Wprowadzenie do modeli liniowych 2

Krystyna Grzesiak

Uogólniony model liniowy

$$g(\mathbb{E}(Y)) = \mathrm{X}\beta \tag{1}$$

Parametry:

- Y - wektor losowy o znanej realizacji
- X - znana macierz deterministyczna
- $\beta = (\beta_0, \beta_1, \ldots, \beta_p)$ - nieznany wektor deterministyczny
- $g : \mathbb{R}^n \to \mathbb{R}^n$ - znana funkcja różnowartościowa (funkcja linkująca/wiążąca)

Estymacja: MLE

Rodzina wykładnicza

Definicja: Rodziną wykładniczą nazywamy rodzinę rozkładów, których funkcje gęstości prawdopodobieństwa są postaci:

$$f_\theta(x) = C(\theta)e^{xQ(\theta)}h(x), \theta \in \Theta,$$

gdzie C i Q nie zależą od x, a h nie zależy od θ.

Jeśli w uogólnionym modelu zmienne $Y_1, \ldots, Y_n$ należą do rodziny wykładniczej, to funkcja Q jest kanoniczną funkcją linkującą dla tej rodziny rozkładów.

Przykłady kanonicznych funkcji wiążących

- Rodzina rozkładów zerojedynkowych:

$$f_p(k) = p^k(1-p)^{1-k} = (1-p)\left(\frac{p}{1-p}\right)^k = (1-p)e^{k\log\frac{p}{1-p}}$$

  Stąd mamy $Q(p) = \log\frac{p}{1-p}$ (tzw. logit).

- Rodzina rozkładów Poissona:

$$f_\lambda(k) = e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}e^{k\log\lambda}\frac{1}{k!}$$

  Stąd mamy $Q(\lambda) = \log\lambda$.

Regresja logistyczna

- $Y_i \sim b(1, p_i)$ gdzie $p_i \in (0, 1)$, $i = 1, \ldots, n$

Zauważmy, że przy powyższych założeniach $\mathbb{E}(Y_i) = p_i$ dla każdego $i = 1, \ldots, n$.
Stąd postać ogólnego wzoru modelu

$$g(\mathbb{E}(Y)) = \mathrm{X}\beta$$

dla regresji logistycznej jest następująca

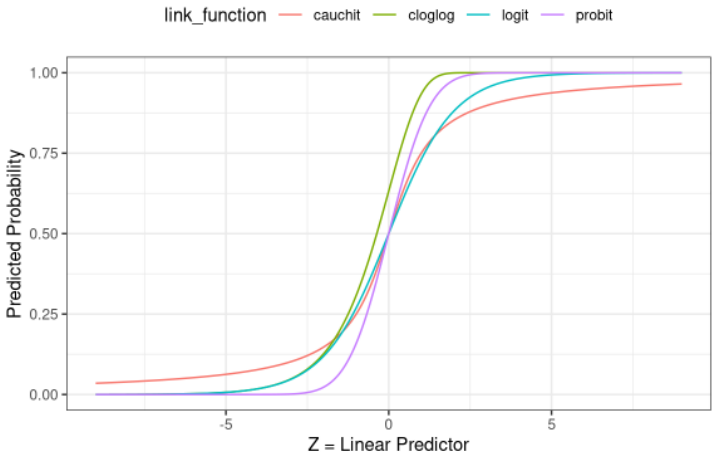
$$g(p) = \mathrm{X}\beta$$

gdzie $p = (p_1, \ldots, p_n)$.

Funkcje linkujące w modelu logistycznym

Name	Link Function	Response Probability	Properties
Logit	$z = \log\left(\frac{p}{1-p}\right)$	$p = \frac{e^z}{1+e^z}$	Coefficients explained using odds; canonical link for binary family
Probit	$z = \Phi^{-1}(p)$	$p = \Phi(z)$	Coefficients explained as impact on z-score for Normal distribution
Cauchit	Na	$p = \frac{1}{\pi}\arctan(z) + \frac{1}{2}$	Heavier tails than logit or probit
Cloglog	Na	$p = 1 - e^{-e^z}$	Inverse cdf of extreme value distribution; curv near probability of 1 is sharp

Funkcje linkujące w modelu logistycznym



Predykcja w modelu logistycznym

- **Predykcja** dotyczy wnioskowania na temat wartości oczekiwanej zmiennej Y: $\mathbb{E}(Y)$
- **Klasyfikacja** dotyczy "przydzielania"wartości zmiennej Y dla danego zestawu cech X na podstawie predykcji

Predykcja w modelu logistycznym

- **Predykcja** dotyczy wnioskowania na temat wartości oczekiwanej zmiennej Y: $\mathbb{E}(Y)$
- **Klasyfikacja** dotyczy "przydzielania"wartości zmiennej Y dla danego zestawu cech X na podstawie predykcji

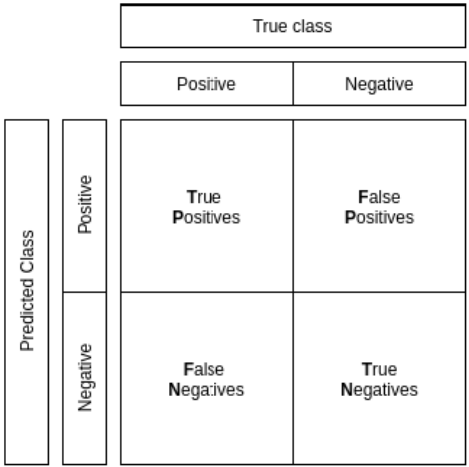
Klasyfikacja - confusion matrix

$$accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

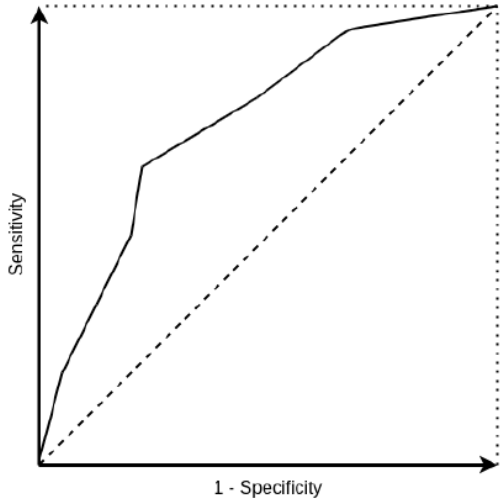
$$precision = \frac{TP}{TP + FP}$$

$$sensitivity = recall = \frac{TP}{TP + FN}$$

$$specificity = \frac{TN}{TN + FP}$$



ROC curve - Receiver Operating Characteristic curve



AUC - Area Under the ROC Curve

- AUC < 0.5
- AUC $= 0.5$
- AUC > 0.5

AUC może być używane do porównywania modeli.

Regresja Poissona

- $Y_i \sim Poiss(\lambda_i)$ $i = 1, \ldots, n$
- $\mathbb{E}(Y_i) = \lambda_i$
- $g(\lambda) = \log(\lambda)$

Formuła:

$$log(\lambda) = X\beta,$$

gdzie $\lambda = (\lambda_1, \ldots, \lambda_i)$.

Overdispersion - nadmierna dyspersja?

Przy założeniach rozkładu Poissona zachodzi $VarY = EY$.

Ocena:

- Czy w zbiorze danych znajdują się obserwacje różniące się znacznie od pozostałych?
- Czy $\frac{\mathbb{E}Y}{VarY} \approx 1$?
- Co mówią testy statystyczne?

Jak sobie z tym radzić?

- zero-inflated models
- zero-truncated models
- negative binomial distribution

Kaczy alarm

Bibliografia

- https://sdcastillo.github.io/PA-R-Study-Manual/glms-for-classification.html
- Tom Fawcett. "An introduction to ROC analysis". en. In: Pattern Recognition Letters. ROC Analysis in Pattern Recognition 27.8 (June 2006), pp. 861–874. issn: 0167-8655. doi: 10.1016/j.patrec.2005.10.010. url: https://www.sciencedirect.com/science/article/pii/S016786550500303X (visited on 07/26/2021).