



POLITECHNIKA WARSZAWSKA
WYDZIAŁ MATEMATYKI I NAUK INFORMACYJNYCH



PRACA DYPLOMOWA MAGISTERSKA
NA KIERUNKU MATEMATYKA

STATYSTYCZNE METODY REGRESJI PORZĄDKOWEJ

AUTOR:
MARTA SOMMER

PROMOTOR:
PROF. NZW. DR HAB.
PRZEMYSŁAW GRZEGORZEWSKI

WARSZAWA, CZERWIEC 2015

.....
podpis promotora

.....
podpis autora

Spis treści

Wstęp	5
1. Opis teoretyczny dostępnych metod	7
1.1. Metoda zaproponowana przez E. Franka i M. Halla	7
1.2. Sieci bayesowskie	7
Literatura	9

Wstęp

Regresja porządkowa (ang. *ordinal regression*) jest jednym z działów uczenia maszynowego. Od problemu klasycznej regresji różni ją to, że zmienna odpowiedzi jest dyskretna, natomiast od problemu klasyfikacji to, że zmienna odpowiedzi ma pewien naturalny porządek. Regresja porządkowa zajmuje się zatem uczeniem i oceną jakości predyktora, który modeluje zmienną uporządkowaną i skończoną. Problem regresji porządkowej rozwija się dość szybko m.in. dlatego, że ma on bardzo wiele zastosowań, choćby w systemach rekomendacji, czy bardzo popularnych wyszukiwarkach internetowych. Prześledźmy to na konkretnym przykładzie. Wyobraźmy sobie sytuację, że chcielibyśmy określić, w jakim stopniu danemu człowiekowi spodoba się sprzedawany przez nas produkt. Mamy do dyspozycji zbiór treningowy składający się z wektora zmiennej objaśniającej $\mathbf{x} = (x_1, \dots, x_d)$, gdzie x_i są różnymi cechami określającymi daną osobę (np. płeć, wiek, wykształcenie, ...). Cechy te – podobnie jak w przypadku zwykłej regresji – mogą być zarówno ciągłe, jak i dyskretne. Mamy również dostęp do zmiennej objaśnianej $\mathbf{y} = (y_1, \dots, y_r)$, będącej wektorem zero-jedynkowym, wskazującym która klasa została przypisana danemu rekordowi. W naszym przykładzie, zmienną odpowiedzi mogłyby być na przykład: *zdecydowanie mi się nie podoba*, *nie podoba mi się*, *nie mam zdania*, *podoba mi się*, *zdecydowanie mi się podoba*. Widać wyraźnie, że są one uporządkowane.

Najprostszym podejściem do tego typu problemu byłoby zignorowanie kolejności zmiennej odpowiedzi i potraktowanie go, jak zwykłą klasyfikację. W takim przypadku tracimy jednak pewną informację, która prawdopodobnie mogłaby przyczynić się do poprawy naszego klasyfikatora. Idąc w drugą stronę, można potraktować nasz problem, jak zwykłą regresję, zamieniając zmienną odpowiedzi na pewną zmienną ciągłą i to ją modelując, a następnie z powrotem dyskretyzować. Pojawia się tu jednak problem, jak optymalnie zrobić taką transformację, uwzględniając chociażby fakt, że nasza odpowiedź niekoniecznie jest monotoniczna (tzn. np. różnica między *nie podoba mi się* a *nie mam zdania* wcale mnie musi być taka sama, jak między *podoba mi się* a *zdecydowanie mi się podoba*).

Możemy wyróżnić dwa główne nurty w regresji porządkowej:

- prognoza konkretnej obserwacji (nacisk kładziony jest tu na wyznaczenie konkretnego \mathbf{y} dla konkretnego \mathbf{x} np. czy potencjalnemu klientowi spodoba się dany produkt),
- uszeregowanie kilku obserwacji (celem nie jest poznanie estymacji konkretnej zmiennej odpowiedzi, ale takie uszeregowanie kilku rekordów, by te najbardziej preferowane znalazły się na samej górze, a te najmniej na samym dole np. w jakiej kolejności powinny wyświetlić się znalezione strony w wyszukiwarce).

W mojej pracy zajmować się będę przede wszystkim pierwszym punktem, lecz nakreślę też kilka podejść dotyczących drugiego.

Rozdział 1

Opis teoretyczny dostępnych metod

W tym rozdziale opracuję kilka znanych i opisanych w literaturze podejść do regresji porządkowej. Można je podzielić na kilka grup:

- korzystające z dostępnych metod klasyfikacji, m.in.:
 - metoda zaproponowana przez E. Franka i M. Halla,
- modyfikujące dostępne metody klasyfikacji, m.in.:
 - SVM,
 - sieci neuronowe,
 - procesy gaussowskie,
- metody stworzone specjalnie dla regresji porządkowej, m.in.:
 - model proporcjonalnych szans.

1.1. Metoda zaproponowana przez E. Franka i M. Halla

Podejście Franka i Halla (por. [1]) do zagadnienia regresji porządkowej opiera się nie na stworzeniu nowego modelu, ale na odpowiednim przedefiniowaniu zbioru danych, a następnie na sprowadzeniu zadania do problemu zwykłej klasyfikacji z dwoma klasami. Przekształcamy zatem r -klasowy model regresji porządkowej do $(r - 1)$ -dwuklasowych problemów klasyfikacji.

bla bla bla... przepisać to, co było na slajdach ;D

1.2. Sieci bayesowskie

Sieci bayesowskie to bardzo proste i szeroko stosowane narzędzie zarówno w problemach regresji, jak i klasyfikacji. Znalazło również swoje zastosowanie w regresji porządkowej (por. [2]). Idea jest podobna do modelu proporcjonalnych szans.

Literatura

- [1] Frank E., Hall M., A simple approach to ordinal classification, *Proceedings of the European Conference on Machine Learning*, Freiburg, Niemcy, 2001, str. 146–156.
- [2] Cheng J., Wang Z., Pollastri G., A neutral network approach to ordinal regression, *Neutral Networks*, Hong Kong, 2008.

Marta Sommer
Nr albumu 237503

Warszawa, 20 kwietnia 2015

Oświadczenie

Oświadczam, że pracę magisterską pod tytułem „Statystyczne metody regresji porządkowej”, której promotorem jest prof. nzw. dr hab. Przemysław Grzegorzewski wykonałem samodzielnie, co poświadczam własnoręcznym podpisem.

.....
Marta Sommer