Statystyczna Analiza Danych – laboratorium

Regresja logistyczna

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 9
4/5 maja 2023

Idea zajęć – co i po co będziemy robić?

- Kontynuujemy temat **klasyfikacji** – na podstawie obserwowalnych cech (predyktorów) będziemy chcieli przyporządkować obiekty do k rozłącznych klas (wiemy, co to za klasy)
- Metoda KNN nie zwracała nam informacji o wpływie poszczególnych predyktorów na przyporządkowanie do klasy
- Regresja logistyczna zapewni nam predykcję do $k = 2$ klas oraz wyjaśnialność modelu.

Dyskretne zmienne zależne – ogólna idea

- Wyjaśniamy z wykorzystaniem zmiennych niezależnych prawdopodobieństwa stanów, korzystając z modelu dwumianowego

$$\mathbb{P}(y_i|x_i) = \begin{cases} 1 - p(x_i) & dla \quad y_i = 0 \\ p(x_i) & dla \quad y_i = 1 \end{cases}$$

- $p(x_i)$ będzie dystrybuantą rozkładu dla zmiennej ciągłej ($p(x_i) = F(x_i\beta)$)
- Modele szacować będziemy z wykorzystaniem MNW:

$$L(\theta) = \Pi_{i=1}^{N}[p(x_i)]_i^y[1 - p(x_i)]^{1-y_i}$$

Dyskretne zmienne zależne – zmienna ukryta

- Prawdopodobieństwo uzyskania wartości w zmienna 0-1, którą obserwujemy jest zależne od działania nieobserwowalnej *zmiennej ukrytej* (latentnej).
- To właśnie tę zmienną modelujemy – nie mamy dostępu do jej wartości, ale wiemy, że steruje prawdopodobieństwem uzyskania sukcesu ($y_i = 1$)
- Gdy ta zmienna przyjmie wartości przekraczające pewien próg (*próg odcięcia*), uzyskamy sukces
- Ze zmiennymi ukrytymi się już spotkaliśmy (zadanie o bogactwie mieszkańców)

Modele regresyjne dla zmiennej y binarnej

- $F(x_i\beta) = \frac{exp(x_i\beta)}{1+exp(x_i\beta)} = \Lambda(x_i\beta)$ – regresja logistyczna (logit)
- $F(x_i\beta) = \Phi(x_i\beta)$ – probit
- $F(x_i\beta) = x_i\beta$ – Liniowy Model Prawdopodobieństwa

Zadanie 1

- Będziemy chcieli przewidzieć, czy osoba, która otrzymała pożyczkę spłaci ją (1), czy nie (0) (problem scoringu kredytowego)
- Załaduj dane z pliku `pozyczka.csv`. Obejrzyj dane: określ, ile jest obserwacji, zmiennych, które zmienne są jakościowe, a które ilościowe
- Zbuduj model regresji logistycznej, który na podstawie wszystkich zmiennych poza pozyczka wykona predykcję dla zmiennej pozyczka. Zinterpretuj wydruk.

Logit w R

```
# skladnia podobna do KMRL, tylko tym razem UOGOLNIONY (generalized) model liniowy

model <- glm(pozyczka ~ ., dane, family = binomial)
summary(model)
```

Interpretacja + co można wyczytać z wydruku

- Interpretujemy znaki oszacowań
- Dokładniejsza interpretacja ilościowa – ilorazy szans
- Co się stało z R^2?...

Zadanie 1 – ocena jakości predykcji

- Wygeneruj wektor predykowanych wartości "0" (brak spłaty) i "1" (spłacone), przyjmując jako poziom odcięcia wartość prawdopodobieństwa 0.5.
- Oblicz średni błąd klasyfikatora (w tym zadaniu zbadamy jedynie błąd treningowy naszego klasyfikatora).
- Utwórz macierz błędu i oblicz czułość oraz specyficzność wykrywania udzielenia pożyczki. Zinterpretuj wyniki.
- Utwórz krzywą ROC

Dla zainteresowanych – rozszerzenia

- Uporządkowana zmienna dyskretna y – uporządkowany probit, uporządkowany logit
- Nieuporządkowana zmienna dyskretna y – wielomianowy logit
- Inne **modele wyborów dyskretnych** – warunkowy logit...

Zadanie #2 – do kontynuacji za tydzień (:

- Proszę stworzyć małe zespoły (ok. 3-osobowe)
- Będziemy pracować na zbiorze danych biopsy z pakietu MASS (załadować i obejrzeć zawartość zbioru)
- Proszę naszykować kody, które będą służyły klasyfikacji złośliwości nowotworu (zmienna class – 1 – nowotwór złośliwy, 0 – nowotwór łagodny), z wykorzystaniem wszystkich zmiennych w zbiorze poza ID
- Zakres technik: KNN, regresja logistyczna, KMRL, (dla chętnych) LDA
- Przygotować kody do oceny jakości predykcji (tablice klasyfikacyjne, obliczenia accuracy, precision, czułości, specyficzności)
- Naszykować krzywe ROC dla technik
- Która z technik najlepiej radzi sobie z klasyfikacją typu nowotworu?