

## POSUDEK VEDOUCÍHO DIPLOMOVÉ PRÁCE

**Název:** Logistická regrese s misklasifikovanou odezvou

**Autor:** Martin Farda

### Shrnutí:

Martin Farda měl za úkol zabývat se ve své práci problematikou logistické regrese s misklasifikovanou odezvou a prozkoumat metody umožňující v takové situaci získat konsistentní odhady parametrů logistického modelu. Téma bylo inspirováno dobou covidovou, kdy se masově prováděly různé druhy testů na přítomnost viru v organismu, přičemž každý z nich mohl v určitých případech vydávat falešně pozitivní či falešně negativní výsledky. Při analýzách covidových dat se tento fakt téměř vždy zanedbával bez ohledu na to, jak závažné důsledky mohou takové chyby v datech mít na jakékoli výsledky analýz a závěry z nich činěné.

Protože binární odezva nemůže mít jiné než alternativní rozdělení, lze celý problém řešit v rámci parametrických modelů pomocí teorie maximální věrohodnosti. Prvním úkolem diplomata tedy bylo odvodit věrohodnostní funkci pro daný problém v situaci, kdy sensitivita a specificita (čili pravděpodobnosti determinující misklasifikaci odezvy) jsou známy. Přestože tato věrohodnost byla již dříve publikována, diplomant byl veden k tomu, aby ji odvodil samostatně a teprve pak výsledek porovnal s tím, co existuje v literatuře.

Protože tato věrohodnost nemá analyticky jednoduchý tvar, je nutné odhad hledat pomocí nějaké numerické metody, např. EM algoritmu. Martin Farda byl tedy pověřen tím, aby odvodil EM algoritmus pro tento problém a pak teprve porovnal výsledek s tím, který je k dispozici v literatuře. Vzhledem ke složitému tvaru věrohodnosti pak byla provedena podrobná numerická analýza věrohodnostní funkce dovedená až do absurdně malých rozsahů výběru, abychom se přesvědčili, zdali je maximálně věrohodný odhad vůbec rozumně determinován a zdali nemá věrohodnostní funkce řadu lokálních maxim či minim, které by numerické hledání globálního maxima komplikovaly. Z výsledků této činnosti byla nakonec do diplomové práce vložena jen velmi malá část.

Pak jsme se začali zabývat situací, kdy sensitivita a specificita nejsou známy. Jsou-li to konstanty, je možné pokusit se je odhadnout z misklasifikovaných dat zároveň s parametry logistické regrese. V praxi jsou ale takové odhady dosti nepřesné. Mnohem lepší je získat informace o těchto parametrech výběrem validační skupiny získané z původních dat pomocí Bernoulliho výběru. Na validační skupině se pak k misklasifikované odezvě doměří i skutečná odezva. Věrohodnostní funkce se upraví tak, aby zahrnovala jak validační skupinu tak zbytek datového souboru a všechny parametry se odhadují zároveň metodou maximální věrohodnosti.

Pro tuto situaci Martin Farda navrhl a implementoval svůj vlastní EM-algoritmus a doplnil jej výpočtem odhadu asymptotického rozptylu a metodikou pro testování hypotéz a konstrukci intervalů spolehlivosti. Tento algoritmus pak implementoval v softwaru R a využil k provedení simulační studie vyhodnocující výsledky navrženého algoritmu v realistických situacích. V literatuře takový algoritmus neexistuje; věrohodnostní funkce pro tento problém sice publikována byla, ale autoři ji navrhují maximalizovat čistě numerickými metodami. Martin Farda na simulacích ukázal, že jím navržený EM-algoritmus v některých situacích dává mnohem spolehlivější výsledky než ryze numerický přístup, o výpočetní náročnosti ani nemluvě.

**Téma práce.** Téma práce vyžaduje schopnost aplikovat tvůrčím způsobem teorii maximální věrohodnosti a EM-algoritmy na konkrétní problém chybějících dat. Zadáání práce bylo bezpochyby splněno.

**Příspěvek diplomanta.** Diplomant samostatně odvodil i ty výsledky, které už byly předtím publikovány. K tomu navrhl nový EM-algoritmus pro výpočet maximálně věrohodného odhadu při neznámých misklasifikačních parametrech za přítomnosti validační skupiny, odvodil odhad asymptotického roz-

ptylu, implementoval tyto metody v softwaru a provedl simulační studii. Tyto části práce lze dokonce považovat za potenciálně publikovatelný originální vědecký příspěvek.

**Matematická úroveň.** Matematická úroveň je výborná a zcela naplňuje požadavky na diplomovou práci.

**Formální úroveň.** Formální úroveň práce je vynikající. V práci nejsou závažné překlepy či typografické prohřešky.

S prací Martina Fardy jsem naprosto spokojen, považuji ji za vynikající a doporučuji ji uznat jako práci diplomovou.

doc. Mgr. Michal Kulich, PhD.

KPMS MFF UK

26. srpna 2024