Statystyczna Analiza Danych – laboratorium

Walidacja krzyżowa i bootstrap

Dorota Celińska-Kopczyńska

Uniwersytet Warszawski

Zajęcia 10
12/13 maja 2022

Idea zajęć – co i po co będziemy robić?

- Zajmując się modelami predykcyjnymi/klasyfikacyjnymi, chcemy, żeby te modele dobrze działały nie tylko na zbiorze treningowym, ale przede wszystkim na testowym
- Jednokrotne wylosowanie danych do zbioru testowego z reguły nie pozwala nam uogólnić, czy nasz model działa zadowalająco – mogliśmy trafić na "pasującą" próbkę, do tego, co już było
- Walidacja krzyżowa pozwala wygenerować wiele prób testowych, pozwalając nam wyciągnąć "odporniejsze" wnioski
- Przyjrzymy się też metodzie symulacyjnej, jaką jest bootstrap

Walidacja krzyżowa

- Dzielimy zbiór na rozkłaczne zbiory testowe.
- Dla każdego zbioru trenujemy model korzystając z pozostałych danych i oceniamy jego błąd
- k-krotna walidacja – każda obserwacja znajduje się tylko w jednym zbiorze testowym, może zaburzać estymację błędu
- walidacja MC – tworzymy wiele zbiorów testowych (np. po 10% obs), losując bez zwracania

Zadanie 1

- Zastosuj 10-krotną walidację krzyżową (implementację z pakietu `caret`) do zbadania błędu testowego dla klasyfikacji modelem logit na danych biopsy
- Wykorzystaj class jako zmienną y oraz wszystkie zmienne poza ID jako zmienne x. Usuń braki danych.

k-krotna walidacja krzyżowa z caret

```
library(caret)

# obiekt, ktory okesla, jak chcemy kontrolowac model
train_control <- trainControl(method='cv', number=k)

# wytrenowanie modelu i przetestowanie go
# prosze zwrocic uwage, ze pierwsze 4 argumenty sa prawie takie same jak w glm
kfold_train <- train(y ~ x, data=biopsy,
method='glm', family=binomial, trControl=train_control)

# podsumowanie
print(kfold_train)
```

Kryteria informacyjne

- Dobry model powinien spełniać dwa podstawowe warunki: być dobrze dopasowany do danych i możliwie jak najprostszy
- Kryteria informacyjne służą wyborowi modelu uwzględniając kary za złe dopasowanie i zbytnią złożoność:
  - $AIC = -\frac{2l(\hat{\theta})}{N} + \frac{2K}{N}$
  - $BIC = -\frac{2l(\hat{\theta})}{N} + \frac{K \log N}{N}$

  k to liczba parametrów w modelu, N to liczba obs, l logarytm funkcji wiarogodności dla oszacowanego MNW modelu
- Wybieramy model o najniższych wartościach kryteriów

Zadanie 2

- Utwórz klasyfikator typu nowotworu w oparciu o regresję logistyczną i dane biopsy.
- Przeprowadź wybór modelu poprzez minimalizację AIC oraz minimalizację BIC (korzystając z funkcji stepAIC)
- Porównaj modele poprzez k-krotną walidację krzyżową, z samodzielnie wybraną wartością parametru k.

Zadanie 3

- Oszacuj średnią zawartość cukru w landrynkach oferowanych przez Włodzimierza Bielskiego Meksykaninowi Tuco (plik walter.csv).
- Oszacuj wariancję estymatora średniej korzystając z metody bootstrap:
  - Napisz funkcję, która przyjmie oryginalne dane oraz wektor indeksów, których użyje do obliczenia średniej
  - Wylosuj ze zwracaniem 1000 prób zawierających pomiary cukru. Każda próba powinna być takiej samej liczności jak oryginalne dane
  - Wykorzystaj wylosowane próby, aby zbadać rozkład estymatora średniej.
- Porównaj swoje wyniki z funkcją `boot::boot`. Sposób jej użycia znajdziesz w dokumentacji.