14: Diagram spolehlivosti

Brierovo skóre

Statistikou hodnotící spolehlivost předpovědi je *Brierovo skóre*. V základním tvaru jde o střední kvadratickou odchylku předpokládané pravděpodobnosti defaultu od skutečnosti

$$\mathrm{BS} = \frac{1}{\widetilde{n}} \sum_{i=1}^{\widetilde{n}} (\widetilde{y}_i - q_{\rho(\boldsymbol{x}_i)})^2,$$

po třídách

$$\mathrm{BS} = \sum_{r=1}^{R} \frac{\widetilde{n}_r}{\widetilde{n}} \left(\widetilde{q}_r (1 - q_r)^2 + (1 - \widetilde{q}_r) q_r^2 \right).$$

Brierovo skóre lze rozepsat na tvar

$$\mathrm{BS} = \sum_{r=1}^{R} \frac{\widetilde{n}_r}{\widetilde{n}} (\widetilde{q}_r (1 - \widetilde{q}_r)) + \sum_{r=1}^{R} \frac{\widetilde{n}_r}{\widetilde{n}} (q_r - \widetilde{q}_r)^2,$$

kde první člen shrnuje vnitroskupinovou variabilitu a druhý spolehlivost předpovědi. Jiná podoba

$$\mathrm{BS} = \widetilde{q}(1 - \widetilde{q}) + \sum_{r=1}^{R} \frac{\widetilde{n}_r}{\widetilde{n}} (q_r - \widetilde{q}_r)^2 - \sum_{r=1}^{R} \frac{\widetilde{n}_r}{\widetilde{n}} (\widetilde{q} - \widetilde{q}_r)^2$$

dále rozděluje vnitroskupinovou variabilitu na variabilitu validační množiny (nezávisející na ratingovém modelu) a na variabilitu meziskupinovou.

Pro potřeby srovnání spolehlivosti předpovědi je vhodné Brierovo skóre standardizovat na *Brier Skill Score*

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\widetilde{q}(1-\widetilde{q})}.$$

Není-li použit žádný ratingový model (všechna pozorování jsou zařazena do jediné třídy), je BSS přibližně nula, v ideálním modelu je BSS rovno jedné.

4.3 Hodnocení diskriminační schopnosti

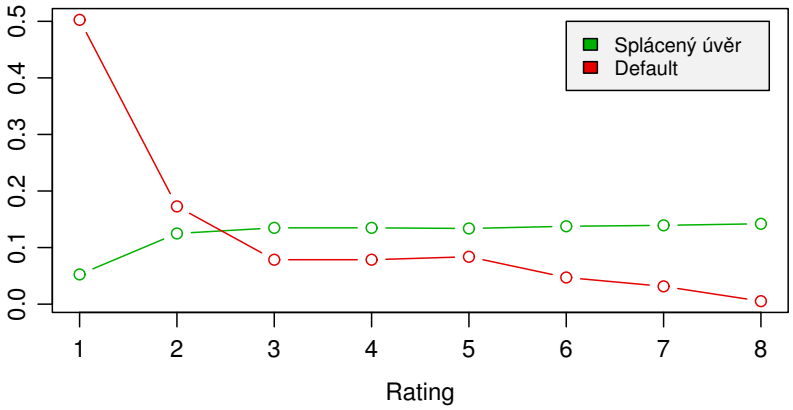
Při hodnocení diskriminační schopnosti se zjišťuje, jak ratingový model rozlišuje defaulty a splácené úvěry. Vychází se z (empirických) podmíněných hustot, které mohou vypadat podobně jako ty na obrázku 15.

Jednotlivé ratingy jsou považovány za možné hranice pro poskytnutí úvěru. A je možné se ptát, kolika defaultům by banka zamezila (*hit rate*) a o kolik splácených úvěrů by přišla (*false alarm rate*), kdyby odmítla všechny uchazeče s ratingem maximálně r. To vyjadřují (podmíněné) distribuční funkce

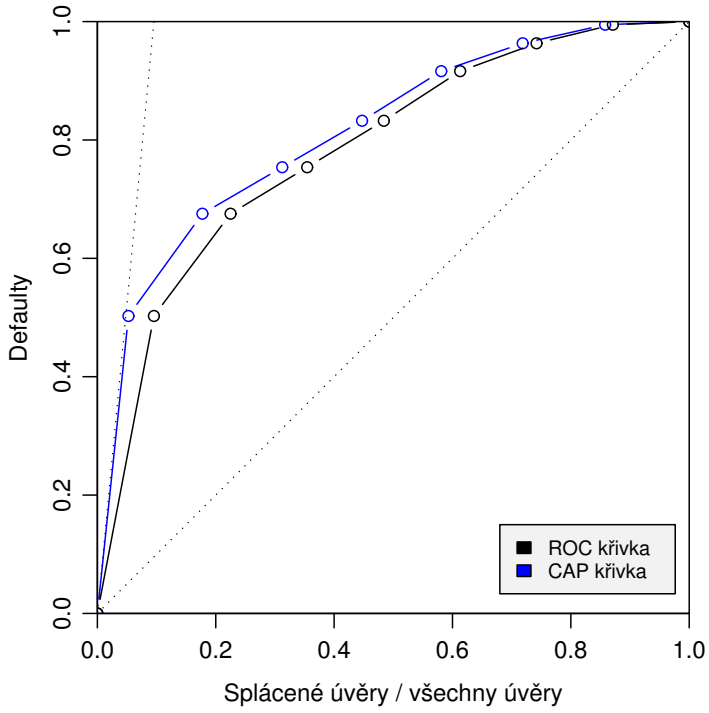
$$\begin{aligned}\mathrm{HR}(r) &= \mathrm{P}(\rho(\boldsymbol{X}) \le r \,|\, Y = 1\,),\\ \mathrm{FAR}(r) &= \mathrm{P}(\rho(\boldsymbol{X}) \le r \,|\, Y = 0\,).\end{aligned}$$

Celkový podíl odmítnutých (*total rate*) udává (nepodmíněná) distribuční funkce

$$\mathrm{TR}(r) = \mathrm{P}(\rho(\boldsymbol{X}) \le r).$$



Obrázek 15: Splácené úvěry a defaulty v jednotlivých ratingových třídách



Obrázek 16: ROC a CAP křivka

ROC křivka

Podmíněné distribuční funkce graficky srovnává *ROC křivka* (černě na obrázku 16). Na vodorovné ose je FAR, na svislé HR.

Ideální model by měl všem defaultům přiřadit ratingy nízké, všem spláceným úvěrům vysoké. ROC křivka v takovém případě vede z rohu levého dolního do levého horního a pak do pravého horního. Naopak ratingový model s náhodnou diskriminací má ROC křivku shodnou s diagonálou. Skutečný model je mezi těmito dvěma extrémy.

Další vlastností ROC křivky je konkávnost. Jednotlivé části ROC křivky odpovídají ratingům a směrnice této úsečky je určena relativním poměrem defaultů a splácených úvěrů v dané třídě. Tento poměr s vyšším ratingem klesá. U méně kvalitních ratingových modelů se však stává, že zmíněný poměr mezi dvěma po sobě jdoucími třídami vzroste a ROC křivka pak konkávní není.

Zkratka "ROC" znamená *receiver operating characteristics*. Podle [13] vyjadřuje, že příjemce (receiver) může využít jako hranici (operate) libovolný bod této křivky. Poslední slovo (characteristics) se vztahuje k rysům pozorování, na jejichž základě byla křivka zkonstruována.

Statistika AUC

Statistikou shrnující vlastnosti ROC křivky do jednoho čísla je AUC (*area under curve*). Tato veličina má význam nejen geometrický, ale i pravděpodobnostní. Platí totiž

$$\begin{aligned}
\mathrm{AUC} &= \sum_{r=1}^{R} \frac{\mathrm{HR}(r-1)+\mathrm{HR}(r)}{2}(\mathrm{FAR}(r)-\mathrm{FAR}(r-1)) \\
&= \sum_{r=1}^{R} \frac{P(\rho(\boldsymbol{X}_1)\le r-1|Y_1=1)+P(\rho(\boldsymbol{X}_1)\le r|Y_1=1)}{2} P(\rho(\boldsymbol{X}_2)=r|Y_2=0) \\
&= \sum_{r=1}^{R} P(\rho(\boldsymbol{X}_1)\le r-1, \rho(\boldsymbol{X}_2)=r|Y_1=1,Y_2=0) \\
&+ \frac{1}{2}\sum_{r=1}^{R} P(\rho(\boldsymbol{X}_1)=r, \rho(\boldsymbol{X}_2)=r|Y_1=1,Y_2=0) \\
&= P(\rho(\boldsymbol{X}_1)<\rho(\boldsymbol{X}_2)|Y_1=1,Y_2=0)+\frac{1}{2}P(\rho(\boldsymbol{X}_1)=\rho(\boldsymbol{X}_2)|Y_1=1,Y_2=0)
\end{aligned}$$

AUC tedy vyjadřuje, jaká je pravděpodobnost, že náhodně zvolený default bude mít nižší rating než náhodně zvolený splácený úvěr (v případě rovnosti ratingů se započítává polovina). A přesně takto se AUC odhaduje - jako průměr přes všechny dvojice typu default a splácený úvěr.

Na takto definované statistice je založen *Mann-Whitneyův test*, modifikace *Wilcoxonova testu* ([2], str. 237 - 241, nebo [15], `wilcox.test`). K testům existují tabelované hodnoty, ale ty předpokládají spojitou odezvu. Proto se vychází z normální aproximace, která umožňuje jak výpočet intervalových odhadů, tak i test významnosti diskriminace porovnáním AUC s číslem 0.5 (hodnota pro ratingový model s náhodnou diskriminací). Konfidenční intervaly je také možné zkonstruovat metodou bootstrapu (bez předpokladu normality).

AUC lze rovněž využít ke srovnání dvou modelů. Lepší je ten, jehož AUC je významně větší. Asymptoticky platí

$$\frac{(\widehat{\mathrm{AUC}}_1-\widehat{\mathrm{AUC}}_2)^2}{\mathrm{var}(\widehat{\mathrm{AUC}}_1)+\mathrm{var}(\widehat{\mathrm{AUC}}_2)-2\,\mathrm{cov}(\widehat{\mathrm{AUC}}_1,\widehat{\mathrm{AUC}}_2)} \sim \chi^2_1.$$

Vzorce pro výpočet variance a kovariance jsou uvedené v [8], str. 13 - 15. Např. variance se odhadne

$$\mathrm{var}(\widehat{\mathrm{AUC}}) = \frac{\widehat{P}_1+(n_0-1)\widehat{P}_2+(n_1-1)\widehat{P}_3-4(n_0+n_1-1)(\widehat{AUC}-0.5)^2}{4(n_0-1)(n_1-1)},$$

kde n_0 a n_1 je počet splácených úvěrů a defaultů. Dále musí být odhadnuty pravděpodob-

nosti

$$
\begin{aligned}
P_1 &= P(\rho(\boldsymbol{X}_1) \neq \rho(\boldsymbol{X}_2)|Y_1 = 0, Y_2 = 1),\\
P_2 &= P(\rho(\boldsymbol{X}_1) < \rho(\boldsymbol{X}_3), \rho(\boldsymbol{X}_2) < \rho(\boldsymbol{X}_3)|Y_1 = 0, Y_2 = 0, Y_3 = 1)\\
&\quad + P(\rho(\boldsymbol{X}_1) > \rho(\boldsymbol{X}_3), \rho(\boldsymbol{X}_2) > \rho(\boldsymbol{X}_3)|Y_1 = 0, Y_2 = 0, Y_3 = 1)\\
&\quad - P(\rho(\boldsymbol{X}_1) < \rho(\boldsymbol{X}_3), \rho(\boldsymbol{X}_2) > \rho(\boldsymbol{X}_3)|Y_1 = 0, Y_2 = 0, Y_3 = 1)\\
&\quad - P(\rho(\boldsymbol{X}_1) > \rho(\boldsymbol{X}_3), \rho(\boldsymbol{X}_2) < \rho(\boldsymbol{X}_3)|Y_1 = 0, Y_2 = 0, Y_3 = 1),\\
P_3 &= P(\rho(\boldsymbol{X}_1) < \rho(\boldsymbol{X}_3), \rho(\boldsymbol{X}_2) < \rho(\boldsymbol{X}_3)|Y_1 = 1, Y_2 = 1, Y_3 = 0)\\
&\quad + P(\rho(\boldsymbol{X}_1) > \rho(\boldsymbol{X}_3), \rho(\boldsymbol{X}_2) > \rho(\boldsymbol{X}_3)|Y_1 = 1, Y_2 = 1, Y_3 = 0)\\
&\quad - P(\rho(\boldsymbol{X}_1) < \rho(\boldsymbol{X}_3), \rho(\boldsymbol{X}_2) > \rho(\boldsymbol{X}_3)|Y_1 = 1, Y_2 = 1, Y_3 = 0)\\
&\quad - P(\rho(\boldsymbol{X}_1) > \rho(\boldsymbol{X}_3), \rho(\boldsymbol{X}_2) < \rho(\boldsymbol{X}_3)|Y_1 = 1, Y_2 = 1, Y_3 = 0).
\end{aligned}
$$

CAP křivka a statistika AR

Všechny úvěry a defaulty graficky srovnává *CAP křivka*. Na vodorovné ose je TR, na svislé pak HR. Oproti ROC křivce jsou tedy splácené úvěry nahrazeny celou validační množinou, skutečný tvar obou křivek je však celkem podobný (jak je vidět na obrázku 16).

CAP křivka ideálního modelu míří z levého dolního rohu prudce nahoru (se směrnicí odpovídající inverzi podílu defaultů na validační množině, viz jedna z tečkovaných čar na obrázku 16). Model bez diskriminační síly má CAP křivku (stejně jako ROC křivku) shodnou s diagonálou.

Statistikou založenou na CAP křivce je AR (*accuracy ratio*). Je-li plocha pod křivkou ideálního modelu a_i, pod křivkou modelu s náhodnou diskriminací a_n a pod křivkou reálného modelu a_r, pak

$$\mathrm{AR} = \frac{a_r - a_n}{a_i - a_n}.$$

Tato veličina tedy udává, jakého podílu maximální možné diskriminace ratingový model dosahuje.

Statistiky AUC a AR poskytují stejnou informaci. Platí totiž

$$\begin{aligned}
a_i &= 1 - \frac{\pi_1}{2}, \\
a_n &= \frac{1}{2}, \\
a_r &= \sum_{r=1}^{R} \frac{\mathrm{HR}(r-1) + \mathrm{HR}(r)}{2} (\mathrm{TR}(r) - \mathrm{TR}(r-1)) \\
&= \sum_{r=1}^{R} \frac{\mathrm{HR}(r-1) + \mathrm{HR}(r)}{2} P(\rho(\boldsymbol{X}) = r) \\
&= \pi_0 \sum_{r=1}^{R} \frac{\mathrm{HR}(r-1) + \mathrm{HR}(r)}{2} P(\rho(\boldsymbol{X}) = r | Y = 0) \\
&+ \pi_1 \sum_{r=1}^{R} \frac{\mathrm{HR}(r-1) + \mathrm{HR}(r)}{2} P(\rho(\boldsymbol{X}) = r | Y = 1) \\
&= \pi_0 \sum_{r=1}^{R} \frac{\mathrm{HR}(r-1) + \mathrm{HR}(r)}{2} (\mathrm{FAR}(r) - \mathrm{FAR}(r-1)) \\
&+ \pi_1 \sum_{r=1}^{R} \frac{\mathrm{HR}(r-1) + \mathrm{HR}(r)}{2} (\mathrm{HR}(r) - \mathrm{HR}(r-1)) \\
&= \pi_0 \ \mathrm{AUC} + \frac{\pi_1}{2} \sum_{r=1}^{R} (\mathrm{HR}(r)^2 - \mathrm{HR}(r-1)^2) \\
&= \pi_0 \ \mathrm{AUC} + \frac{\pi_1}{2} (\mathrm{HR}(R)^2 - \mathrm{HR}(0)^2) \\
&= (1 - \pi_1) \, \mathrm{AUC} + \frac{\pi_1}{2}.
\end{aligned}$$

Celkově se dostane

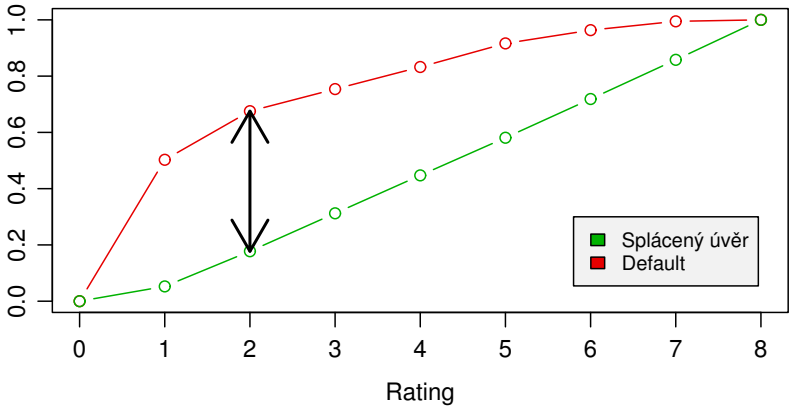
$$\begin{aligned}
\mathrm{AR} &= \frac{(1-\pi_1)\,\mathrm{AUC} + \pi_1/2 - 1/2}{1 - \pi_1/2 - 1/2} = \frac{2(1-\pi_1)\,\mathrm{AUC} + \pi_1 - 1}{1 - \pi_1} \\
&= 2 \ \mathrm{AUC} - 1.
\end{aligned}$$

Výsledky platné pro AUC lze proto využít i v případě AR.

Kolmogorov-Smirnovův test

Kolmogorov-Smirnovův test v základní variantě ([2], str. 243 - 245) srovnává, zda se empirická distribuční funkce rovná skutečné. Kritériem je maximální vzdálenost těchto funkcí. V [15], `ks.test`, je však možné testovat i rovnost dvou empirických distribučních funkcí.

Pro potřeby hodnocení diskriminační schopnosti je vhodné porovnat FAR a HR (viz obrázek 17). Kolmogorov-Smirnovův test si všímá jen maximálního rozpětí mezi těmito křivkami, využívá tedy méně informace než test založený na AUC.



Obrázek 17: Kolmogorov-Smirnovův test

4.4 Informační kritéria

Informační kritéria přistupují k hodnocení ratingových modelů odlišným způsobem. Udávají, kolik informace je potřeba k plné znalosti (tedy k situaci, kdy je o každém úvěru známo, zda je či není defaultem).

(Získanou) informaci měří veličina I. Je-li A nějaký jev a p_A jeho pravděpodobnost, pak $\mathrm{I}(p_A)$ je informace, která se získá, když k jevu A dojde. Je-li p_A jedna, je výsledek předem známý a $\mathrm{I}(p_A)$ je nula. Je-li p_A polovina, pak se získá jeden bit informace a $\mathrm{I}(p_A)$ je proto jedna.

Pro další odvozování se předpokládá "sčítací" vlastnost informace. Je-li B další jev (s pravděpodobností p_B) nezávislý na A, platí

$$\mathrm{I}(p_A p_B) - \mathrm{I}(p_A) = \mathrm{I}(p_B),$$

na obou stranách je totiž informace, která se získá, když dojde k jevu B. Pro libovolné přirozené číslo n se tedy dostane

$$\mathrm{I}(p^n) = n\,\mathrm{I}(p).$$

Protože dále platí

$$\mathrm{I}(p) = \mathrm{I}((p^{1/n})^n) = n\,\mathrm{I}(p^{(1/n)}),$$

lze pro libovolné racionální číslo (a po spojitém rozšíření i pro libovolné reálné číslo) využít vztah

$$\mathrm{I}(p^x) = x\,\mathrm{I}(p).$$

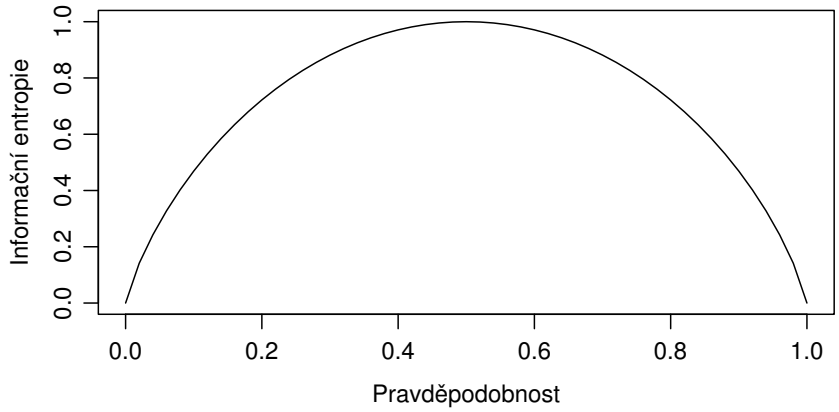
Je-li p rovno $(1/2)^x$ (a tedy x je $-\log_2(p)$), platí

$$\mathrm{I}(p) = \mathrm{I}((1/2)^x) = x\,\mathrm{I}(1/2) = -\log_2(p).$$

Dále se předpokládá, že jev A' nastane právě tehdy, nenastane-li jev A. První možnost má pravděpodobnost $1-p$, druhá p. Střední množství informace potřebné k plné znalosti je

$$p\,\mathrm{I}(p) + (1-p)\,\mathrm{I}(1-p) = -p\log_2(p) - (1-p)\log_2(1-p), \tag{17}$$

tato veličina se nazývá *informační entropie*. Závislost informační entropie na pravděpodobnosti ukazuje obrázek 18.



Obrázek 18: Informační entropie

Informační entropie má jednu zajímavou souvislost s metodou neuronových sítí. Po srovnání vzorců (12) a (17) se zjistí, že parametry neuronové sítě jsou odhadovány tak, aby bylo maximalizované množství informace poskytované neuronovou sítí. (Ve vztazích jsou sice odlišné základy logaritmů, ale protože platí $\log_2(x) = \log(x)/\log(2)$, stačí vydělit konstantou.)

Informační entropie ratingového modelu

Ratingový model, který zařazuje všechna pozorování do jediné třídy, má informační entropii danou vzorcem (17), tedy

$$\mathrm{IE}_0 = -\widetilde{q}\log_2(\widetilde{q}) - (1-\widetilde{q})\log_2(1-\widetilde{q}).$$

Tuto statistiku lze považovat za odhad informační entropie validační množiny.

Informační entropie v jednotlivých třídách skutečného ratingového modelu odhaduje výraz

$$-\widetilde{q}_r\log_2(\widetilde{q}_r) - (1-\widetilde{q}_r)\log_2(1-\widetilde{q}_r),$$

průměrná informační entropie je pak

$$\mathrm{IE}_1(\rho) = -\sum_{r=1}^{R}\frac{\widetilde{n}_r}{\widetilde{n}}(\widetilde{q}_r\log_2(\widetilde{q}_r) + (1-\widetilde{q}_r)\log_2(1-\widetilde{q}_r)).$$

Literatura žádný analytický vztah pro intervalové odhady informační entropie neuvádí. Když byla při implementaci pro tento účel využita metoda bootstrapu, byl odhalen problém s podhodnocováním této veličiny. Podrobnější informace jsou v příloze A.2.

Statistika IER_1

Statistika IER_1 (*information entropy ratio*) je normovanou informační entropií. Je definována vztahem

$$\mathrm{IER}_1(\rho) = 1 - \frac{\mathrm{IE}_1(\rho)}{\mathrm{IE}_0}$$

a udává, jaká část informační entropie validační množiny je ratingovým modelem vysvětlena.

V článku [10] je tato veličina pojmenovaná CIER, tedy *conditional information entropy ratio*. Název zdůrazňuje, že se hodnotí informační entropie podmíněná ratingovým modelem.

Statistika IER_2

Pro srovnání se odhadne informační entropie dvou ratingovým modelů použitých najednou

$$\mathrm{IE}_2(\rho, \rho') = -\sum_{r=1}^{R}\sum_{r'=1}^{R'} \frac{\widetilde{n}_r}{\widetilde{n}} \frac{\widetilde{n}_{r'}}{\widetilde{n}'} (\widetilde{q}_r \widetilde{q}_{r'} \log_2(\widetilde{q}_r \widetilde{q}_{r'}) + (1-\widetilde{q}_r)(1-\widetilde{q}_{r'}) \log_2((1-\widetilde{q}_r)(1-\widetilde{q}_{r'}))).$$

Kolik z informační entropie původního modelu vysvětluje nový model (relativně vzhledem k informační entropii validační množiny), udává statistika

$$\mathrm{IER}_2(\rho, \rho') = \frac{\mathrm{IE}_1(\rho) - \mathrm{IE}_2(\rho, \rho')}{\mathrm{IE}_0}.$$

Tato veličina nemá žádnou vlastnost typu symetrie. $\mathrm{IER}_2(\rho, \rho')$ i $\mathrm{IER}_2(\rho', \rho)$ nabývají kladných hodnot, ty však nejsou provázány nějakou jednoduchou závislostí.

4.5 Metoda bootstrapu

Statistika obvykle činí na základě náhodného výběru závěry o celé populaci. Metoda bootstrapu ke stejnému účelu využívá velké množství výběrů s vracením se stejným rozsahem jako má skutečný výběr (tedy tzv. *bootstrapových výběrů*). Dále předpokládá, že vztah výběru bootstrapového ke skutečnému je analogický vztahu skutečného výběru k populaci.

Tato práce se nezabývá teoretickými vlastnostmi metody bootstrapu (ty jsou uvedeny v [7]), ale popisuje některé z možných aplikací při tvorbě a validaci ratingových modelů.

Intervalové odhady

Intervalové odhady představují pásmo, které s velkou pravděpodobností pokrývá skutečnou hodnotu parametru. Dají se ale chápat i jako interval, do kterého náleží většina bodových odhadů vytvořených na základě náhodného výběru. A pokud se skutečné výběry nahradí výběry bootstrapovými, dostane se velké množství bodových odhadů. Z nich se odvozují odhady intervalové.

Nejjednodušší možností je předpokládat normalitu odhadů. Vypočte se rozptyl bootstrapových odhadů a ten se využije ke konstrukci konfidenčního intervalu. V tomto případě postačuje 25 - 200 bootstrapových výběrů ([7], str. 52).

Jinou variantou je vytvořit z bootstrapových odhadů empirickou distribuční funkci. Intervalové odhady se pak vymezují jejími příslušnými kvantily. Ty mohou být i nesymetrické, což je výhodné zvláště v případě klasifikace. Pro dostatečnou přesnost je však potřeba větší množství bootstrapových výběrů (podle [7], str. 170 - 174, by jich mělo být 1000 - 2000).

Nejlepší možnost představuje BC_a metoda ([7], 184 - 188). Ta zvyšuje přesnost intervalových odhadů pomocí dvou veličin nazývaných *bias correction* a *acceleration* (z nichž název metody vychází). Bias correction vyjadřuje nesoulad mezi mediánem bootstrapových odhadů a odhadem vytvořeným na základě celé validační množiny. Acceleration měří změnu směrodatné odchylky odhadu při změně hodnoty odhadovaného parametru.

Validace bez validační množiny

Obecnou statistiku S (tou může být např. AUC nebo IER) hodnotící kvalitu ratingové funkce ρ lze spočítat i na základě trénovací množiny. Ratingová funkce je však vytvořena tak, aby pro trénovací data dávala co nejlepší výsledky, odhady jsou tedy *vychýlené*. A právě proto se využívá validační množina - odhady na ní založené by měly být nestranné.

Považuje-li se statistika S za funkci ρ, je možné pokusit se odhadnout

$$\mathrm{E}_{(\tilde{\boldsymbol{\mathcal{X}}},\tilde{\boldsymbol{\mathcal{Y}}})}\,\mathrm{S}(\rho) - \mathrm{E}_{(\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{Y}})}\,\mathrm{S}(\rho). \tag{18}$$

Pokud by se tento rozdíl připočetl k hodnotě S pro trénovací množinu, získal by se odhad stejně dobrý jako ten založený na validační množině. Validační množina by přitom existovat nemusela.

Bootstrapové výběry zastupují výběry skutečné, proto lze (18) nahradit výrazem

$$\mathrm{E}_{(\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{Y}})}\,\mathrm{S}(\rho_k) - \mathrm{E}_{(\boldsymbol{\mathcal{X}}_k,\boldsymbol{\mathcal{Y}}_k)}\,\mathrm{S}(\rho_k) \tag{19}$$

(index k označuje bootstrapový výběr nebo ratingovou funkci vytvořenou na základě tohoto výběru). Zatímco (18) bez validační množiny odhadnout nelze, v případě (19) to možné je - bootstrapových výběrů je totiž velké množství. Stačí spočítat průměr

$$\frac{1}{K}\sum_{k=1}^{K}(\mathrm{S}(\rho_k,\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{Y}}) - \mathrm{S}(\rho_k,\boldsymbol{\mathcal{X}}_k,\boldsymbol{\mathcal{Y}}_k))$$

(K je počet bootstrapových výběrů).

Je-li statistika S průměrem přes jednotlivá pozorování, tedy

$$\mathrm{S}(\rho,\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{Y}}) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{S}'(\rho,\boldsymbol{x}_i,y_i)$$

lze její průměr pro bootstrapové ratingové funkce zapsat ve tvaru

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{K}\sum_{k=1}^{K}\mathrm{S}'(\rho_k,\boldsymbol{x}_i,y_i)\right).$$

Tento zápis přenáší důraz z bootstrapových výběrů na jednotlivá pozorování. Bootstrap dává dodatečnou informaci jen v situaci, kdy pozorování není prvkem bootstrapového výběru. Pravděpodobnost, že se tak nestane, je

$$1-\left(1-\frac{1}{n}\right)^{n},$$

limitně $1-\exp(-1)$, zaokrouhleně 0.632. Proto je doporučeno použít odhad

$$0.368\ \mathrm{S}(\rho,\boldsymbol{\mathcal{X}},\boldsymbol{\mathcal{Y}})+0.632\,\frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{K_i}\sum_{k_i=1}^{K_i}\mathrm{S}'(\rho_{k_i},\boldsymbol{x}_i,y_i)\right),$$

"vnitřní" průměry jsou vypočteny pouze z těch výběrů, které dané pozorování neobsahují. Tato metoda je podle [9], str. 124, jedna z nejlepších a díky použité konstantě se nazývá 632 bootstrap.

5 Aplikace

Tato část se zabývá aplikací diskutovaných teoretických poznatků. Nejdříve jsou popsána data, která byla pro tento účel k dispozici. Následují poznámky k implementaci a k provedené simulaci, omezené na logický pohled (z technického hlediska se implementací zabývá příloha A.3). Simulace poskytla velké množství nejrůznějších výsledků. Zbytek této části je proto věnován interpretaci některých z nich.

Data

Bylo by určitě velmi zajímavé využít skutečná data, ty se ale získat nepodařilo. V České národní bance však byl naprogramován generátor portfolia úvěrů, jehož výstupy byly využity i v této práci. Zde jsou zmíněné některé jeho vlastnosti.

Generování portfolia probíhá po krocích chápaných jako měsíční období. Každý sledovaný úvěr se v určitém okamžiku nachází v jednom ze sedmi stavů, mezi kterými přechází s pravděpodobnostmi přechodu určenými pevnou transformační maticí. Stavy jsou značeny čísly $1, 2, \ldots, 7$, vyšší číslo znamená větší pravděpodobnost defaultu.

Pro každý úvěr je dále stanovena jistina a doba splatnosti. Je-li úvěr v některém z prvních šesti stavů, je považován za splácený. V každém kroku se pak od nesplacené části jistiny odečítá měsíční splátka. Po splacení celé jistiny úvěr přestává být sledován.

Je-li úvěr v posledním (sedmém) stavu, je pokládán za nesplácený. Trvá-li tato situace stanovenou dobu (standardně pět měsíců), je označen za default a dále není sledován. Jinak dochází k posunutí doby splatnosti o jeden měsíc.

V závislosti na aktuálním stavu je úvěrům přiděleno osm vstupů. Je-li úvěr i v čase t ve stavu $s^{i,t}$, platí

$$\begin{aligned}
x_1^{i,t} &\sim |N(5s^{i,t}, 1)|,\\
x_2^{i,t} &\sim |N(0,1)|,\\
x_3^{i,t} &\sim \log(|N(2s^{i,t}, 1)|),\\
x_4^{i,t} &\sim |N(0,1)|,\\
x_5^{i,t} &\sim N(0,1),\\
x_6^{i,t} &\sim 1/s^{i,t} + N(0,1),\\
x_7^{i,t} &\sim N(0,1),\\
x_8^{i,t} &\sim I_{\{-(s^{i,t}+2R(0,1)-1)>3\}}
\end{aligned}$$

(I značí indikátor a R rovnoměrné rozdělení). Situace je tedy opačná než ve skutečnosti - na základě výstupů se rozhoduje o vstupech. Vektor $\boldsymbol{x}$ navíc obsahuje i znaky, které s aktuálním stavem nijak nesouvisí, což v praxi obvyklé nebývá.

Pro zjednodušení byl generátor vytvořen tak, aby další vývoj každého úvěru nezávisel na historii, ale pouze na současném stavu. Proto není chybou zařadit do vytvářeného datového souboru různá (disjunktní) časová období jednoho úvěru. Jako vysvětlující pro-

měnná tedy slouží $\boldsymbol{x}^{i,t}$ v čase přidělení úvěru a dále vždy po dvanácti měsících. Odezvou je přítomnost defaultu do dvanácti měsíců od okamžiku získání těchto vstupů.

Implementace

K výpočtům je použit statistický jazyk R ([15]), výsledky jsou ukládány do databáze. K jejich prohlížení slouží webové rozhraní. Výsledky provedené simulace jsou dostupné i na internetu (http://diplomka.pm13.org/).

Vlastní výpočty se provádí spuštěním jednoho ze sedmi skriptů - `create.sample.R`, `do.lda.R`, `do.logit.R`, `do.nnet.R`, `predict.R`, `validate.R` a `compare.R`. Každý z nich má několik parametrů řídících průběh výpočtu.

Mezi hlavní složky vytvořeného systému patří datový soubor, výběr, ratingový model, predikce, validace a srovnání.

Datový soubor obsahuje všechny potřebné záznamy o úvěrech. Dá se chápat jako matice s $p + 1$ sloupci (prvních p pro vstupy, poslední pro odezvu). Systém může obsahovat velké množství datových souborů. Není tedy závislý na souboru vytvořeném generátorem portfolia úvěrů. Pokud by byla k dispozici skutečná data, bylo by triviální záležitostí je doplnit, celou simulaci zopakovat a dojít k realističtějším závěrům.

K jednomu datovému souboru může existovat několik *výběrů*. K jejich tvorbě slouží skript `create.sample.R` (parametrem je identifikátor datového souboru). Výběr má stejný tvar jako datový soubor, řádky jsou však "přeházeny". Prvních n řádků slouží jako množina trénovací, posledních $\widetilde{n}$ je zase množina validační. Tento mechanismus jednoduchým způsobem reprezentuje náhodný výběr.

Skripty `do.lda.R`, `do.logit.R` a `do.nnet.R` (parametry jsou identifikátor výběru a velikost trénovací množiny) vytváří *ratingové modely*. Jako metody jsou použity lineární diskriminační analýza, logistická regrese a neuronové sítě. Ve všech případech jsou nalezeny váhy, metodou bootstrapu jsou odhadnuty některé statistiky (AUC, IE, IER) a je vytvořena ratingová funkce (takovým způsobem, aby prvky trénovací množiny byly rozděleny mezi jednotlivé ratingové třídy přibližně rovnoměrně).

Predikci provádí skript `predict.R` (parametry jsou identifikátor ratingového modelu a vstupy úvěru). Pro zadaný úvěr je odhadnuta pravděpodobnost defaultu včetně intervalových odhadů vytvořených metodou bootstrapu a delta metodou (u lineární diskriminační analýzy delta metoda použita není).

O *validaci* se stará skript `validate.R` (parametry jsou identifikátor ratingového modelu a velikost validační množiny). Prvky validační množiny jsou zařazeny do ratingových tříd a je zhodnocena kvalita modelu - graficky (ROC křivka, diagram spolehlivosti) i číselně (AUC, IE, IER).

Srovnání dvou ratingových modelů provádí skript `compare.R` (parametry jsou identifikátory ratingových modelů a velikost validační množiny). Průběh je podobný jako v případě validace.

Některé parametry (např. počet bootstrapových výběrů nebo počet ratingů) jsou společné pro všechny skripty. Konkrétní hodnoty jsou nastaveny v souboru `require.R`.

Simulace

K vygenerovanému datovému souboru bylo vytvořeno deset náhodných výběrů. Každý výběr byl využit ke konstrukci třiceti dvou ratingových modelů - pomocí čtyř metod (lineární diskriminační analýza, logistická regrese, neuronová síť s chybovou funkcí součet čtverců a neuronová síť s chybovou funkcí entropie) a osmi různých velikostí výběru (500, 1000, 2500, 5000, 7500, 10000, 12500, 14957). Každý ratingový model byl použit k predikci dvaceti úvěrů a poté třikrát zhodnocen - s využitím validačních množin poloviční, stejné a dvojnásobné velikosti. Nakonec byly některé z modelů srovnány.

Pro zrychlení výpočtu byly změněny některé konstanty. Počet bootstrapových výběrů byl snížen na sto. Všechny neuronové sítě měly stejný počet skrytých uzlů (3) a stejnou velikost regularizačního parametru (0.01).

Všechny závěry, zformulované na základě této simulace, jsou podmíněny použitým datovým souborem. Toto omezení je třeba brát v úvahu. A to i v případě, kdy tento fakt explicitně zmíněn není.

Dále je nutno upozornit, že simulace nebyla prováděna s nějakým konkrétním záměrem. Cílem bylo ověřit, že diskutované poznatky jsou prakticky použitelné. Z tohoto důvodu výsledky spíše naznačují, jaké problémy by mohly být řešeny, než aby dávaly odpovědi na konkrétní otázky.

Velikost trénovací a validační množiny

Jeden z hlavních problémů při tvorbě ratingového modelu je potřebné množství dat. Jako kritérium může sloužit variabilita AUC - statistiky hodnotící diskriminační schopnost ratingového modelu. Dá se vyjít z hodnoty AUC pro různé výběry nebo z konfidenčních intervalů. V druhém případě je možné požadovat, aby dolní mez tohoto intervalu přesáhla určitou hranici nebo aby byl konfidenční interval dostatečně úzký.

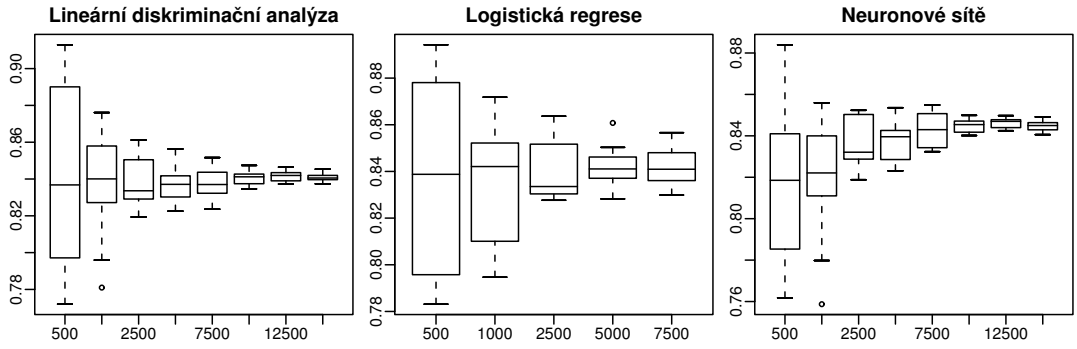
Obrázek 19 srovnává AUC pro tři různé metody (u neuronových sítí je chybovou funkcí entropie). Obrázek 20 ukazuje dolní mez konfidenčního intervalu vytvořeného pomocí normální aproximace se spolehlivostí 95 % pokrývající skutečnou hodnotu AUC. Velikosti obou množin (trénovací a validační) se neliší a jsou udané na vodorovné ose (celkový počet úvěrů je tedy dvojnásobný).

Zdá se, že 2000 úvěrů nestačí. Simulace naznačuje, ze minimální počet může být 3000 - 6000. V celém datovém souboru je 9.16 % defaultů, při tvorbě ratingového modelu lze tedy doporučit použití datových souborů s počtem defaultů v řádu několika set.

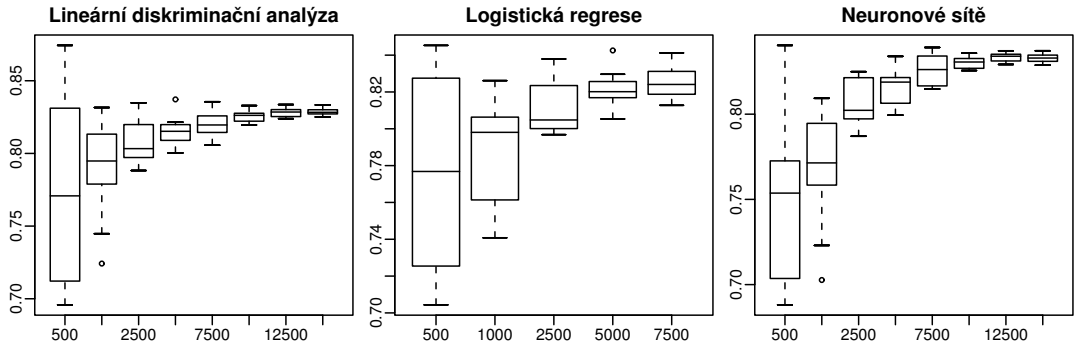
Dělení dat na trénovací a validační množinu

Je-li k dispozici jen omezené množství dat, je důležité správně rozhodnout, jak výběr rozdělit na trénovací a validační množinu. Kritériem opět může být ukazatel diskriminační schopnosti.

Výsledky simulace umožňují porovnat, zda je lepší rozdělit výběr obsahující 1500 úvěrů na trénovací a validační množinu v poměru 1:2 nebo 2:1. Z obrázku 21 lze usoudit, že hodnota AUC je v obou případech přibližně stejná. Je-li však v trénovací množině 500

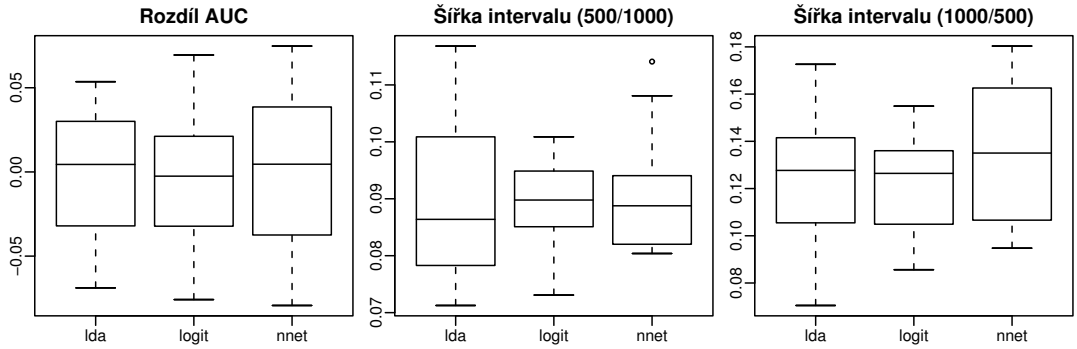


Obrázek 19: AUC v závislosti na použité metodě a na velikosti trénovací a validační množiny



Obrázek 20: Dolní mez pro AUC (normální aproximace).

úvěrů, konfidenční intervaly jsou užší. Z tohoto důvodu se jako výhodnější jeví dělení v poměru 2:1.



Obrázek 21: Rozdíl AUC pro dvě různá dělení datového souboru (kladné hodnoty znamenají výhodnost dělení 500/1000) a šířka konfidenčního intervalu u obou variant.

Z výsledků však už není možné zjistit, zda by nebylo nejlepší použít poměr 1:1. A také není zřejmé, jak by se měl rozdělit větší počet úvěrů.

Chybová funkce pro neuronové sítě

K tréninku neuronové sítě lze využít dvě různé chybové funkce - součet čtverců nebo entropii. Teoretické úvahy naznačují, že entropie by mohla být pro odhad binární odezvy vhodnější.

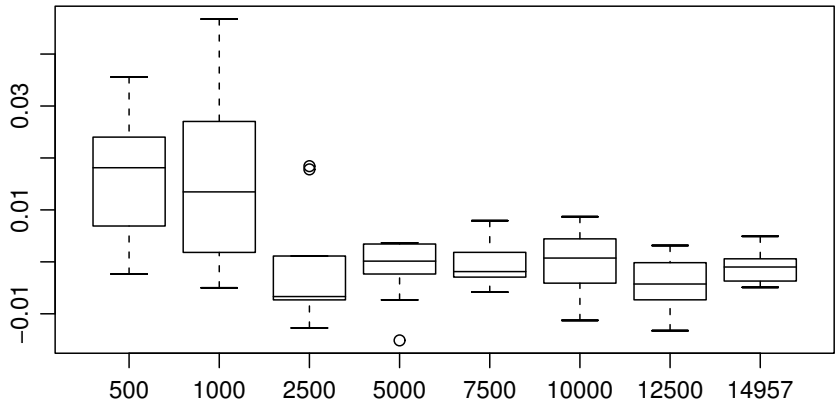
Jenže výsledky simulace tento názor nepotvrzují. Jak ukazuje obrázek 22, při použití větší trénovací množiny (validační množina obsahuje ve všech případech 2000 úvěrů) mají neuronové sítě stejnou diskriminační sílu bez ohledu na volbu chybové funkce. Obsahuje-li trénovací množina 500 nebo 1000 úvěrů, dokonce se zdá, že je lepší volbou součet čtverců. Rozdíly však nejsou takové, aby byl tento závěr statisticky významný. Bylo by vhodné provést další srovnání.

Všechny neuronové sítě byly vytvořeny se stejným počtem skrytých uzlů a se stejnou hodnotou regularizačního parametru. Bylo by zajímavé zjistit, jaké by byly rozdíly při jiné volbě.

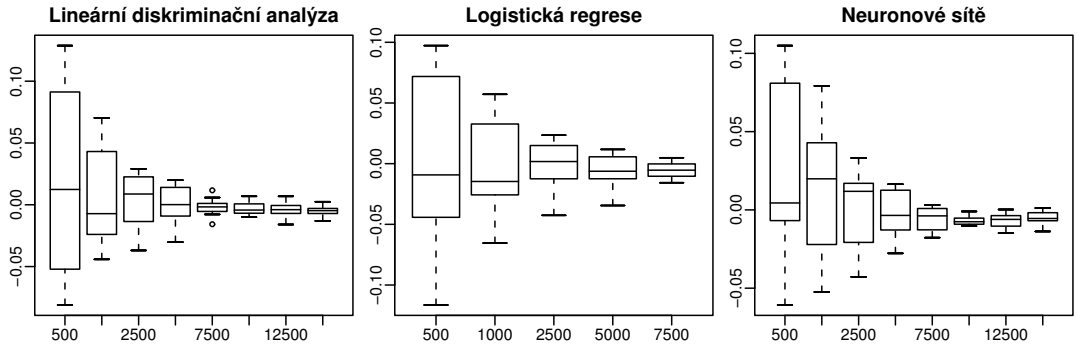
Validace bez validační množiny

Jak je uvedeno v části 4.5, metoda bootstrapu umožňuje zhodnotit kvalitu ratingového modelu bez použití validační množiny. Výsledky simulace ukazují, jak dobré toto hodnocení je.

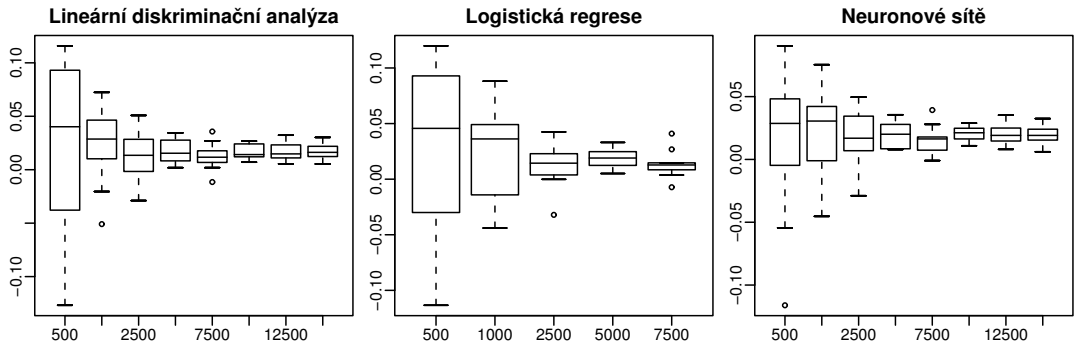
Obrázek 23 naznačuje, že trénovací množina obsahující 2500 úvěrů dává dostatečně dobré odhady AUC. Ke stejnému závěru nelze dojít u odhadu IE (viz obrázek 24), v tomto případě nejspíše bude příčinou problém popsaný v následujících odstavcích.



Obrázek 22: Rozdíl AUC pro dvě neuronové sítě se stejnými daty, ale různou chybovou funkcí. (Kladné hodnoty znamenají, že větší diskriminační sílu mají neuronové sítě trénované s použitím chybové funkce součet čtverců.)



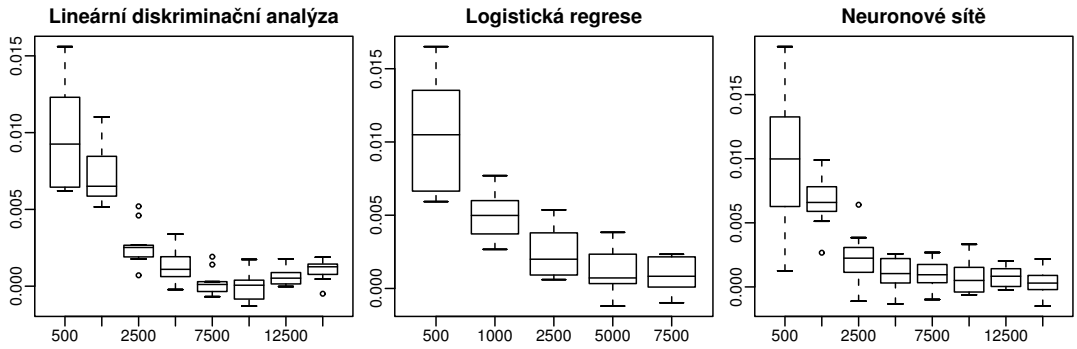
Obrázek 23: Rozdíl mezi odhady AUC založenými na trénovací a validační množině. (Kladné hodnoty znamenají, že odhad AUC založený na trénovací množině je vyšší.)



Obrázek 24: Rozdíl mezi odhady IE založenými na trénovací a validační množině. (Kladné hodnoty znamenají, že odhad IE založený na trénovací množině je vyšší.)

Bootstrapové odhady informační entropie

Simulace odhadla IE ratingových modelů s menší trénovací a validační množinou takovými hodnotami, které ležely v horní části konfidenčního intervalu vytvořeného metodou bootstrapu. Podobný problém je patrný i z obrázku 25.



Obrázek 25: Rozdíl mezi odhadem IE vypočteným na základě celé validační množiny a průměrem bootstrapových odhadů stejné veličiny. (Kladné hodnoty znamenají, že průměr bootstrapových odhadů IE je nižší.)

Situace je dále komentována v příloze A.2.

Delta metoda u logistické regrese

U mnoha obrázků v této části chybí výsledky pro ratingové modely vytvořené logistickou regresí s použitím rozsáhlejší trénovací množiny. Příčinou je násobení ve vzorci (8). Pro odvození rozptylu vah (nutného pro použití delta metody) je totiž nutné provádět příliš náročné maticové operace.

Pokud by byla v těchto případech delta metoda skutečně nutná, mohlo by se postupovat podobně jako u neuronových sítí. I v případě logistické regrese totiž lze využít algoritmus back-propagation.

6 Závěr

Tato diplomová práce popsala tři různé statistické metody používané v bankovnictví pro klasifikaci a tvorbu ratingových modelů. Podrobně rozebrala jejich principy a předpoklady. Ukázala, co mají jednotlivé metody společného a čím se naopak liší. Uvedla, jak se tyto metody dají využít k odhadování pravděpodobnosti defaultu a ke klasifikaci.

Dále se práce zaměřila na ratingové modely. Naznačila, jak se tyto modely vytváří a kalibrují. Rozebrala, jak je možné vzniklý model zvalidovat. Využila k tomu dvě různá kritéria - diskriminační schopnost modelu a informaci poskytovanou modelem. Odůvodnila, proč je druhé kritérium často nepřesné.

Závěrem se práce zabývala aplikací teoretických poznatků. Popsala provedenou simulační studii. Uvedla i některé výsledky, které tato simulace přinesla, včetně několika problémů, které by mohly být dále řešeny. Mezi ně patří především potřebná velikost datového souboru a jeho rozdělení na trénovací a validační množinu. Bylo by také zajímavé prozkoumat, zda je možné datový soubor nedělit a učinit všechny potřebné závěry pouze na základě trénovací množiny.

A Přílohy

A.1 Data pro vizualizaci

Tato trénovací množina obsahuje 330 pozorování z normálních rozdělení s různými parametry. Polovina z 300 splácených úvěrů má střední hodnotu $\begin{pmatrix}0.3\\0.3\end{pmatrix}$ a rozptyl $\begin{pmatrix}0.15 & 0\\0 & 0.15\end{pmatrix}$. Varianční matice se u druhé poloviny neliší, střední hodnota je $\begin{pmatrix}0.7\\0.7\end{pmatrix}$. Defaulty mají rozptyl $\begin{pmatrix}0.1 & 0\\0 & 0.1\end{pmatrix}$. Střední hodnota je ve 25 případech $\begin{pmatrix}0.2\\0.8\end{pmatrix}$, ve zbylých 5 pak $\begin{pmatrix}0.8\\0.2\end{pmatrix}$.

Ztráta plynoucí z chybného zařazení do třídy 0 je desetkrát větší než ztráta plynoucí z chybného zařazení do třídy 1.

Na základě těchto dat není v žádném případě možné činit závěry o popisovaných metodách. Parametry jsou totiž zvoleny tak, aby byla ilustrována odlišnost neuronových sítí. Při jiných hodnotách byl rozdíl mezi lineárními metodami a neuronovými sítěmi zanedbatelný.

A.2 Informační entropie a metoda bootstrapu

Simulace odhalila problém týkající se informační entropie. Při nižším počtu pozorování byly výsledkem metody bootstrapu podhodnocené odhady této statistiky.

Proto byla provedena nová simulace, která předpokládala, že úvěry pocházejí z alternativních rozdělení s pravděpodobnostmi defaultu uvedenými v tabulce 1. Pravděpodobnosti příslušnosti k jednotlivým ratingovým třídám byly rovnoměrné. Za této situace informační entropie vychází 0.3449.

1	2	3	4	5	6	7
0.5	0.1	0.05	0.04	0.03	0.02	0.01

Tabulka 1: Pravděpodobnosti defaultu podle ratingu

Opakovaně (10000krát) bylo vygenerováno n úvěrů a na základě každého výběru byla odhadnuta informační entropie. Průměrné hodnoty obsahuje tabulka 2. Je vidět, že odhady jsou dostatečně přesné až při počtu pozorování v řádu desetitisíců.

100	200	500	1000	2000	5000	10000	20000
0.2973	0.3185	0.3341	0.3398	0.3425	0.3439	0.3443	0.3447

Tabulka 2: Průměrný odhad informační entropie v závislosti na počtu pozorování

Problém vysvětluje tabulka 3. Při pravděpodobnosti defaultu 0.01 je informační entropie 0.080. Ale střední hodnota odhadu založeného na pětičlenném náhodném výběru je pouze 0.007. Za takovéto situace totiž existuje velká pravděpodobnost, že ve výběru nebude žádný default a informační entropie bude nulová.

	Pravděpodobnost jevu	Odhad pravděpodobnosti defaultu	Odhad informační etropie
0	0.9509900499	0.0	0.00
1	0.0096059601	0.2	0.72
2	0.0000970299	0.4	0.97
3	0.0000009801	0.6	0.97
4	0.0000000099	0.8	0.72
5	0.0000000001	1.0	0.00

Tabulka 3: Odhad informační entropie v závislosti na počtu defaultů

Podobné závěry platí i pro odhad informační entropie vycházející z validační množiny. Nepřesné mohou být i další statistiky založené na informační entropii.

A.3 Implementace

Implementace se dá rozdělit na dvě části. K výpočtům je použit statistický jazyk R (http://www.r-project.org/). K prohlížení výsledků slouží webové rozhraní vytvořené v jazyce PHP (http://www.php.net/). Obě složky využívají (pro práci s daty) databázový systém MySQL (http://www.mysql.com/).

Software uvedený v předchozím odstavci existuje pro více operačních systémů. Ale testován byl pouze systém Linux. Navíc jsou pro některé úkony využívány ještě další skripty (pro shell a python), které v jiných systémech pravděpodobně fungovat nebudou.

Dále je potřeba upozornit, že všechny soubory jsou uloženy v kódování UTF-8. Při použití jiné znakové sady (ISO-8859-2, windows-1250) jsou tedy slova s diakritikou nečitelná.

Nejzajímavější soubory se nacházejí v adresáři `run`. Především v něm jsou (velmi krátké) skripty popsané v části 5. Příkladem může být soubor `create-sample.R`:

```
# Nový náhodný výběr.

data.set.id = 1 # identifikátor datového souboru
description = 'První náhodný výběr' # popis výběru

parse.arguments('data.set.id', 'description')
create.sample(data.set.id, description)
```

Podobnou strukturu mají i ostatní skripty. Nejdříve jsou zadány parametry (buď úpravou souboru nebo z příkazové řádky) a pak je zavolána funkce provádějící výpočet. Ta může postupně využívat velké množství funkcí uložených v podadresářích adresáře `run`.

V souboru `require.R` se nachází kód společný pro více skriptů. V prostředí R tedy lze použít příkaz `source('require.R'); source('create-sample.R');` (místo řetězce `create-sample.R` může být název některého z ostatních skriptů). Jinou možností je využití shellovského programu `run`, jehož parametrem je název skriptu (případné další parametry jsou předány spuštěnému skriptu).

Celou simulaci provádí programy `run-all.py` a `compare-more.py` (tím, že opakovaně s různými parametry spouští skript `run`). Všechny tři programy (`run`, `run-all.py` a `compare-more.py`) očekávají, že aktuálním adresářem je `run`.

Soubory týkající se databáze se nachází v adresáři `database`. Je zde např. `schema.xml` (popisující schéma databáze), `data.txt` (obsahující výstup generátoru portfolia úvěrů), `data2db.py` (ukládající výstup generátoru portfolia úvěrů do databáze) a `simulation.sql` (obsahující všechna data vytvořená v průběhu simulace).

Pro jednodušší vytvoření webového rozhraní byl použit perzistentní framework Propel (http://propel.phpdb.org/, verze 1.0.x). Ten umožňuje ze souboru popisujícího schéma databáze vygenerovat jak skript v jazyce SQL, tak i množství PHP tříd výrazně ulehčujících práci s perzistentními objekty. Soubory potřebné pro tento úkol jsou v adresáři `persistence`.

Webové rozhraní se nachází v adresáři `web`. Do podadresáře `images` se ukládají obrázky vygenerované v průběhu výpočtu. Soubory potřebné pro Propel (a soubory jím generované) jsou v podadresáři `persistence`. Vlastní webové stránky jsou v podadresáři `view`.

Pro ulehčení některých častých úkonů bylo vytvořeno několik shellovských skriptů uložených v adresáři `bin`. Skript `persistence` generuje soubory pro perzistenci. Skript `database` připravuje databázi. Skript `restore` uvádí databázi a adresáře pro ukládání obrázků do stavu, ve kterém je systém prázdný (neobsahuje žádný ratingový model) a připravený pro výpočty. Skript `run` spouští simulaci. Všechny skripty musí být spouštěny přímo v adresáři `implementation`.

Pro případné otestování je potřeba provést tyto kroky:

1. Zkontrolovat, zda je v systému přítomen požadovaný software:
   - R od verze 2.0.x, včetně balíčků DBI a RMySQL,
   - Apache od verze 2.0.x, včetně modulů mod_php5 a mod_rewrite,
   - MySQL od verze 4.1.x,
   - Python od verze 2.3.x.
2. Vytvořit novou databázi.
3. Upravit soubory obsahují název databáze, uživatelské jméno a heslo:
   - `bin/database`,
   - `bin/restore`,
   - `bin/run`,
   - `database/database.sql`,
   - `persistence/persistence-properties.xml`,
   - `run/utilities/get.connection.R`.
4. Spustit skript `bin/persistence`.

5. (Pro prohlížení výsledků simulace) zkopírovat obsah adresáře `web` do kořenového adresáře serveru a vložit do databáze `database/simulation.sql`.

6. (Pro provádění výpočtů) spustit skript `bin/restore` a učinit adresář `web` kořenovým adresářem serveru Apache.

A.4 Obsah přiloženého CD

Přiložené CD obsahuje tři adresáře - `graphs`, `implementation` a `simulation`. V adresáři `graphs` jsou skripty, které generují grafy použité v práci. Nejdůležitější adresář `implementation` obsahuje soubory blíže popsané v příloze A.3. V adresáři `simulation` jsou skripty provádějící simulaci z přílohy A.2.

Literatura

[1] Agresti Alan (1990): Categorical Data Analysis. John Wiley & Sons.

[2] Anděl Jiří (2002): Základy matematické statistiky. Univerzita Karlova v Praze, Matematicko-fyzikální fakulta.

[3] Anderson T. W. (2003): An Introduction to Multivariate Statistical Analysis, Third Edition. John Wiley & Sons.

[4] Basel Committee on Banking Supervision (2004): International Convergence of Capital Measurement and Capital Standards, A Revised Framework. Bank for International Settlements.

[5] Bishop Christopher M. (1995): Neural Networks for Pattern Recognition. Oxford University Press.

[6] Dybowski Richard, Roberts Stephen J. (2001): Confidence Intervals and Prediction Intervals for Feed-Forward Neural Networks. *Clinical Applications of Artifical Neural Networks, 298 - 326.*

[7] Efron Bradley, Tibshirani Robert J. (1993): An Introduction to the Bootstrap. Chapman & Hall/CRC.

[8] Engelmann Bernd, Hayden Evelyn, Tasche Dirk (2002): Measuring the Discriminative Power of Rating Systems. Deutche Bundesbank.

[9] Hand David J. (1997): Construction and Assessment of Classification Rules. John Wiley & Sons.

[10] Keenan S. C., Sobehart J. R. (1999): Performance Measures for Credit Risk Models. Moody's Risk Management Services.

[11] Mays Elizabeth (1998): Credit Risk Modeling: Design and Application. Amacom.

[12] McLachlan Geoffrey J. (1992): Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons.

[13] Metz Charles E. (1978): Basic Principles of ROC Analysis. *Seminars in Nuclear Medicine, 283 - 298.*

[14] Oesterreichische Nationalbank (2004): Rating Models and Validation. Oesterreichische Nationalbank.

[15] R Development Core Team (2004): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

[16] Ripley B. D. (1996): Pattern Recognition and Neural Networks. Cambridge University Press.