Podstawy modelowania w R

Mateusz Staniak
eRementarz
Wrocław, 7.02.2020

Cele modelowania

Cele modelowania

Uczenie nadzorowane [znane etykiety / prawdziwe wartości]

- **przewidywanie** przyszłych cen akcji [regresja]
- **estymacja** wpływu leczenia na przebieg choroby [regresja]
- **identyfikacja** genów wpływających na występowanie choroby [regresja + selekcja zmiennych]
- modelowanie prawdopodobieństwa kliknięcia w reklamę przez użytkownika [regresja + klasyfikacja]

Uczenie nienadzorowane [brak znanych etykiet]

- **identyfikacja grup** klientów o podobnym profilu [klasteryzacja]
- **wykrywanie** oszustw przy transakcjach kartą płatniczą [anomaly detection]

Podejścia do modelowania

Każde z tych zadań wymaga innego podejścia do modelowania.

- Istotność (lub nie) założeń statystycznych.
- Ocena jakości modelu (podział na zbiór uczący i testowych, walidacja krzyżowa, bootstrap, statystyki jakości dopasowania).
  [Ryzyko **przeuczenia**]
- Imputacja brakujących danych.
- Redukcja wymiaru, korelacja między zmiennymi.
- Transformacja danych.

Jak zbudować model?

Regresja logistyczna

Modelujemy prawdopodobieństwo sukcesu dla ustalonego zestawu zmiennych objaśniających:

$$\Pr(G=1|X=x) = \frac{\exp(\beta_0+\beta^T x)}{1+\exp(\beta_0+\beta^T x)},$$
$$\Pr(G=2|X=x) = \frac{1}{1+\exp(\beta_0+\beta^T x)}. \qquad (4.1)$$

Here the monotone transformation is the *logit* transformation: $\log[p/(1-p)]$, and in fact we see that

$$\log\frac{\Pr(G=1|X=x)}{\Pr(G=2|X=x)} = \beta_0+\beta^T x. \qquad (4.2)$$

Elements of statistical learning

Model

- formuła definiująca model (alternatywnie dwie macierze)
- summary (i inne metody)

Zadania:

- dopasuj model ze zmienną oznaczającą rok,
- **BONUS**: dopasuj model z przekształceniami zmiennych

```
> glm_3var = glm(is_HS ~ budget + duration + coinvestigators,
+                data = hs_st_train)
> summary(glm_3var)

Call:
glm(formula = is_HS ~ budget + duration + coinvestigators, data = hs_st_train)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.95724  -0.37719  -0.08255   0.38156   1.34301

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.149e+00  4.288e-02  26.790  < 2e-16 ***
budget          -5.405e-07  2.212e-08 -24.436  < 2e-16 ***
duration        -1.254e-02  1.366e-03  -9.182  < 2e-16 ***
coinvestigators -7.792e-03  2.196e-03  -3.548 0.000393 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1835074)

    Null deviance: 962.39  on 3942  degrees of freedom
Residual deviance: 722.84  on 3939  degrees of freedom
AIC: 4510.4

Number of Fisher Scoring iterations: 2
```

Jak ocenić model?

Jakość modelu

- Miary statystyczne
- Macierz błędów:

		Przewidywana klasa	
		0 [ST]	1 [HS]
Prawdziwa klasa	0 [ST]	True negative [**TN**]	False positive [**FP**]
	1 [HS]	False negative [**FN**]	True positive [**TP**]

Miary oparte na macierzy błędów

- skuteczność (accuracy): (TP + TN) / (TP + TN + FP + FN)
- specyficzność (specificity): TN / (TN + FP)
- czułość (sensitivity, recall): TP / (TP + FN) [-> AUROC]
- precision: TP / (TP + FP) [-> precision-recall curve]
- F1: 2TP / (2TP + FN + FP)

Zadanie:

- napisać funkcję do obliczenia dowolnej z tych miar [lub kilku],
- na tej podstawie ocenić dopasowane modele.
- **BONUS**: zrobić to dla 1000 podziałów na zbiór uczący i testowy [**bootstrap**]

Więcej informacji

Pakiety R-owe

- **mlr**: Machine Learning in R - jednolitych interfejs do dziesiątek metod uczenia nadzorowanego i nienadzorowanego
- **broom**: jednolite przedstawienie i przetwarzanie wyników modelowania statystycznego
- **finalfit** i **modelsummary**: wizualizacje i tabele dla modeli liniowych
- **auditor**: diagnostyki dla modeli liniowych i nie tylko
- **gam** i **mgcv**: modele addytywne - pozwalające modelować zależności nieliniowe

Więcej informacji

- „biblia” modelowania predykcyjnego: *Elements of statistical learning* https://web.stanford.edu/~hastie/ElemStatLearn/
- o modelach regresyjnych: *Extending linear model with R* https://people.bath.ac.uk/jjf23/ELM/index.html
- statystyczne podejście: *Linear mixed effects models using R* https://link.springer.com/book/10.1007/978-1-4614-3900-4
- krytyka macierzy błędów w ocenie modelu: https://www.fharrell.com/post/mlconfusion/

Dziękuję za uwagę!