Predicting League of Legends match outcome with Bayesian methods

Z. Nagelj

April 21, 2019

1 Uvod

Logisticna regresijska se uporablja za analizo povezavo med kategoricno odvisno spremenljivko in neodvisnimi spremenljivkami.Polege logisticne regresije se za analizo kategoricnih spremenljivk uporablja diskriminantna analiza, ki za razliko od logisticne regresije predpostavlja normalno porazdelitev neodvisnih spremenljivk.

2 Logisticna regresija

Kot vhodni podatek za logisticno regresijo dobimo podatkovni set N tock. Vsaka i-ta tocko sestavlja set m-tih neodvisnih spremenljivk in kategoricna odvisna spremenljivka Y_i z dvema možnima izidoma.

2.1 Logit in logisticna transormacija

Naši kategoricni odvisni spremenljivki najprej dodelimo numericne vrednost (0 in 1). Povprecje na vzorcu predstavlja delež ugodnih izidov p, razmerje $p/(1-p)$ pa obeti (odds). Logit transormacijo definiramo kot logaritem obetov (log odds):

$$l = logit(p) = \log \frac{p}{1-p}$$

S pomocjo te transformacije preidemo iz omejene zaloge vrednosti p na intervalu $[0, 1]$, na obete $p/(1-p)$ omejene z zalogov vrednosti $[0, \infty)$ in na koncu na logaritem obetov z zalogo vrednosti $(-\infty, \infty)$. Inverzno transormacijo imenujemo logisticna transformacija:

$$p = \text{logistic}(l) = \frac{\exp l}{1 + \exp l}$$

S transformacijo se izognemo problemu omejene zaloge vrednosti odvisne spremenljivke. Potencialno bi lahko izbrali tudi kakšno drugo transormacijo (probit).

2.2 Logisticni model

Kategoricno odvisno spremenljivko definiramo kot slucajno spremenljivko $Y_i$porazdeljeno po Bernoulliju s pricakovano vrednostjo p_i. Vsak izid je torej dolocen s svojo neznano verjetnostjo p_i, ki je dolocena na podlagi neodvisnih spremenljivk.

$$Y_i|x_{1,i},...,x_{m,i} \sim \text{Bernoulli}(p_i)$$
$$E[Y_i|x_{1,i},...,x_{m,i}] = p_i$$
$$P(Y_i = y|x_{1,i},...,x_{m,i}) = p_i^y(1-p_i)^{(1-y)}$$

Ideja je zelo podobna kot pri linearni regressiji, torej verjetnost p_i modeliramo kot linearno kombinacijo neodvisnih spremenljivk. Razlika je v tem, da verjetnosti transformiramo s pomocjo logit funkcije. V modelu nastopa dodaten intercept clen, zato imamo $m+1$ regresijskih koficientov β.

$$\text{logit}(E[Y_i|X_i]) = logit(p_i) = \log \frac{p_i}{1-p_i} = X_i\beta$$

Oziroma:

$$E[Y_i|X_i] = p_i = \text{logit}^{-1}(X_i\beta) = \frac{1}{1+\exp^{-X_i\beta}}$$
$$P[Y_i = y_i|X_i] = p_i^y(1-p_i)^{1-y} = \frac{\exp^{y_iX_i\beta}}{1+\exp^{X_i\beta}}$$

2.3 Dolocanje vrednosti regresijski koficientov

Regresijski koficienti in verjetnosti p_i so dolocene optimizacijo, na primer MLE. Za lažjo predstavo si najprej ogledamo MLE za enostaven primer Bernoullija:

2.4 MLE za Bernoulli(p)

Zapišemo enacbo za verjetje in jo logaritmiramo.

$$L = \prod_{i=1}^{n} p^{Y_i}(1-p)^{1-Y_i} = p^{\sum_{i=1}^{n} Y_i}(1-p)^{n-\sum_{i=1}^{n} Y_i}$$

$$l = \log(L) = \sum_{i=1}^{n} Y_i \log(p) + (n - \sum_{i=1}^{n} Y_i)\log(1-p)$$

Cenilko za $\hat{p}$ dolocimo s prvim parcialnim odvodom, ki ga enacimo z 0. Za asimptotski interval zaupanja cenilke dolocimo Fisherjevo informacijsko matriko. Saj populacijske vrednosti p ne poznamo Fisherjevo informacijo dolocimo z oceno $\hat{p}$.

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$
$$I(p) = E[-\frac{\partial^2 l}{\partial p^2}] = \frac{1}{p(1-p)}$$
$$(\hat{var}) = \frac{1}{nI(\hat{p})} = \frac{\hat{p}(1-\hat{p})}{n}$$
$$(\hat{SE}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Za velik n je naša cenilka porazdeljena približno normalno:

$$\hat{p} \sim_{CLI} \text{Normal}(p, \frac{p(1-p)}{n})$$

2.5 MLE za Bernoulli(p_i)

Saj ima pri logisicne regresije vsak izid svojo verjetnost p_i je naša enacba za verjetje naslednja:

$$L = \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i} = p_i^{\sum_{i=1}^{n} Y_i}(1-p_i)^{n-\sum_{i=1}^{n} Y_i} = (\frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}})^{\sum_{i=1}^{n} Y_i}(\frac{1}{1+\exp^{X_i\beta}})^{n-\sum_{i=1}^{n} Y_i}$$

$$l = \log(L) = \sum_{i=1}^{n} Y_i \log(\frac{\exp^{X_i\beta}}{1+\exp^{X_i\beta}}) + (n - \sum_{i=1}^{n} Y_i)\log(\frac{1}{1+\exp^{X_i\beta}})$$

Posamezne regresijske koficiente dobimo s parcialnim odvodom logaritma verjetja. Saj rešitev v zakljuceni formi ne obstaja uporabimo numerice metode (npr. Newtnova metoda). Prav tako dolocimo informacijsko matriko za dolocanje asimptotske kovariancne matrike in intervalov zaupanja.

3 Statisticno testiranje regresijskih koficientov

Za testiranje hipotez nenicelnosti regresijski koficientov sta v uporabi Waldov test in test razmerja verjetij.

3.1 Waldov test

Waldov test se uporablja za dolocanje statisticne znacilnosti posameznih regresijskih koficientov (podobno kot t-test pri linearni regresiji). Za koficiente pridobljene z MLE je testna statistika naslednja:

$$Z = \frac{\hat{\beta}_i}{\hat{SE}}$$

Nicelelna hipoteza testira $H_0 : \beta_i = 0$. Kot velja za vse cenilke pridobljene po metodi najvecjega verjetja so te asimptotsko nepristranske (dosledne) in normalno porazdeljene okoli prave vrednosti z varianco $\frac{1}{nI(p)}$.

3.2 Test razmerja verjetij

Pri testu nas zanima logaritem Wilksovega lambda pri dveh razlicnih modelih, pri cemer je en model gnezden (podmnožica drugega). Primerjali bomo polni model s **k** regresijskimi koficienti in delni model z **m** regresijskimi koficienti, kjer je **m** < **k**. Nicelna hipoteza (delni model) trdi da so testirani regresijski koficienti enaki 0. Alternativna hipoteza (polni model) pa trdi, da so vsi regresijski koficienti razlicni od 0. Pod nicelno hipotezeo torej testiramo nicelnost k - m β. Naša testrna statistika je:

$$LR = -2\log\Lambda = -2\log\frac{L(H_0)}{L(H_A)} = -2(l(H_0) - l(H_A)) = -2(l(\hat{\beta}^{(0)}) - l(\hat{\beta}))$$

Asimptotsko je ta porazdeljena s χ^2 z k - m stopinjami prostosti. Test razmerja verjetij se od Waldovega testa razlikuje po tem, da je potrebno narediti dva modela pod razlicnimi hipotezami.

4 Naloga

4.1 Opis

Na vzorcu bolnikov z rakom primerjamo dve vrsti operacije. Zanima nas ali obstaja povezanost s stadijem bolnika in ali ena vrsta operacije povzroca manj zapletov. Cilj naloge je analizirati napako I. reda pri testiranja hipotez na nacin, da testiramo vsako spremenljivko posebaj.

4.2 Postopek testiranje

Sledili bomo naslednjim korakom:

4.2.1 Stadij

1. Generiranje podatkov pod dano hipotezo
2. Izdelava 4 modelov logisticne regresije z informacijo o posameznem stadiju
3. Pridobimo p-vrednosti Waldovega testa vseh 4 modelov z delnimi podatki
4. Izdelava 1 modela logisticne regresije z informacijo o vseh stadijh (1 spremenljivka)
5. Pridobimo p-vrednosti testa razmerja verjetij za model z vsemi podatki
6. Analiziramo porazdelitve p-vrednosti in primerjamo deleže zavrnjenih hipotez

4.2.2 Zaplet

1. Generiranje podatkov pod dano hipotezo
2. Izdelava 10 modelov logisticne regresije z informacijo o posameznem zapletu
3. Izdelava 1 modela logisticne regresije z informacijo o vseh zapletih (10 spremenljivk)
4. Pridobimo p-vrednosti Waldovega testa vseh 11 modelov
5. Analiziramo porazdelitve p-vrednosti in primerjamo deleže zavrnjenih hipotez

4.3 Pricakovani rezultati

Zagotovo, pricakujemo, da bo naveden nacin testiranja pri stadiju napacen, saj so stadiji med seboj neodvisni, cesar v modelih ne zajamemo. Prav tako v model ne vkljucimo informacije o ostalih stadijih in jih obravnavamo kot enakovredne. Pravilno bi stadij tako, da ga v celoti vkljucimo v model ter z testom razmerja verjetij testiramo nenicelnost regresijskega koficienta.
Pri obravnavi zapletov operacij, ob predpostavki da so si te neodvisni, predvidevam, da tak nacin testiranja ne bi bil napacen. V realnost pa temu zagotovo ni tako in so posamezni zapleti med seboj povezani, zato bi zagotovo naleteli na enake probleme kot pri testiranju posameznega stadija.

5 Podatki

5.1 Generiranje

Saj je v nalogi doloceno, da je vzorec velik 300 pacientov, kjer vsaka polovica prejme en tip operacije, najprej generiramo vecji vzorec $k = 10000$ bolnikov iz katerga bomo vzorcili. Pridobiti moramo vrednosti spremenljivke stadij in desetih spremenljivk zaplet.

5.1.1 Stadij

Definiramo populacijske verjenosti za vsak stadij (Tabela 1), ter glede na njih s funkcijo sample izžrebamo stadij vsakega bolnika in dolocimo modelsko matriko. Verjetnosti stadijev se morajo sešteti v 1, saj so med seboj odvisni.

	verjenost
1	0.60
2	0.25
3	0.10
4	0.05

Table 1: Populacijski verjenosti posameznega stadija raka

5.1.2 Zaplet

Definiramo populacijske verjenosti za vsak zaplet (Tabela 2), ter glede na njih s funkcijo rbinom izžrebamo ali se je posamezen zaplet zgodil.

	verjenost
Stadij 1	0.10
Stadij 2	0.12
Stadij 3	0.14
Stadij 4	0.16
Stadij 5	0.18
Stadij 6	0.20
Stadij 7	0.30
Stadij 8	0.40
Stadij 9	0.50
Stadij 10	0.60

Table 2: Populacijski verjenosti posameznega zaplete pri operaciji

5.1.3 Vzorec

Ko imamo dolocene vrednosti spremenljiv na podlagi definirane hipoteze (oz. regresijskih koficientov) dolocimo linearne kombinacije spremenljivk ter s pomocjo logit transformacije verjetnost za tip operacije za vsakega izmed k bolnikov. Izbran tip operacij pridobimo iz Bernoullijeve porazdelitve, glede na p_i.
Da zadostimo specifikacijam naloge iz vzorca 10000 bolnikov ob vsaki izmed $m = 1000$ iteracij nakljucno izžrebamo 150 operacij vsakega tipa.

5.2 Hipoteze

Podatke generiramo glede na 2 hipoteze, nicelno in alternativno. Glede na nicelno hipotezo so vsi regresijski koficienti enaki 0. Vrednosti regresijski koficientov pod alternativno hipotezo so vidni v Tabeli 3.

	stadij	zaplet
beta0	0.00	0.00
beta1	0.00	-3.50
beta2	0.10	3.00
beta3	0.60	-2.50
beta4	-1.20	2.00
beta5		-1.50
beta6		1.00
beta7		-0.80
beta8		0.60
beta9		-0.40
beta10		0.20

Table 3: Vrednosti regresijski koficientov pri HA glede na neodvisno spremenljivko