Bc. Jana Hricová:
Metody konstrukce klasifikátorů vhodných pro segmentaci zákazníků
(Posudek oponenta)

Předložená práce se zabývá, jak požaduje zadání, metodami konstrukce klasifikátorů, klasifikačními stromy a lesy. Navíc obsahuje pasáže o podobné úloze, kdy místo klasifikace předpovídáme hodnotu spojité závisle proměnné. Zmíněné je obsahem 2. a 3. kapitoly. Rozsáhlá poslední kapitola obsahuje ukázky použití oněch metod. Na doprovodném CD jsou uvedena veškerá použitá data i skripty z prostředí R.

Poměrně rozsáhlá práce se snaží vysvětlit nejznámější metody a pojmy v nich vystupující. Jde o metody, které nejsou obsahem běžných přednášek, takže bylo třeba nastudovat potřebnou literaturu a vše v jednotném značení souvisle vysvětlit. Ne vždy se to podařilo. V úvodu se píše o R hodnotách závisle proměnné, jinde je jich J. Podobně je v Úvodu klasifikátor označen symbolem $d()$, kdežto ve 2. kapitole je to $C(.)$ Chyba se zřejmě vloudila do zavedení funkcí $CV(d)$ a $CV(d, \alpha)$ na str. 22, přičemž není jednoznačný význam α. Na str. 23 má symbol K dokonce dvojí význam. V textu se pracuje s nedefinovanými pojmy. Na str. 23 jsou zavedeny chyby $e(T), e'(T)$ způsobem, kterému neodpovídá hned následující příklad 2.1.1. Nejasný je význam slova nezávislost (někde asi disjunktnost množin), Některé části jsou téměř doslovným překladem článku [4] do slovenštiny, navíc na str. 27 nahoře s vynecháním důležitého slova, které mění význam.

Část 2.4 vychází z Breimanova článku [6], ale navíc obsahuje předpoklad, že v případě spojité závisle proměnné Y ná ϵ normální rozdělení. Proč? Zmíněný článek pracuje v případě klasifikační úlohy s nestandardní terminologií, na což bylo vhodné v diplomce upozornit.

Velkým kladem předložené práce je 4. kapitola, v níž autorka práce aplikuje metody vysvětlené v předchozích kapitolách na data. Zejména datový soubor `kredit` odpovídá zadání práce. Škoda, že jsem nenašel osvětlení původu těchto dat. Diplomantka podrobně ukazuje aplikaci jednotlivých metod, výsledky podrobně komentuje. Za jistý nedostatek považuji to, že s prediktory tu fakticky zachází jako se spojitými veličinami, přestože se u některých nejedná ani o měřítko ordinální, spíše jen nominální. Pokusil jsem se použít klasickou logistickou regresi dvěma způsoby: modelem bylo použití řady prediktorů v nominálním měřítku, podmodelem pak jejich použití jako spojitých veličin. Model však dal prokazatelně lepší výsledek.

Počet technických nedopatření či překlepů odpovídá délce práce, je spíše malý. Za zmínku, kromě zmíněné definice $CV(d)$, stojí dvojí označení vychýlení (bias vers. *Bias* na str. 33), n resp. N na str. 20, chybějící pravá závorka v prvním vzorci na str. 30 či *identifikátor* místo indikátor na str. 22. Také některé formulace jsou obtížně srozumitelné (např. první věta na str. 46).

Při obhajobě očekávám vysvětlení některých nedokazovaných tvrzení, např. nerovnosti $0 \leq R^2 \leq 1$ uvedené na str. 31.

Práce je v zásadě kompilační, ovšem 4. kapitola uvádí velmi náročnou aplikaci. Zadání práce bylo nepochybně splněno. **Doporučuji uznat předloženou práci jako práci diplomovou.**

V Praze dne 14. srpna 2013 Karel Zvára