Klasyfikowanie zmiennej jakościowej za pomocą regresji logistycznej

Marzena Nowakowska
Wydział Zarządzania i Modelowania Komputerowego

Metody modelowania nadzorowanego

- klasyfikacja nadzorowana

  Zmienna objaśniana jest cechą opisową

- analiza regresji

  Zmienna objaśniana jest cechą liczbową, ciągłą

- analizie przetrwania

  Zmienna objaśniana jest czasem do zajścia pewnego zdarzenia

Istnieje regresja logistyczna, która jest metodą używaną do klasyfikacji nadzorowanej i odwołuje się do metod i pojęć analizy regresji

Zadania regresji logistycznej

Regresja logistyczna pozwala dokonać prognozy dyskretnej wartości pewnej zmiennej (określić prawdopodobieństwo przynależności do jakiejś kategorii zmiennej dyskretnej) na podstawie znajomości zbioru wartości wielu zmiennych niezależnych, które mogą mieć charakter ciągły lub dyskretny (wielowartościowy lub dwuwartościowy).

Pojęcie szansy
w badaniach statystycznych

Iloraz prawdopodobieństwa sukcesu π do prawdopodobieństwa porażki (1- π) oznaczony przez *odds* nosi nazwę szansy

$$odds = \pi /(1- \pi)$$

gdzie π jest prawdopodobieństwem sukcesu

Prawdopodobieństwo sukcesu (porażki) jest liczone z próby i oznaczane symbolem *p* (*1-p*).

Szansa w przykładzie

Niech będzie dana grupa pacjentów badana na okoliczność wystąpienia u nich udaru mózgu. Jest to grupa 200 mężczyzn po czterdziestce, obciążonych dziedziczne chorobami krążenia, którzy nałogowo palą papierowy. W tej grupie liczba osób, u których wystąpił udar mózgu jest równa 136.

- Szacowane prawdopodobieństwo wystąpienia udaru mózgu u mężczyzny należącego do takiej grupy jest równe:
  $p = 136/200 = 68\%$.
- Szansa na to, że mężczyzna należący do takiej grupy jest narażony na udar mózgu jest równa:
  $p/(1-p) = 68\%/(1-68\%) = 2{,}13$
  – szansa sukcesu ponad 2 x większe niż porażki

Iloraz szans

Iloraz szans pojawia się najczęściej, gdy bada się dwie grupy na okoliczność pewnego zdarzenia, przy czym grupy te różnią się wartością określonej cechy. W każdej z tych grup można określić szansę sukcesu oznaczoną odpowiednio *odds1* i *odds2*. Iloraz tych wartości nosi nazwę wskaźnika szans lub ilorazu szans IR (*odds ratio – OR*):

$$IR = OR = odds1/odds2$$

Iloraz szans w przykładzie

	Atak serca		
Grupa	**Nie (porażka)**	**Tak (sukces)**	**Razem**
Grupa 1: *Placebo*	10845	189	11034
Grupa 2: *Aspiryna*	10933	104	11037

Szacowanie sukcesu:

p1 = 189/11034 = 0.0171 *p2 = 0.0094.*

Szacowane szanse ataku serca są równe:

p1/(1-p1)= 189/10845=0.0174

p2/(1-p2)= 104/10933=0.0095

Szacowany iloraz szans: *0.0174/0.0095=1.832*

Szanse ataku serca w grupie pierwszej są blisko 2 razy większe niż w grupie drugiej

Przypadek klasyczny regresji logistycznej – prognozowanie sukcesu

Zmienna zależna Y może przyjąć dwie wartości, zazwyczaj nazywane sukcesem (wartość 1) i porażką (wartość 0)

$$\pi = \exp(U)/(1+\exp(U))$$

gdzie:

$$\pi = P(Y=1|\,[X_1, X_2,,..., X_k])$$

$$U = \beta_0 + \sum_{j=1}^{k} \beta_j X_j$$

k jest liczbą zmiennych objaśniających

Funkcja logistyczna dla dwuwartościowej zmiennej objaśnianej

Logit szans $\longrightarrow \ln(\pi/(1-\pi)) = \beta_0 + \sum_{j=1}^{k} \beta_j X_j$

Iloraz szans dla dwuwartościowej zmiennej objaśnianej:

$$OR(X_j) = \frac{P(Y=1 \mid X_j = x_{j2}) / P(Y=0 \mid X_j = x_{j2})}{P(Y=1 \mid X_j = x_{j1}) / P(Y=0 \mid X_j = x_{j1})}$$

$OR(X_j) = \exp(\beta_j)$

Szansa, że zmienna objaśniana przyjmie wartość sukcesu zmienia się w skali $\exp(\beta_j)$, gdy wartość objaśniającej zmiennej ilościowej X_j zwiększa się o jeden

$OR(X_{j\,w_i vs.wo}) = \exp(\beta_j)$

Szansa, że zmienna objaśniana przyjmie wartość sukcesu jest dla wartości w_i jakościowej zmiennej objaśniającej X_j:

- $\exp(\beta_j)$ razy większa, gdy $\exp(\beta_j) > 1$,
- o $\{(1-\exp(\beta_j))*100\%\}$ procent mniejsza, gdy $\exp(\beta_j) < 1$

niż dla wartości odniesienia *wo* tej zmiennej (tzn. zmiennej X_j)

Oceny ilorazu szans (Odds ratio estimates)

Iloraz szans mówi, jaki jest wpływ wybranej zmiennej objaśniającej na szansę sukcesu przy nie zmieniających się wartościach dla pozostałych zmiennych objaśniających (predyktory).

- Dla zmiennej objaśniającej ilościowej *X* wartość *IS(X)* mówi w jakiej skali zmieniają się (zwiększają lub zmniejszają) szanse sukcesu jeżeli wartość zmiennej *X* zwiększy się o 1, przy ustalonych wartościach pozostałych predyktorów.
  Przykład: *IS*(*wiek*) = *1,2* – szansa sukcesu zwiększa się o 20% gdy wiek zwiększa się o jeden rok
- Dla zmiennej objaśniającej jakościowej *X* wartość $IS(X_{kategoria\ 1\ vs\ kategoria\ o})$ mówi w jakiej skali zmieniają się (zwiększają lub zmniejszają) szanse sukcesu dla *kategorii 1* cechy X w stosunku do *kategorii o* tej cechy, przy ustalonych wartościach pozostałych predyktorów.
  Przykład: $IS(płeć_{K\ vs.M}) = 0,34$ – szansa sukcesu dla kobiet jest taka jak 0,34 szans dla mężczyzn, co oznacza, że szansa sukcesu dla kobiet jest o 66% mniejsza niż dla mężczyzn

Ocena modelu logistycznego

Model regresji logistycznej jest weryfikowany w oparciu o:

- testowanie istotności modelu
- testowanie istotności zmiennych wejściowych (niezależnych, predyktorów) modelu (często mówi się o nich efekty)
- testowanie parametrów strukturalnych modelu

Testowanie hipotez (przypomnienie):

1. Zdefiniowane hipotezy zerowej H0 i hipotezy alternatywnej H1
2. Wyznaczenie wartości statystyki testowej *WST* na podstawie próby i odpowiadającej jej wartości prawdopodobieństwa testowego (*p-value = p-wartość*)
3. Porównanie *p-wartości* z α lub położenia statystyki testowej względem obszaru krytycznego OK i podjęcie decyzji odnośnie hipotezy H0:
   Jeśli p-value < α ($\Leftrightarrow$ WST $\in$ OK) H0 należy odrzucić

Testowanie istotności modelu

Test ilorazu wiarygodności globalnej hipotezy zerowej: BETA=0

(1)

Porównanie modelu tylko z wyrazem wolnym (stałą) oraz modelu z wyrazem wolnym i zmiennymi objaśniającymi:

Jaki jest model M_P w porównaniu z modelem M_0 ?

gdzie: M_0 model tylko z wyrazem wolnym,

M_P model z wyrazem wolnym i ze zmiennymi objaśniającymi (predyktorami)

Statystyka -2LOGL – Log-Likelihood

Jeżeli $-2LOGL(M_0) > -2LOGL(M_P)$ to M_P jest lepszy niż M_0.

(2)

Testowanie hipotezy zerowej, że wszystkie parametry regresji są zerami wobec hipotezy alternatywnej, że co najmniej jeden parametr jest istotnie różny od zera

H0: BETA= 0

H1: BETA ≠ 0

Statystyka testowa Chi-kw ilorazu wiarygodności (Likelihood Ratio Chi-Square) ma rozkład chi-kwadrat

Jeśli *p-value* < α to H0 należy odrzucić

Testowanie istotności efektów

Analiza efektów typu 3 (*Type 3 Analysis of Effects*)

Tablica podaje:

- nazwę efektu (zmiennej wejściowej),
- liczbę stopni swobody LLS (DF – Degree of Freedom),
- statystykę testową dla efektu,
- prawdopodobieństwo testowe

Testowana hipoteza zerowa mówi, że wpływ efektu na zmienną celu jest nieistotny.

Statystyka testowa Wald Chi-Square ma asymptotyczny rozkład chi-kwadrat.

Jeśli prawdopodobieństwo testowe *p-value* < α, to H0 należy odrzucić, co oznacza że badana zmienna wejściowa jest istotna statystycznie (ma wpływ na zmienną objaśnianą (zależną, celu))

Ocena parametrów strukturalnych modelu

Analiza ocen maksymalnej wiarygodności (*Analysis of Maximum Likelihood Estimates*)

Tablica podaje między innymi:

- nazwę zmiennej wejściowej (zmienna jakościowa jest kodowana), przy której stoi parametr,
- liczbę stopni swobody (DF),
- wartość estymatora parametru strukturalnego i jego błąd standardowy,
- wyniki testu (statystka testowa oraz prawdopodobieństwo testowe),

Treść testowanej hipotezy zerowej jest następująca:

$H0: \beta_i=0$

Statystyka testowa dla hipotezy to statystyka Walda

Jeśli prawdopodobieństwo testowe *p-value* < α to H0 należy odrzucić, co oznacza że parametry strukturalne są istotne statystycznie (istotnie różnią się od zera).

Ocena ilorazów szans oraz przedział ufności dla ilorazu szans (również wskazanie na istotność wyniku).

Tabela klasyfikacji

	Prognozowane porażki (ujemne)	Prognozowane sukcesy (dodatnie)	Liczebności obserwowane
Obserwowane porażki (negatywne)	**A** (prawdziwie negatywne)	B (fałszywie pozytywne)	A + B
Obserwowane sukcesy (pozytywne)	C (fałszywie negtywne)	**D** (prawdziwie pozytywne)	C + D

Prawdziwie ujemne (True Negative Rate) TNR= A / (A + B)
TNR = (Liczba porażek sklasyfikowanych jako porażki)/(Ogólna liczba porażek)

Fałszywie dodatnie (False Positive Rate) FPR = B / (A + B)
FPR = (Liczba porażek sklasyfikowanych jako sukcesy)/(Ogólna liczba porażek)

Fałszywie ujemne (False Negative Rate) FNR = C / (C + D)
FNR = (Liczba sukcesów sklasyfikowanych jako porażki)/(Ogólna liczba sukcesów)

Prawdziwie dodatnie (True Positive Rate) TPR = D / (C + D)
PD = (Liczba sukcesów sklasyfikowanych jako sukcesy)/(Ogólna liczba sukcesów)

Poprawne klasyfikacje: (A + D) / (A + B + C + D)
Błędne klasyfikacje: (B + C) / (A + B + C + D)

Zasięg i precyzja – ocena jakości klasyfikacji

	Prognozowane porażki (ujemne)	Prognozowane sukcesy (dodatnie)	Liczebności obserwowane
Obserwowane porażki (negatywne)	**A** (prawdziwie negatywne)	B (fałszywie pozytywne)	A + B
Obserwowane sukcesy (pozytywne)	C (fałszywie negtywne)	**D** (prawdziwie pozytywne)	C + D

Miary zasięgu (pokrycia, dotarcia)

Specyficzność (Specificity) - Prawdziwie ujemne (True Negative Rate)

$$TNR = A / (A + B)$$

Czułość Sensitivity - Prawdziwie dodatnie (True Positive Rate)

$$TPR = D / (C + D)$$

Miary precyzji

Precyzja przewidywania negatywnego (Negative Predictive Value)

$$NPV = A / (A + C)$$

NPV = (Liczba porażek sklasyfikowanych jako porażka)/ (Ogólna liczba prognozowanych porażek)

Precyzja przewidywania pozytywnego (Positive Predictive Value)

$$PPV = D / (D + B)$$

PPV = (Liczba sukcesów sklasyfikowanych jako sukcesy)/(Ogólna liczba prognozowanych sukcesów)