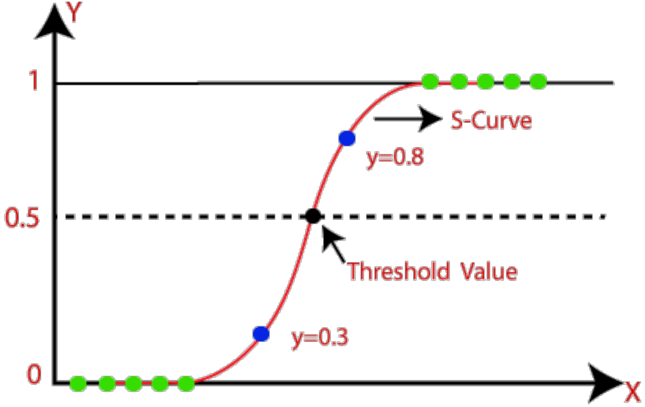
POLITECHNIKA GDAŃSKA

Regresja logistyczna oraz korelacja zmiennych objaśniających

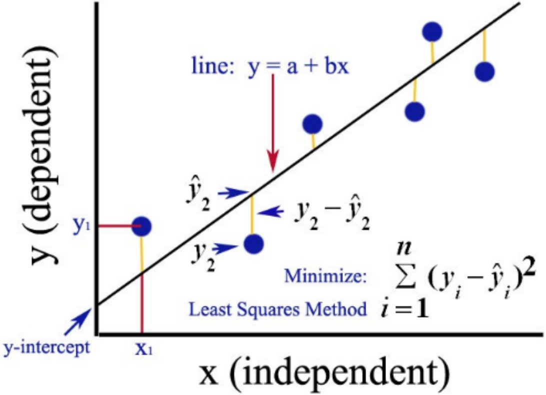
Jędrzej Smulski

18 maja 2022

- Należy do GLM.
- Problem klasyfikacji.
- Wyznacza prawdopodobieństwo wystąpienia danej klasy (wartości z przedziału $[0, 1]$).
- Linia dopasowana do danych jako predykcja.



- Metoda najmniejszych kwadratów.
- Współczynnik determinacji R^2.
- Wyznacza równanie prostej $y = a + b \cdot x$ (wartości z $\mathbb{R}$).



Szansa (ang. odds)

Wzór:

$$\frac{\text{Liczba zdarzeń gdzie występuje zdarzenie (np. drużyna wygra)}}{\text{Liczba zdarzeń gdzie występuje przeciwne zdarzenie (np. drużyna przegra)}}$$

Prawdopodobieństwo zajścia zdarzenia (np. drużyna wygra):

$$p = \frac{\text{Liczba zdarzeń gdzie występuje zdarzenie (np. drużyna wygra)}}{\text{Liczba wszystkich zdarzeń}}$$

Prawdopodobieństwo zajścia zdarzenia przeciwnego (np. drużyna przegra):

$$q = 1 - \text{Prawdopodobieństwo zajścia zdarzenia (np. drużyna wygra)} = 1 - p$$

Wtedy wzór na odds:

$$\frac{p}{1-p}$$

Szansa - przykład

Śzansa (odds) na wygranie meczu to 5 do 3":

$$odds = \frac{5}{3}$$

Prawdopodbieństwo wygrania meczu:

$$p = \frac{5}{5+3} = \frac{5}{8}$$

Prawdopodobieństwo przegrania meczu:

$$q = 1 - p = \frac{3}{8}$$

Wtedy:

$$\frac{p}{q} = \frac{p}{1-p} = \frac{\frac{5}{8}}{\frac{3}{8}} = \frac{5}{3}$$

Funkcja logit - po co?

Wraz ze wzrostem liczby zdarzenia przeciwnego wartość szansy spada i dąży do 0. Maksymalna wartość wynosi wtedy 1 - wystąpiło jedno zdarzenie przeciwne (przy zachowaniu wystąpienia jednego zdarzenia oczekiwanego).

Teraz przy zachowaniu jednego zdarzenia przeciwnego wraz ze wzrostem oczekiwanych zdarzeń osiągnięte zostaną wartości z przedziału $[1, \infty)$.

Wniosek: Występuje asymetria - trudność w porównaniu ze sobą szans.

Na przykład:

Szansa na wystąpienie zdarzenia przeciwnego wynosi 1 do 6, czyli: $1/6{=}0.17$.

Szansa na wystąpienie zdarzenia oczekiwanego wynosi 6 do 1 $=$ 6.

Biorąc logarytm naturalny (funkcja logit) z szansy pozbywa się asymetrii:

$$\ln\frac{1}{6} = -1.79 \qquad \ln\frac{6}{1} = 1.79$$

Powrót do regresji logistycznej

Transformacja z regresji logistycznej do regresji liniowej - transformacja z "prawdopodobieństwa zajścia zdarzenia"do "$\ln$(szansa na zajście danego zdarzenia)". Wtedy zmapowana zostanie wartość osi y z $[0, 1]$ na $(-\infty, \infty)$.

$$\ln(\text{szansa zajścia danego zdarzenia}) = \ln \frac{p}{1-p}$$

, gdzie p jest prawdopodobieństwem zajścia zdarzenia (powrót do slajdu 1). Podstawiając kolejno:

$$p = 0.5 \Rightarrow \ln \frac{p}{1-p} = \ln(1) = 0$$

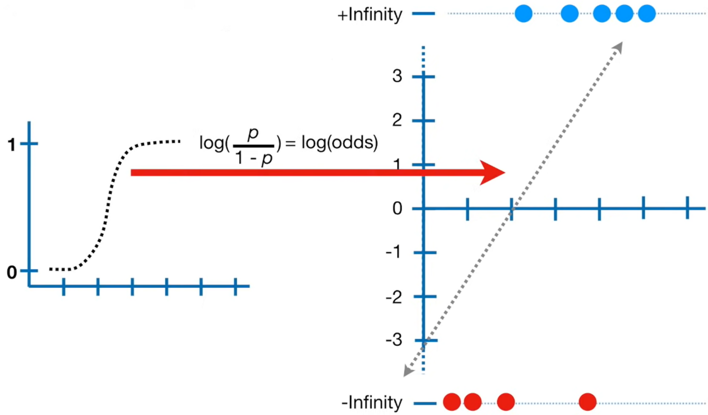
$$p = 0.731 \Rightarrow \ln \frac{p}{1-p} = \ln(2.717) = 1$$

$$p = 0.88 \Rightarrow \ln \frac{p}{1-p} = \ln(7.33) = 2$$

$$p = 1 \Rightarrow \ln \frac{p}{1-p} = \ln(1) - \ln(0) \rightarrow \infty$$

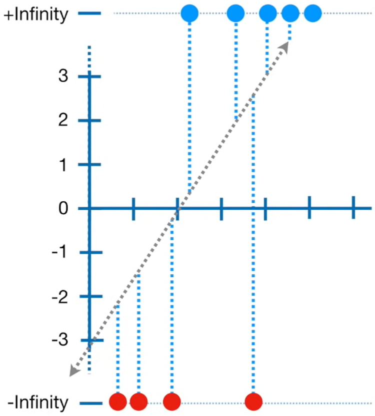
Ciąg dalszy

Transformacja na rysunku:



Ustalanie współczynników

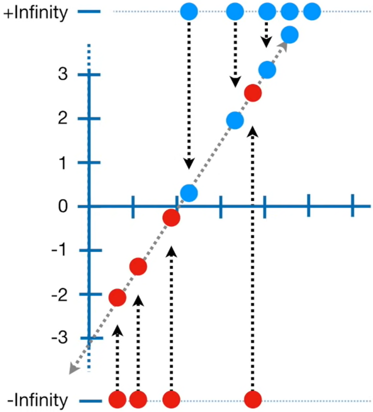
Zadaniem jest znalezienie najlepszej linii dopasowania do danych (najlepszej linii z prawego rysunku z poprzedniego slajdu).



Nie można użyć metody najmniejszych kwadratów.

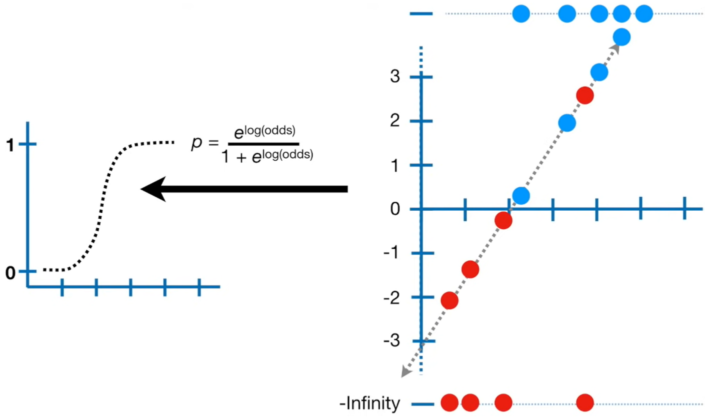
Maximum Likelihood

Używa się maximum likelihood. W pierwszym kroku zmapowane zostają wartości na prostą będącą kandydatem. w ten sposób dostaje się logarytm z szansy na zajście danego zdarzenia.



Maximum Likelihood

Następnie tworzy się krzywą predykcji modelu regresji logistycznej (wyprowadzenie p na kolejnym slajdzie).



Wzór na p

$$\ln\left(\frac{p}{1-p}\right) = \ln(odds)$$

$$\frac{p}{1-p} = \exp^{\ln(odds)}$$

$$p = (1-p)\exp^{\ln(odds)}$$

$$p = \exp^{\ln(odds)} - p\exp^{\ln(odds)}$$

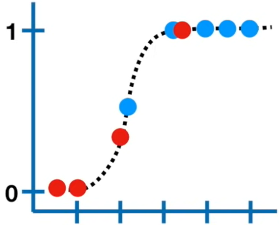
$$p + p\exp^{\ln(odds)} = \exp^{\ln(odds)}$$

$$p = \frac{\exp^{\ln(odds)}}{1+\exp^{\ln(odds)}}$$

Maximum Likelihood

Następnie liczy się iloczyn wartości* dla każdej z obserwacji.
Wartość obliczana w następujący sposób:

- dla zdarzenia (obserwacji) oczekiwanego: prawdopodobieństwo (wartości z osi y)
- dla zdarzenia (obserwacji) przeciwnego: 1 - prawdopodobieństwo (wartości z osi y)



Zazwyczaj zamiast maksymalizować likelihood to maksymalizuje się logarytm z likelihood. Wtedy zamiast iloczynu występuje suma zlogarytmizowanych wartości dla każdej z obserwacji.

Link function - Binomial

Weźmy rozkład dwumianowy z ustaloną/znaną liczbą prób (równą n):

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Logarytmujemy obie strony:

$$\ln(f(x)) = \ln\binom{n}{x} + x\ln(p) + (n-x)\ln(1-p)$$

Następnie nakładamy exp:

$$f(x) = \exp(x(\ln(p) - \ln(1-p)) + n\ln(1-p) + \ln\binom{n}{x})$$

Upraszczając:

$$f(x) = \exp\left(x\ln(\frac{p}{1-p}) + n\ln(1-p) + \ln\binom{n}{x}\right)$$

Link function - Binomial

Podstawiając:

$$\theta = \ln(\frac{p}{1-p})$$

Dostajemy:

$$e^{\theta} = \frac{1}{1-p}$$

$$p = \frac{e^{\theta}}{1+e^{\theta}}$$

$$(1-p) = \frac{1+e^{\theta}-e^{\theta}}{1+e^{\theta}} = \frac{1}{1+e^{\theta}} = (1+e^{\theta})^{-1}$$

$$n\ln(1-p) = -n\ln(1+e^{\theta})$$

$$f(x) = \exp\left(x\theta - n\ln(1+e^{\theta}) + \ln\binom{n}{x}\right)$$

Link function - Binomial

Zestawiając z (struktura funkcji z rodziny eksponentalnych):

$$f(y) = \exp(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi))$$

Dostajemy

$$f(x) = \exp\left(x\theta - n\ln(1 + e^{\theta}) + \ln\binom{n}{x}\right)$$

$$b(\theta) = n\ln(1 + e^{\theta})$$

$$a(\phi) = 1$$

$$c(x, \phi) = \ln\binom{n}{x}$$

Link function - Binomial

Następnie

$$\mu = b'(\theta) = n\frac{e^\theta}{1+e^\theta} \Rightarrow \frac{\mu}{n} = p$$

Wówczas link function:

$$\begin{aligned}
g(\mu) &= \theta \\
\frac{\mu}{n} &= \frac{e^\theta}{1+e^\theta} \\
\frac{\mu}{n}(1+e^\theta) &= e^\theta \\
\frac{\mu}{n} &= e^\theta - \frac{\mu}{n}e^\theta \\
\frac{\frac{\mu}{n}}{\frac{n}{n} - \frac{\mu}{n}} &= e^\theta \\
\ln\frac{\mu}{n-\mu} &= \theta
\end{aligned}$$

Funkcja logit jest link function dla rozkładu dwumianowego.

Link function - Poisson

Teraz wyprowadzona zostanie link function dla rozkładu Poissona.
Tak jak w przypadku rozkładu dwumianowego rozkład Poissona zostanie sprwadzony do rodziny funkcji eksponentalnych.

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!},$$

gdzie

- $\lambda > 0$
- $x = 0, 1, 2, \ldots$

Nakładamy logarytm:

$$\ln(f(x)) = x\ln(\lambda) - \lambda - \ln(x!)$$

Nakładamy funkcję exp:

$$f(x) = \exp(x\ln(\lambda) - \lambda - \ln(x!))$$

Link function - Poisson

Podstawiamy:

$$\theta = \ln(\lambda) \Rightarrow \lambda = e^{\theta}$$

Zestawiając z (struktura funkcji z rodziny eksponentalnych):

$$f(y) = \exp(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi))$$

Dostajemy

$$f(x) = \exp(x\theta - \lambda - \ln(x!))$$
$$b(\theta) = e^{\theta}$$
$$a(\phi) = 1$$
$$c(x, \phi) = -\ln(x!)$$

Link function - Poisson

Następnie

$$\mu = b'(\theta) = e^{\theta} = \lambda$$

Wówczas link function:

$$g(\mu) = \theta$$

$$\theta = \ln(\lambda) = \ln(\mu)$$

Wniosek:
Logarytm naturalny jest link function dla rozkładu Poissona.

Link function - Gamma

Teraz wyprowadzona zostanie link function dla rozkładu Gamma.
Tak jak w przypadku rozkładu dwumianowego i Poissona rozkład Gamma zostanie sprwadzony do rodziny funkcji eksponentalnych.

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \qquad x, \alpha, \beta > 0$$

Nakładamy logarytm:

$$\ln(f(x)) = \alpha \ln(\beta) - \ln(\Gamma(\alpha)) + (\alpha - 1)\ln(x) - \beta x$$

Nakładamy funkcję exp:

$$f(x) = \exp(-\beta x + \alpha \ln(\beta) + (\alpha - 1)\ln(x) - \ln(\Gamma(\alpha)) =$$

$$= \exp(\frac{-\frac{\beta}{\alpha} x + \frac{\alpha}{\alpha}\ln(\beta)}{\frac{1}{\alpha}} + (\alpha - 1)\ln(x) - \ln(\Gamma(\alpha)) =$$

$$= \exp(\frac{\frac{\beta}{\alpha} x - \ln(\beta)}{-\frac{1}{\alpha}} + (\alpha - 1)\ln(x) - \ln(\Gamma(\alpha))$$

Link function - Gamma

Podstawiamy:

$$\theta = \frac{\beta}{\alpha} \qquad \phi = \frac{1}{\alpha}$$

Wówczas:

$$\beta = \theta\alpha = \frac{\theta}{\phi}$$
$$\ln(\beta) = \ln(\theta) - \ln(\phi)$$

Link function - Gamma

Zestawiając z (struktura funkcji z rodziny eksponentalnych):

$$f(y) = \exp(\frac{\theta y - b(\theta)}{a(\phi)} + c(y, \phi))$$

Dostajemy

$$f(x) = \exp(\frac{-\theta x + \ln\theta}{\phi} + \frac{\ln(\phi)}{\phi} + (\frac{1}{\phi} - 1)\ln(x) - \ln(\Gamma(\frac{1}{\phi})))$$
$$b(\theta) = -\ln(\theta)$$
$$a(\phi) = \phi$$

Link function - Gamma

Następnie

$$\mu = b'(\theta) = -\ln(\theta) = -\theta^{-1}$$

Wówczas link function:

$$g(\mu) = \theta$$

$$\theta = -\mu^{-1}$$

Wniosek:
Funkcja odwrotna jest link function dla rozkładu Gamma.

Czemu korelacja jest użyteczna:

- pomaga predykować wartość zmiennej w oparciu o innej (przydatne np. w usuwaniu nulli w danych),
- może wskazać na istnienie związku przyczynowego,
- jest wykorzystywana jako jedna z podstawowych metryk przy badaniu danych.

Im korelacja większa (o ile jest większa od zera) tym większa zależność między danymi - wraz ze wzrostem jednej wartości rośnie też druga.
Im korelacja mniejsza (o ile jest mniejsza od zera) tym większa zależność między danymi - wraz ze wzrostem jednej wartości maleje druga.
Korelacja oscylująca wokół zera mówi o braku zależności między zmiennymi.

Korelacja

Multicollinearity występuje wtedy gdy zmienne objaśniające w modelu ze sobą korelują liniowo.
W niektórych modelach duża korelacja między zmiennymi nie jest przeszkodą. Na przykład biorąc do drzewa losowego (model drzewiasty) silnie skorelowane zmienne, to model skorzysta tylko z jednej z nich.
A w jakich może stanowić problem:

- Regresja liniowa,
- Regresja logistyczna,
- KNN,
- SVM with Linear Kernel,
- klasteryzacja oparta o odległości,
- Naive Bayes.

Jak sobie radzić z tym problemem? PCA wydaje się być ok.

Bibliography

- *StatQuest*, https://www.youtube.com/channel/UCtYLUTtgS3k1Fg4y5tAhLbw.
- https://en.wikipedia.org.
- https://www.youtube.com/channel/UCTqZYOHuGOqJkwaUmZthsPg.
- http://sfb649.wiwi.hu-berlin.de/fedc_homepage/xplore/tutorials/xlghtmlnode38.html
- https://stats.stackexchange.com/questions/369985/deriving-the-canonical-link-for-a-binomial-distribution
- https://towardsdatascience.com/why-feature-correlation-matters-a-lot-847e8ba439c4
- https://datascience.stackexchange.com/questions/12554/does-xgboost-handle-multicollinearity-by-itself

POLITECHNIKA
GDAŃSKA