Metody analizy danych

Agnieszka Sołtys

Klasyfikacja

Przykłady zastosowań

- scoring kredytowy,
- predykcja churnu - przewidywanie na podstawie cech klienta (opisujących m.in. jego zachowanie) czy klient odejdzie w najbliższym czasie (np. w ciągu miesiąca),
- rozpoznawanie choroby - rozpoznawanie na podstawie parametrów medycznych czy pacjent zachoruje lub jest chory,
- klasyfikacja tematyczna tekstu,
- rozpoznawanie zawartosci obrazów.

Regresja logistyczna

- „Regresja” w nazwie pochodzi od podobieństwa do regresji liniowej, ale to model do klasyfikacji!
- Nazwa „logistyczna” pochodzi od funckji logit używanej w modelu.

Regresja logistyczna

Zakładamy, że obserwacje y są niezależne i pochodzą z rozkładu Bernoulliego (dwupunktowego):

$$Y \sim \mathbf{B}\left(p(x)\right), \; p(x) = \frac{1}{1 + exp(-x^T\beta)},$$

czyli:

$$Y = \begin{cases} 1 & z \; prawdopodobienstwem \; p(x) \\ 0 & z \; prawdopodobienstwem \; 1 - p(x) \end{cases}$$

Predykcja dla nowych danych:

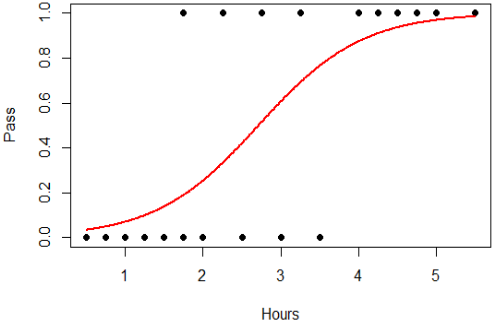
$$\widehat{y} = \begin{cases} 1 & jesli \; p(x) \geq \frac{1}{2} \\ 0 & jesli \; p(x) < \frac{1}{2} \end{cases}$$

Regresja logistyczna

$$p(x) = \frac{1}{1 + exp(-x^T\beta)} = \frac{1}{1 + exp(4.1 - 1.5x)}$$

Np. dla Hours $= 4$, $p(4) = 0.87$, $\widehat{y} = 1$.

Hours	Pass
1.75	1.00
2.00	0.00
2.25	1.00
2.50	0.00
2.75	1.00
...	...



Mamy dane n obserwacji (wielkość próby):

- X macierz planu dla p zmiennych objaśniających:

$$X = \begin{pmatrix} X_1 \\ \dots \\ X_n \end{pmatrix} = \begin{pmatrix} x_{11} & \dots & x_{1p} \\ \dots & \dots & \dots \\ x_{n1} & \dots & x_{np} \end{pmatrix}$$

- zmienną objaśnianą, $y_i \in \{0, 1\}$

$$y = \begin{pmatrix} y_1 \\ \dots \\ y_n \end{pmatrix}$$

Żeby dopasować model, trzeba znaleźć najlepsze wartości dla parametrów $\beta = \begin{pmatrix} \beta_1 \\ \dots \\ \beta_p \end{pmatrix}$. Estymatory $\widehat{\beta}$ metody największej wiarogodności maksymalizują fukcję wiarogodności (optymalizacja numeryczna):

$$\mathbf{L}(\beta) = \prod_{i=1}^{n} (p(X_i))^{y_i} (1 - p(X_i))^{1-y_i}$$

Regresja logistyczna - regularyzacja

Dodanie ograniczeń dla parametrów β (ściąganie).

- Regresja grzbietowa:

$$\widehat{\beta}^{ridge} = \arg\max_{\beta} \mathbf{L}(\beta), \text{ pod warunkiem } \sum_{j=1}^{p} \beta_j^2 \leq t.$$

- LASSO (least absolute shrinkage and selection operator): selekcja zmiennych

$$\widehat{\beta}^{LASSO} = \arg\max_{\beta} \mathbf{L}(\beta), \text{ pod warunkiem } \sum_{j=1}^{p} |\beta_j| \leq t.$$

- Elastic Net: połączenie regresji grzbietowej i LASSO