МАТЕМАТИЧКИ ОМНИБУС, 9 (2021), 167 – 176
http://im-pmf.weebly.com/matematicki-omnibus-kniga-9.html

ЛОГИСТИЧКА РЕГРЕСИЈА И НЕЈЗИНА ПРИМЕНА НА ЗАДАЧИ ОД БИНАРНА КЛАСИФИКАЦИЈА

Филип Николовски [1]

Статистиката како дел од математиката (или како независна наука) меѓу другото се занимава и со опишување на зависностите помеѓу променливите кои ги разгледува. Променливите главно спаѓаат во една од две категории: **независни**, врз кои директно имаме влијание, и **зависни** на чиишто вредности не можеме директно да влијаеме; нивната вредност се менува исклучиво на основа на вредностите на независните променливи. Во контекст на барањата и очекувањата, оваа зависност може да се опише на повеќе начини. Два од нив се: опишувањето на степенот (јачината, моќта) на зависност и конструкција на модел со кој директно би можеле да извршиме предвидување на вредностите на зависната променлива врз основа на вредности на независните променливи. Овој пристап на анализа на зависностите помеѓу променливите главно спаѓа во областа на статистиката наречена *регресиона анализа*. Најчесто користени алатки од оваа област се **коефициентот на корелација *r*** (како мерка за степенот на линеарна зависност помеѓу променливите) и **моделот на линеарна регресија** кој претпоставува постоење на линеарна зависност. Меѓутоа постојат ситуации кога линеарниот регресионен модел не е соодветен, па затоа прибегнуваме кон модели од друг тип. Во овој труд се опишува ситуација која води кон конструкција на таканаречениот модел на **логистичка регресија**. Со помош на овој модел можеме да ги искористиме добрите страни на линеарниот модел, но да ги примениме на типови задачи кои во основа не се во неговиот домен.

Во првата точка накусо ќе се осврнеме на линеарниот регресионен модел, неговата конструкција и неговите особености. Ова ќе ни даде можност да ги согледаме предностите што тој ги има над другите модели. Понатаму, се разгледува задача на бинарна класификација во која зависната променлива не е од нумерички, туку од категориски тип и ќе се обидеме да конструираме модел за класификација преку моделирање на веројатностите со помош на линеарна регресија. Недостатоците на

овој пристап ќе ги исправиме во третата точка каде формулираме модел на логистичка регресија за моделирање на веројатностите преку нивна (привремена) трансформација во *шанси*. Во последната точка даваме едноставен начин за оценка на прецизноста/точноста на конструираниот модел.

1. ЛИНЕАРЕН РЕГРЕСИОНЕН МОДЕЛ

Наједноставниот нетривијален модел кој може да се конструира при работа со податоци е моделот на **линеарна регресија**. Како што кажува самото име, овој модел претпоставува дека врската помеѓу независните променливи (кои се „влезни“) и зависната променлива („излезната“) е линеарна. Да претпоставиме дека сме собрале податоци за n независни променливи во m серии на набљудувања, а во секоја од сериите сме собрале податоци и за независната променлива y. Податоците може да ги структуираме во матрици X и y, на независни и зависни променливи, соодветно, дадени со:

$$X = \left[x_{ij} \right]_{m \times n} \qquad \text{и} \qquad y = \left[y_i \right]_{m \times 1}$$

каде што x_{ij} го означува i-то набљудување за j-та независна променлива, а y_i го означува i-то набљудување за независната променлива. Со помош на овие ознаки линеарниот модел кој ја опишува зависноста на зависната од независните променливи може да се формулира како:

$$y = Xw + \varepsilon \tag{1}$$

каде што со ε е означена случајна компонента која ја содржи грешката во моделот. „Конструкцијата“ на моделот се сведува на наоѓање на коефициентите $w = [w_j]_{n \times 1}$, а откако ќе ги пресметаме овие коефициенти целта ни е да вршиме предвидувања за вредностите на зависната променлива со помош на матричната равенка $\hat{y} = Xw$. Јасно е дека предвидените вредности $\hat{y}$ и вистинските вредности y ќе се разликуваат. Во пракса $m \gg n$ (имаме многу повеќе податоци од број на независни променливи), па наоѓањето на коефициентите w е малку посложено од решавање систем линеарни равенки. Затоа за решавање на оваа задача го користиме *методот на најмали квадрати* кој ни дава таков вектор на

коефициенти w којшто го минимизира вкупното квадратно отстапување помеѓу вистинските вредности на независната променлива, y, и предвидувањата, $\hat{y}$, направени на основа на вредностите на независните променливи и коефициентите. Така, задачата ја сведуваме на задачата на оптимизација без ограничувања:

$$\min_{w\in\mathbb{R}^n} \left\| \hat{y} - y \right\|^2 = \min_{w\in\mathbb{R}^n} \left\| Xw - y \right\|^2 \tag{2}$$

Оваа задача, во општ случај, би морало да се реши со помош на некој од нумеричките методи за оптимизација, но имајќи ја предвид линеарната структура на моделот, во случајов може да се добие експлицитно, затворено решение:

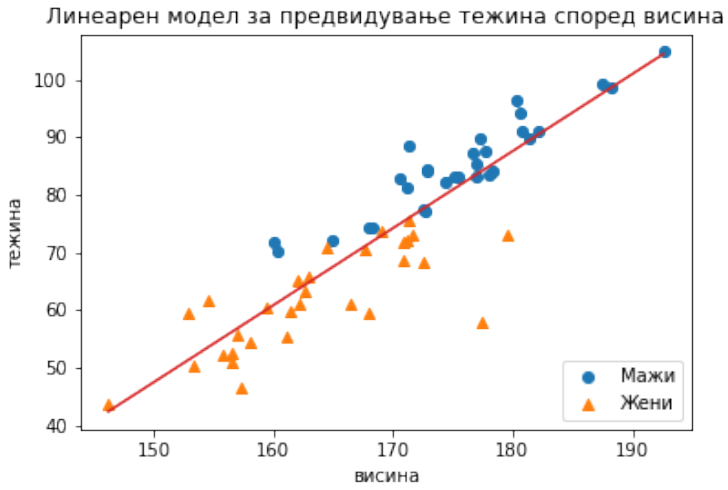
$$w = \left(X^T X \right)^{-1} X^T y,$$

што значи дека пресметката на коефициентите на моделот се сведува на едноставни операции со матрици. Овој факт, заедно со простата структура, го прават линеарниот регресионен модел доста примамлив за користење, па во пракса, наместо конструкција на посложени модели, често се прибегнува кон *линеаризација* на податоците за да се доведат до облик соодветен за моделирање со линеарен регресионен модел. Во продолжение ќе разгледаме пример на примена на ваков модел.

Висина (cm)	**Тежина (kg)**	**Пол**
172.7	77.2	M
172.6	77.4	M
162.7	63.3	F
156.5	51.0	F

Табела 1. Дел од податочното множество кое ќе го користиме.

Пример 1. Дадено ни е податочно множество со набљудувања поврзани со висината, тежината и полот на 60 лица. Дел од податоците се прикажани во Табела 1. Целта е да се конструира линеарен модел врз чија основа може да се изврши предвидување на тежината на лице, ако се знае неговата висина. Податоците се прикажани на дијаграм на расејување на Слика 1 на која со соодветен симбол е означен и полот (ова ќе ни биде потребно подоцна).



Слика 1. Дијаграм на расејување за податоците кои ги моделираме во примерот 1. Црвената линија е линеарниот модел за податоците.

Според податоците, тежината y може да се предвиди на основа на висината x со помош на моделот:

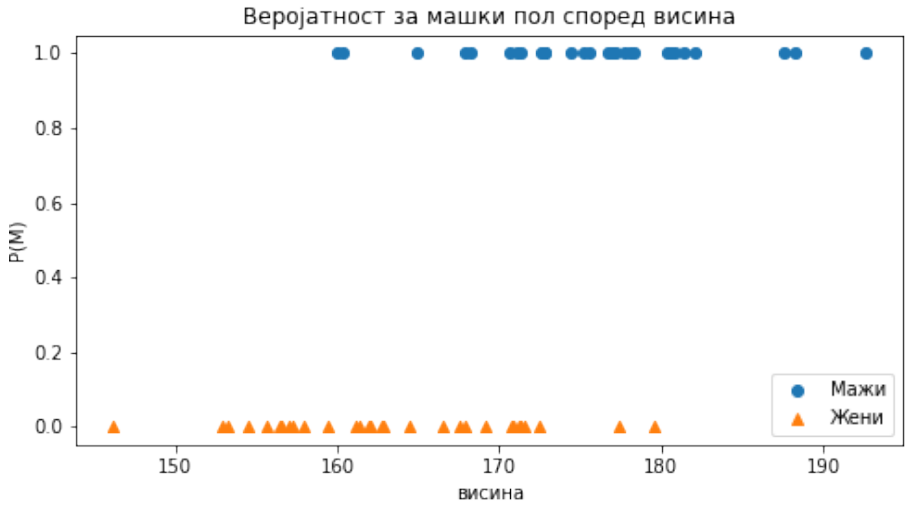
$$\hat{y} = -153.1 + 1.34x$$

Според моделот, предвидувањето е дека лице со висина од 172.7 cm би тежело 78.3 kg, но во првиот ред од Табела 1 гледаме дека лице со оваа висина тежи 77.2 kg, што значи дека моделот прилично добро си ја врши работата. Отстапувањето на предвидената од вистинската вредност се должи на фактот дека моделот кој го користиме е од стохастичка природа, односно вистинската зависност помеѓу влезните и излезната променлива не се чисто линеарни, туку содржат и стохастичка компонента (означена со ε во (1)).

1. МОДЕЛИРАЊЕ ВЕРОЈАТНОСТИ

Видовме дека линеарниот модел се покажа како соодветен избор за опишување на зависноста на висината и тежината на лицата од податочното множество. Но, целта е да конструираме модел кој ќе го предвидува полот на лицето на основа на висината и/или тежината. Прва пречка со која треба да се справиме за во оваа ситуација да може да се примени линеарен модел на оваа ситуација е фактот што полот, завис-

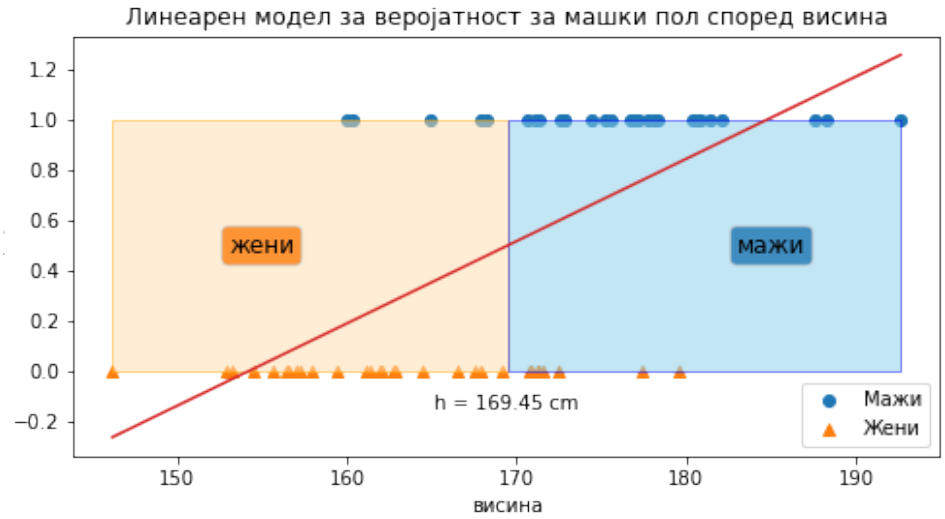
ната променлива, не е од нумеричка природа. Ова може да се надмине лесно ако работиме со *веројатноста* лице да биде од одреден пол, наместо со самата „вредност“ на полот. Така, воведуваме нова променлива-индикатор M за која ќе важи $P(M)=1$, ако соодветното лице е од машки пол, и $P(M)=0$, ако лицето е од женски пол. На овој начин се врши кодирање на категориската променлива *пол* и сега веќе може да конструираме, барем во принцип, линеарен модел кој ќе ја предвидува *веројатноста* лице да биде од машки пол на основа на висината и/или тежината. На Слика 2 е даден дијаграм на расејување на веројатноста $P(M)$ и висината, со и без линеарниот модел.



Слика 2. а) Дијаграм на расејување на $P(M)$ наспрема висината;

Јасно е дека е прилично лесно да се изврши класификација на лицата според нивниот пол со користење на моделот. Има смисла да се користи следново правило: доколку моделот предвидува веројатност од барем 0,5 лицето да биде од машки пол, тогаш го класифицираме како машко; во спротивно го класифицираме како женско. На Слика 2 б) се дадени регионите за класификација: сите лица повисоки од 169,45 cm ќе бидат класифицирани како машки, а сите пониски од таа вредност – како женски.

Очекувано, ќе има и погрешни класификации бидејќи моделот не е совршен (за оценка на моделот ќе дискутираме понатаму), но самиот модел има еден очигледен недостаток.



Слика 2. б) Линеарен модел за $P(M)$ на основа на висината и можна класификација на полот на основа на моделот.

Имено, линеарната функција не е ограничена на нејзината дефиниционa област, па како таква не е соодветна за моделирање веројатности, кои, пак, припаѓаат во интервалот (0, 1). На пример, за многу високи лица моделот ќе даде веројатност поголема од 1, а за многу ниски лица – негативна веројатност. Ова нè наведува да заклучиме дека во принцип е можно веројатности да се моделираат со помош на линеарен регресионен модел, но дека моделот при ваквата примена има недостатоци. Затоа, мора да прибегнеме кон моделирање на друга величина кој е поврзана со веројатноста, но е во склад со неограниченоста на линеарната функција.

2. МОДЕЛИРАЊЕ ШАНСИ

Постои една друга величина која ја содржи истата информација како и веројатноста, но по својата големина е посоодветна за моделирање со помош на линеарен модел. Станува збор за *шансите* (анг. *odds*) да настапи еден настан.

Нека веројатноста да настапи настанот A е еднаква на $p \in (0,1)$. Под „шанса да настапи настанот A“, со ознака o, ќе го подразбираме количникот:

$$o = \frac{p}{1-p}.$$

На пример, ако за веројатноста на некој настан имаме $p = \frac{1}{3}$, тогаш шансата да настапи тој настан е: $o = \frac{\frac{1}{3}}{1-\frac{1}{3}} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2} = 1:2$, односно велиме дека настанот има „еден спрема два“ или „еден во два“ шанси да настапи. Со други зборови, шансите велат дека за секое едно настапување, имаме две ненастапувања на настанот. Од ова се гледа дека шансите се сосема во склад со веројатностите.

Она што може да го заклучиме е следново: од $p \in (0,1)$ следи дека $o \in (0,+\infty)$ и трансформацијата $p \mapsto \frac{p}{1-p}$ е монотоно растечка на (0, 1). Но ова сè уште не е облик кој целосно соодветствува на линеарниот модел. Во последниот чекор, наместо вредноста на шансата, ја земаме вредноста на природниот логаритам од шансата. Така се добива конечната трансформација:

$$p \mapsto \ln \frac{p}{1-p}. \tag{3}$$

Бидејќи $\ln : (0,+\infty) \to (-\infty,+\infty)$, заклучуваме дека трансформацијата дадена во (3) го пресликува интервалот (0, 1) во $(-\infty,+\infty)$. Конечно, заклучуваме дека логаритмот од шансата е величина која е соодветно да се моделира со помош на линеарен регресионен модел. Ова значи дека моделот е:

$$\hat{y} = \ln \frac{p}{1-p},$$

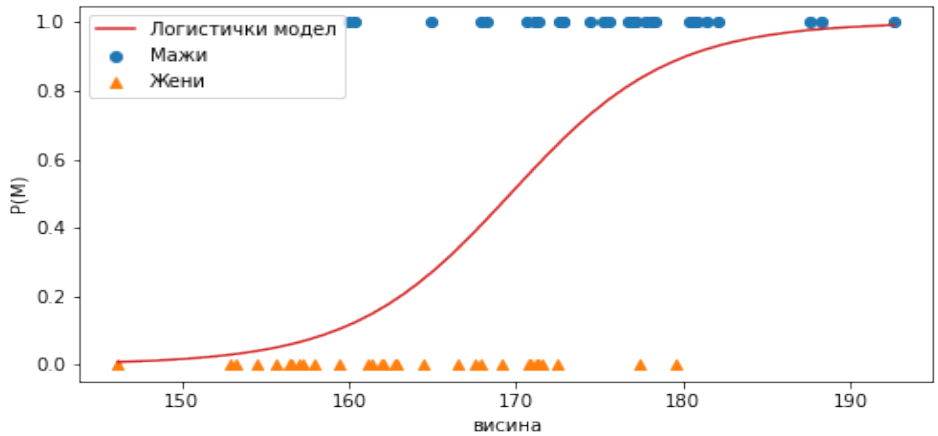
од каде со преуредување на изразот, може да добиеме посреден модел за веројатноста даден со:

$$p = \frac{e^{\hat{y}}}{1+e^{\hat{y}}} \quad \Leftrightarrow \quad p = \frac{1}{1+e^{-\hat{y}}}$$

Овој модел се нарекува *логистички регресионен модел* и, според конструкцијата, секогаш дава вредности во интервалот (0, 1) кои може директно да се толкуваат како веројатности. Да забележиме дека $\hat{y} = Xw$ е линеарниот модел чија конструкција ја дискутиравме во точка 1 и кај

кој коефициентите w се добиени со решавање на задачата на минимизација без ограничувања (2) со помош на методот на најмали квадрати.

Пример 2. Примена на логистичкиот модел на податочното множество со кое работиме резултира со моделот даден на Слика 3. На сликата јасно се гледа дека иако логаритмот од шансата зависи линеарно од висината на лицата, веројатноста не зависи линеарно. Овој модел очигледно е подобар бидејќи (i) ограничен е на (0, 1) што е соодветно за веројатноста и (ii) промената на веројатноста не е константна во однос на промената на висината.



Слика 3. Логистички модел за моделирање на веројатноста $P(M)$ на основа на висината на лицата.

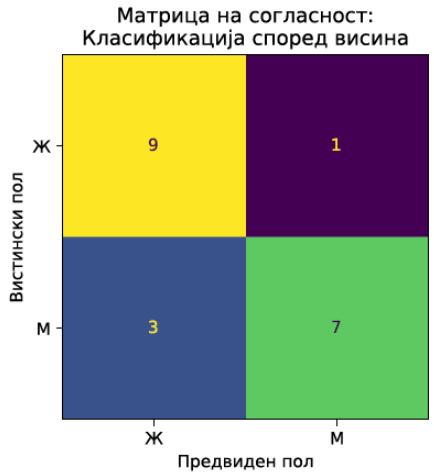
Слично како и претходно, како лице од машки пол го класифицираме секое лице за кое, врз основа на висината, моделот предвидува веројатност поголема од 0,5; во спротивно го класифицираме тоа лице како женско. Да се потсетиме дека ова го правиме поради тоа што избравме моделот да ја моделира веројатноста на настанов „лицето е од машки пол.“

Но, и при вакви услови постои можност моделот да направи погрешна класификација. Ова е факт кој не може да го избегнеме, бидејќи во принцип е невозможно само врз основа на висината на лица да се изврши нивна совршена класификација по пол. Останува уште да извршиме оценка на моделот во контекст на прецизноста на класификацијата.

3. ОЦЕНКА НА МОДЕЛОТ

Постојат повеќе начини на кои може да се оцени моделот. Ние тука ќе изложиме еден мошне едноставен пристап. Треба да го имаме предвид е следново: оценка на моделот не се прави на основа на податоци кои моделот „ги видел“, т.е. податоци на основа врз кои моделот е конструиран; моделот секогаш треба да се оценува на „свежи“ податоци кои се сродни, но не идентични со податоците користени во негова конструкција. Во пракса, се препорачува, околу 25% од податочното множество да се издвои за оваа цел уште пред конструкцијата на моделот. На ваков начин се формира *тестирачко множество*. Ние располагаме со симболично множество од 20 лица, по десет машки и женски.

Оценката на моделот ја вршиме на следниов начин: за секое лице во тестирачкото множество, вршиме класификација со помош на моделот. Потоа ја споредуваме класификацијата на моделот со „вистинската“ состојба и бележиме дали моделот класифицирал точно или не. Резултатите ги прикажуваме во табела-матрица која ја нареквуваме *матрица на согласност* (анг. *confusion matrix*).



Слика 4. Матрица на согласност за моделот на логистичка регресија за класификација на лица според пол на основа на висина.

Она што може да забележиме е дека моделот има глобална точност на класификација од 80% (само четири лица од 20 не се точно класифицирани), но точноста е различна за машки (70%, три од десет неточни) и

женски (90%, само еден од десет неточни). Ова значи дека моделот не е подеднакво добар во препознавање машки и женски и може да заклучиме дека со осетно поголема точност класифицира женски одошто машки. Иако процентот на точно класифицирани лица е висок, сепак ова е прилично скромен резултат. Врз основа на мало податочно и тестирачко множеств, и ваквиот резултат е сосема солиден.

4. ЗАКЛУЧОК

На самиот почеток си поставивме задача да го воведеме моделот на линеарна регресија, да ја опишеме неговата конструкција и неговите особености и да го примениме во два различни контексти. Иако навидум неприродно и непогодно, успеавме да покажеме како може да се надминат недоследностите кои ги носи наивната примена на моделот на линеарна регресија при негова примена на задачата на бинарна класификација. Ова не доведе до соодветен модел за класификација заснован на шанси наместо на веројатност кој се нарекува модел на логистичка регресија. Примената и оценката на овој модел ја илустриравме со негова примена на задачата на класификација на лица според пол на основа на нивната висина.

ЛИТЕРАТУРА

[1] M. P. Deisenroth, A. A. Faisal, C. S. Ong, *Mathematics for Machine Learning*, Cambridge University Press, 2020.

[2] A. B. Downey, *Think Stats (2nd ed)*, O'Reilly Media, 2015.

1 Универзитет „Св. Кирил и Методиј“ во Скопје,
Машински Факултет, Скопје
Руѓер Бошковиќ бр. 18, (1000) Скопје, Р. Северна Македонија
e-mail: filip.nikolovski@mf.ukim.edu.mk

Примен: 7.2.2022
Поправен: 15.2.2022
Одобрен: 16.2.2022
Објавен на интернет: 17.2. 2022