Логистичка регресија

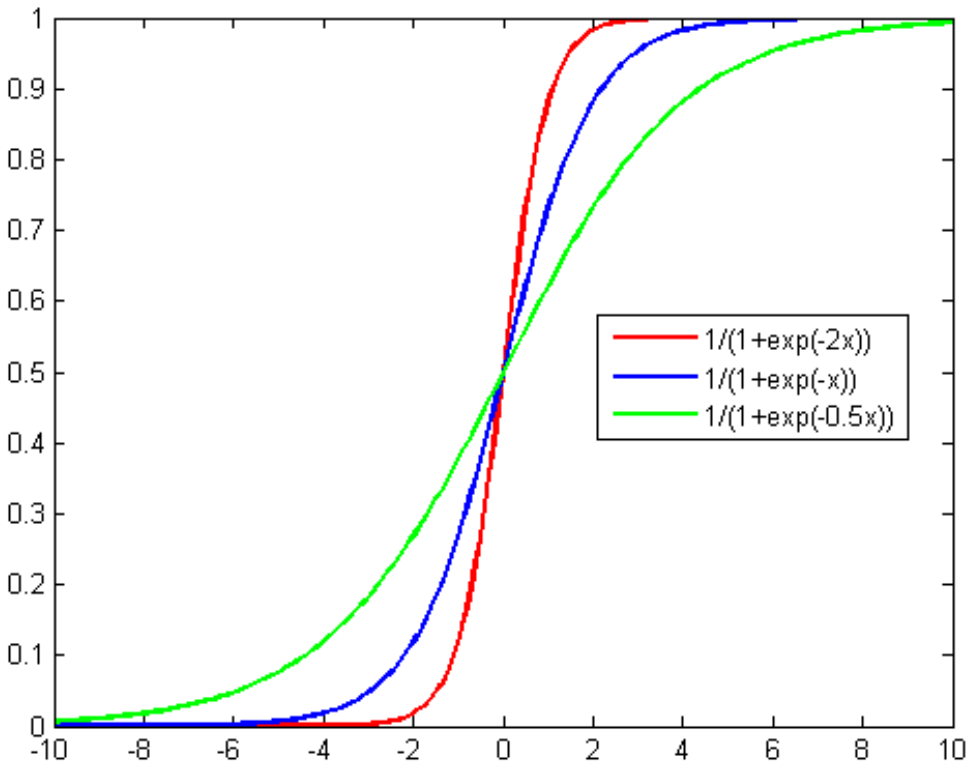
4.час

22. март 2016.

Логистичка расподела

Логистичка расподела је непрекидна расподела вероватноће таква да је њена функција расподеле логистичка функција

$$f(x) = \frac{1}{1 + e^{-\frac{x-m}{s}}}.$$



Историјат

Логистичка функција је настала у 19. веку за потребе моделовања раста различитих популација.

Наиме, различити истраживачи су се још у 18. веку бавили проучавањем и предвиђањем раста популације у некој земљи.

Овај проблем се своди на проучавање неке количине $W(t)$ која, на пример, може да буде величина људске популације у временском тренутку t и њеног прираштаја у јединици времена који се означава са $W'(t)$

$$W'(t) = \frac{dW(t)}{dt}$$

Историјат

Најједноставнија претпоставка која је коришћена у науци још у 18. веку је била да је $W'(t)$ пропорционално са $W(t)$, односно да постоји нека константна β за коју важи

$$W'(t) = \beta W(t), \quad \beta = \frac{W'(t)}{W(t)}.$$

Решавањем ове диференцијалне једначине се долази до закључка да је раст популације експоненцијалан, односно да постоји нека константа A за коју важи

$$W(t) = Ae^{\beta t},$$

где се за A често узима величина популације у почетном тренутку посматрања $W(0)$.

Овај модел се показао као добар при проучавању младих популација, као што је на пример, популација САД-а у првим деценијама по њиховом настанку.

Историјат

Међутим, белгијски математичар Келте (Alphonse Quetelet 1795-1874) и његов млађи сарадник математичар Велхурст (Pierre Francois Velhurst 1804-1849) су приметили да овакво решење после неког времена доводи до нереалних процена и да би требало ограничити прираштај популације на неки начин.

Они су у претходну диференцијалну једначину додали елемент $\phi(W(t))$ који представља отпор популације према даљем расту у тренутку t:

$$W'(t) = \beta W(t) - \phi(W(t)).$$

Историјат

Велхурст је затим експериментисао са различтим облицима за $\phi(W(t))$ и дошао на идеју да уведе константу Ω која би представљала горњу границу засићености за W. Прираштај популације би тада био пропорционалан и тренутној величини, али и њеном простору за даљи раст $\Omega - W(t)$

$$W'(t) = \beta W(t)(\Omega - W(t)).$$

Увођењем смене $P(t) = \frac{W(t)}{\Omega}$ у претходу једначину добијамо диференцијалну једначину

$$P'(t) = \beta P(t)(1 - P(t)),$$

а њено решење је облика

$$P(t) = \frac{e^{\alpha+\beta t}}{1 + e^{\alpha+\beta t}}.$$

Ову функцију је Велхурст назвао логистичком функцијом.

Историјат

Ова истраживања нису привукла велику пажњу математичке јавности.

Тек захваљујући развојем рачунара у другој половини 20. века логистичка расподела стиче широку популарност. Њена предност је у једноставном облику и повољним аналитичким својствима који је чине погодном за израчунавање уз помоћ различитих алгоритама.

Данас је логистичака расподела најпознатија по својој примени у моделима логистичке регресије. Осим тога користи се и у хидрологији за моделовање водостаја, у теорији полупроводника и на многим другим местима.

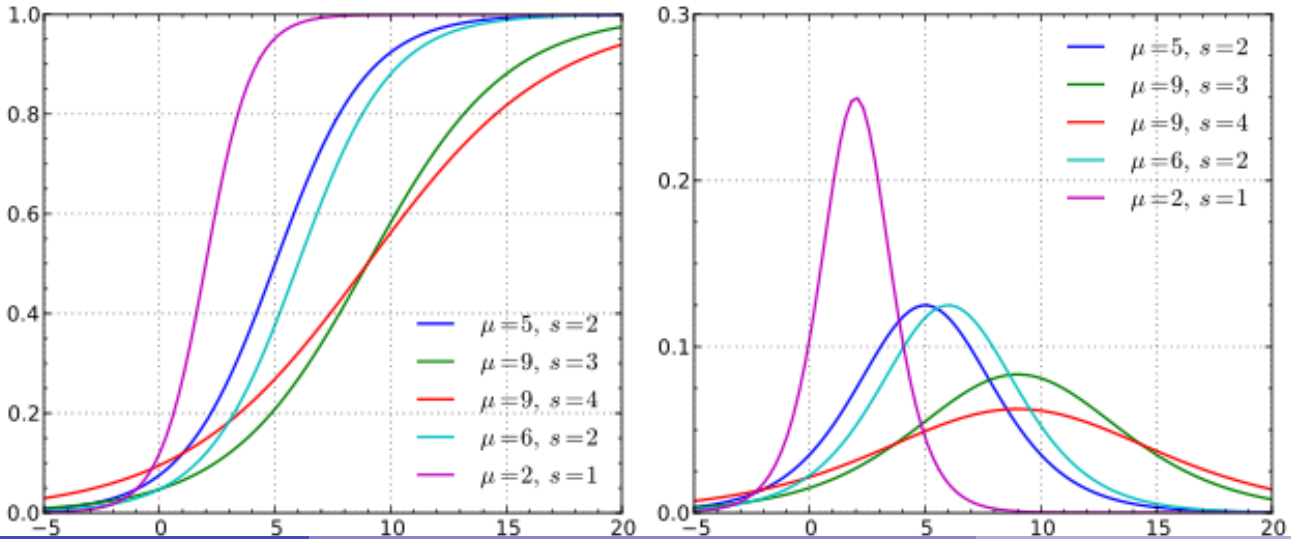
Особине логистичке расподеле

Логистичка расподела је симетрична расподела тешких репова.

Ако је X случајна величина са логистичком расподелом, тада X има следећу функцију и густину расподеле:

$$F(x) = \frac{1}{1 + e^{-\frac{x-m}{s}}}, s > 0, m \in \mathbb{R}, x \in \mathbb{R}.$$

$$f(x) = \frac{e^{-\frac{x-m}{s}}}{s\left(1 + e^{-\frac{x-m}{s}}\right)^2}, s > 0, m \in \mathbb{R}, x \in \mathbb{R}.$$



Особине логистичке расподеле

- **Очекивање**
  $EX = m$
- **Медијана**
  $\mu = m$
- **Мод**
  $mod = m$
- **Дисперзија**
  $DX = \frac{s^2\pi^2}{3}$
- **Коефицијент симетрије**
  $\gamma_1 = 0$
- **Коефицијент спљоштености**
  $\gamma_2 = \frac{6}{5}$
- **Карактеристика расподеле**
  $\ln(1 + e^{-\frac{x-m}{s}}) \sim \epsilon(1)$

Основни модел логистичке регресије

Нека је X независна случајна променљива на основу које треба предвидети вредности за Y и нека Y може да има само две вредности, $\Omega_Y = \{0, 1\}$.

Уместо директног предвиђања којој ће класи припадати Y, идеја логистичке регресије је оцењивање вероватноће да Y припадне свакој од класа ако је вредност за X позната.

Дакле, треба проценити следеће вероватноће:

$$P\{Y = 1|X\}, \quad P\{Y = 0|X\}.$$

Основни модел логистичке регресије

Ако уведемо ознаку $p(X) = P\{Y = 1|X\}$, тада се проблем своди на оцењивање вредности $p(X)$.

Како $p(X)$ представља неку вероватноћу, потребно је да функција којом се ова вредност моделира буде непрекидна, монотона и да узима вредности између 0 и 1.

Многе функције са овим особинама, а у логистичкој регресији се користи логистичка функција облика

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}, \beta_0, \beta_1 \in \mathbb{R}, \beta_1 \neq 0.$$

Једноставном транфсформацијом добијамо једнакост

$$\frac{p(X)}{1 - p(X)} = e^{\beta_0+\beta_1 X}.$$

Основни модел логистичке регресије

Израз $\frac{p(X)}{1-p(X)}$ се назива квотом и може узимати вредности између 0 и ∞.

Квоте се чешће користе од вероватноћа користе у моделовањима система за клађење јер је интуитивније приказују шансе добитка:
вредности близу 0 одговарају веома малим шансама и зато што је вредност квоте већа, то је већа и шанса позитивног исхода клађења.

Применом природног логаритма на претхидну једначину добијамо:

$$\ln\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

Лева страна једначине се назива `logit` трансформацијом од $p(X)$.
Приметимо да је веза између `logit` трансформације и независне променљиве X линеарна.

Оцењивање параметара

Модел логистичке регресије зависи од параметара β_0 и β_1 које је потребно оценити. Оцењивање се врши методом максималне веродостојности. Случајна величина Y у зависности од X има расподелу

$$Y|X : \begin{pmatrix} 0 & 1 \\ 1-p(X) & p(X) \end{pmatrix}$$

која може да се напише и као

$$f(Y|X) = p(X)^Y (1-p(X))^{1-Y}, Y \in \{0,1\}, X \in \mathbb{R}.$$

Функција максималне веродостојности параметара на основу узорка обима n је

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} p(X_i)^{Y_i} (1 - p(X_i)^{1-Y_i}.$$

Оцене $\hat{\beta}_0$ и $\hat{\beta}_1$ параметара β_0 и β_1 се добијају као решења система

$$\frac{\partial \ln L(\beta_0, \beta_1)}{\partial \beta_0} = 0, \quad \frac{\partial \ln L(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

Валдов тест

По добијању оцена за параметре, потребно је проверити да ли је X заправо добар предиктор за вредности за Y, односно да ли постоји статистички значајна повезаност.

Валдовим тестом се тестирају следеће хипотезе

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0.$$

Хипотеза H_0 проверава се формирањем Валдове тест статистике

$$Z^* = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)}$$

која при важењу H_0 има стандардну нормалну расподелу, где је $\hat{\sigma}^2(\hat{\beta}_1)$ оцена стандардне девијације оцене $\hat{\beta}_1$.

Предикција

Када су параметри модела оцењени, оцена вредности $p(X)$ се једноставно добија из формуле

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}.$$

Класификација променљиве Y се затим врши на основу $\hat{p}(X)$

$$\hat{Y} = \begin{cases} 0, & \hat{p}(X) < q \\ 1, & \hat{p}(X) \geq q \end{cases}$$

где је q унапред одређена константа.

Стандардна вредност за q је $\frac{1}{2}$, али постоје и случајеви у којима се узимају друге вредности.

Вишеструка логистичка регресија

Нека су $\mathbf{X} = (X_1, X_2, ..., X_N)$ случајни вектор и Y случајна променљива квалитативног типа која узима вредности из скупа $G = \{G_1, G_2, ..., G_M\}$ и зависна је од случајног вектора $\mathbf{X}$.

Модел логистичке регресије дефинишемо на следећи начин:

$$P\{Y = G_i|\mathbf{X}\} = \frac{e^{\beta_{i0}+\beta_i^T X}}{1 + \sum_{j=1}^{M-1} e^{\beta_{j0}+\beta_j^T X}}, \quad i \in \{1, 2, ..., M-1\},$$

$$P\{Y = G_i|\mathbf{X}\} = \frac{1}{1 + \sum_{j=1}^{M-1} e^{\beta_{j0}+\beta_j^T X}}.$$

где су $\beta_{10}, ..., \beta_{(M-1)0}$ неки реални бројеви и $\beta_1, ..., \beta_{M-1}$ неки N-димензиони вектори. Сви појмови уведени раније важе и овде.

Вишеструка логистичка регресија

Други начин је да се вишеструка логистичка регресија поједностави и сведе на примену неколико основних логистичких регресија.

Овај метод се назива "сам против свих" (one-vs-all) и његова суштина је да се креира M одвојених класификатора који само процењују да ли променљива Y припада некој одређеној класи из G или не.

Ово се постиже увођењем помоћних случајних променљивих $Z_1, ..., Z_M$ које служе као индикатори да ли Y припада одређеној класи из G:

$$Z_i = I\{Y = G_i\} = \begin{cases} 0, & Y \neq G_i \\ 1, & Y = G_i \end{cases}$$

За свако $i = 1, 2, ..., M$

Вишеструка логистичка регресија

За овако уведене променљиве Z_i се формирају модели основне логистичке регресије оцењивањем вредности

$$\hat{p}_i(X) = P\{Z_i = 1|X\}.$$

На основу тих модела променљивој Y се додељује класа за коју је $\hat{p}_i(X)$ највеће

$$\hat{Y} = \{G_i|\hat{p}_i(X) = \max_{j\in\{1,2,\ldots,M\}} \hat{p}_j(X)\}.$$

На овај начин се постижу ефекти вишеструке логистичке регресије без сложене примене у пракси коју она захтева.

Пример

Подаци садрже резултате два испита са пријемног испита на једном универзитету у Америци и информацију да ли је студент примљен на мастер програм.

На основу ових резултата направити модел логистичке регресије који може да предвиди да ли ће студент бити примљен на мастер.

База садржи 100 опсервација и 3 променљиве (`Exam1`, `Exam2`, `Admitted`).

Променљиве које садрже резултате испита су реални бројеви између 0 и 100, а трећа променљива има само две класе - 0 или 1.

Наш модел ће имати облик

$$P\{Admitted = 1 | Exam1, Exam2\} = \frac{e^{\beta_0 + \beta_1 Exam1 + \beta_2 Exam2}}{1 + e^{\beta_0 + \beta_1 Exam1 + \beta_2 Exam2}}.$$

Пример

```
> head(baza)
```

```
     Exam1    Exam2 Admitted
1 34.62366 78.02469        0
2 30.28671 43.89500        0
3 35.84741 72.90220        0
4 60.18260 86.30855        1
5 79.03274 75.34438        1
6 45.08328 56.31637        0
```

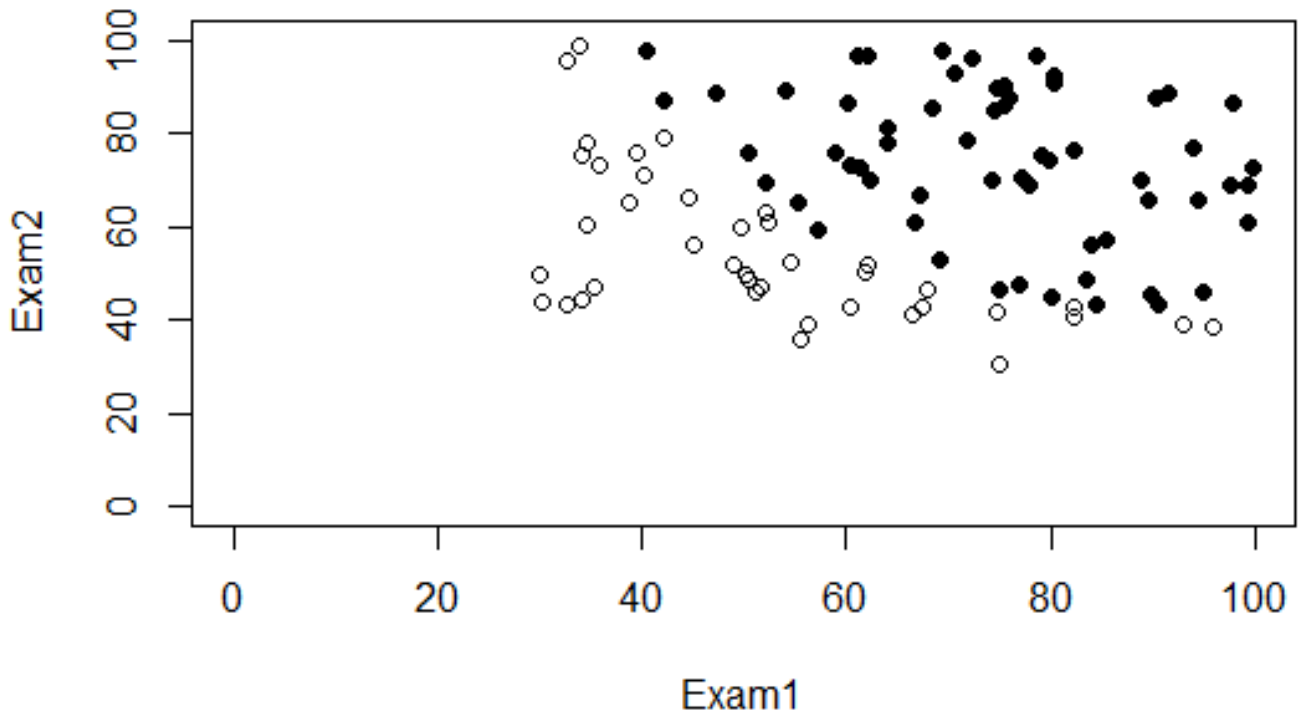
```
> summary(baza)
```

```
     Exam1           Exam2          Admitted
 Min.   :30.06   Min.   :30.60   Min.   :0.0
 1st Qu.:50.92   1st Qu.:48.18   1st Qu.:0.0
 Median :67.03   Median :67.68   Median :1.0
 Mean   :65.64   Mean   :66.22   Mean   :0.6
 3rd Qu.:80.21   3rd Qu.:79.36   3rd Qu.:1.0
 Max.   :99.83   Max.   :98.87   Max.   :1.0
```

Пример

Више информација може нам дати графичко приказивање података. Приказаћемо све податке на графику тако што ће свака тачка имати координате које представљају резултате једног и другог испита, а тип тачке ће носити информацију о укупном успеху на пријемном испиту.

```
> plot(Exam1[Admitted==0], Exam2[Admitted==0], xlab = Exam1", ylab = "Exam2", xlim
= c(0,100), ylim=c(0,100))"
> points(Exam1[Admitted==1], Exam2[Admitted==1], pch = 20)
```



Пример

```
> model <- glm(Admitted~Exam1+Exam2, family = binomial)
> summary(model)


Call:
glm(formula = Admitted ~ Exam1 + Exam2, family = binomial)
Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.19287  -0.18009   0.01577   0.19578   1.78527
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -25.16133    5.79836  -4.339 1.43e-05 ***
Exam1         0.20623    0.04800   4.297 1.73e-05 ***
Exam2         0.20147    0.04862   4.144 3.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
(Dispersion parameter for binomial family taken to be 1)
    Null deviance: 134.6  on 99  degrees of freedom
Residual deviance:  40.7  on 97  degrees of freedom
AIC: 46.7
Number of Fisher Scoring iterations: 7
```

Пример

Коефицијенти модела:

```
> coef(model)
(Intercept)       Exam1       Exam2
-25.1613335   0.2062317   0.2014716
```

Интервали поверења за параметре:

```
> confint(model)
Waiting for profiling to be done...
                  2.5 %      97.5 %
(Intercept) -38.9918822 -15.7757315
Exam1         0.1279764   0.3204597
Exam2         0.1221850   0.3168368
```

Пример

Како су p вредности Валдових тестова веома мале, закључујемо да се хипотезе да је неки од параметара једнак нули одбацују.

Тражени модел је:

$$\hat{P}\{Admitted = 1 | Exam1, Exam2\} = \frac{e^{-25.16+0.21 \cdot Exam1+0.20 \cdot Exam2}}{1 + e^{-25.16+0.21 \cdot Exam1+0.20 \cdot Exam2}}.$$

Пример

Предвиђање на основу добијемог модела:

```
> p.X <- predict(model, type = ”response” )
> y <- rep(0, length(Admitted))
> y[p.X>0.5] <- 1
```

Проверимо да ли ће студент уписати мастер студије ако положи један испит са 50 поена, а други са 80.

```
> newdata <- data.frame(Exam1=50, Exam2=80)
> predict(model,newdata,type = ”response”)
 1
0.7803968
```

Задатак

1. Из базе података `mtcars` (пакет `MASS`) издвојити променљиве `vs`, `mpg` и `am` у нову базу. Наћи најбољи логистички модел за који су `mpg` и `am` независне променљиве помоћу којих предвиђамо зависну променљиву `vs` и предвидети вредности за `vs`.