Metod najmanjih kvadrata

Jovana Arsenović, Nikolina Živanović, Anastasija Golić,
Veljko Radojičić, Stefan Topalov
November 20, 2023

Elementi teorije grešaka

Sve veći broj realnih problema, u svim oblastima života, danas se rešava matematičkim modelovanjem i simulacijom istih, zahvaljujući intezivnom razvoju računara. Ukoliko bismo želeli da modelujemo proces, čija su merenja predstavljena na slici 1, najjednostavnije bismo to učinili postupkom **linearne regresije**, i tada bismo vezu između obeležja predstavili u linearnoj formi:

$$y = a_1 x + a_0, \tag{1}$$

gde su a_1 i a_0 nepoznati parametri, koje želimo da odredimo u optimizacionoj proceduri, pri čemu parametar a_1 predstavlja nagib krive, a parametar a_0 je mesto preseka sa ordinatom.

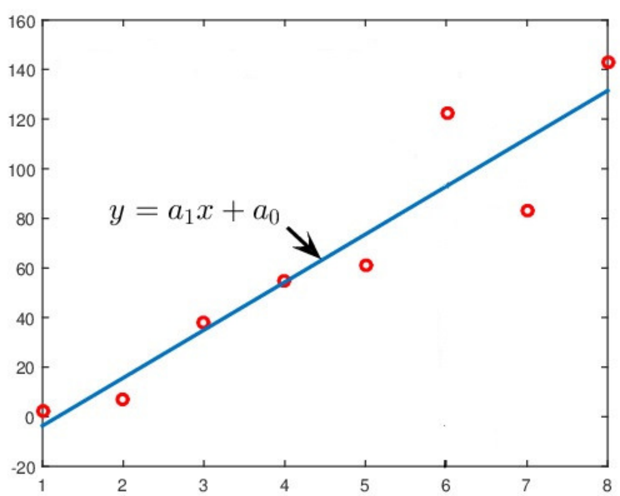

Figure 1: Ilustracija linearne regresije - fitovanje podataka polinomom prvog reda.

U toku optimizacionog postupka potrebno je pronaći one parametre a_1 i a_0 koji će omogućiti da prava najbolje opisuje dati proces, odnosno da razlika između stvarne i aproksimirane vrednosti bude minimalna:

$$E_i = |y - y_i| \,. \tag{2}$$

Ovaj tip greške se naziva **apsolutna greška**[1] i može se predstaviti i na sledeće načine:

1. **Maksimalna apsolutna greška**

$$E_\infty(f) = \max_{1<k<n} |y(x_k) - y_k| .$$

2. **Srednja apsolutna greška**

$$E_1(f) = \frac{1}{n} \sum_{k=1}^{n} (y(x_k) - y_k)^2 .$$

3. **Srednja kvadratna apsolutna greška**

$$E_{LSQ}(f) = \sqrt{\frac{1}{n} \sum_{k=1}^{n} (y(x_k) - y_k)^2} .$$

[1] Ukoliko tačna vrednost nije poznata, a što je slučaj sa numeričkim iterativnim postupcima, tada ne možemo odrediti apsoltnu grešku jer nam nije poznata prava vrednost. U tom slučaju definišemo približnu apsolutnu grešku koja predstavlja razliku između trenutne i prethodne iteracije.

Metod najmanjih kvadrata

Krećemo od postupka linearne regresije uz oslonac na metod najmanjih kvadrata (jer ovi postupci predstavljaju osnovu mašinskog učenja) i predstavićemo ih kao optimizacioni problem. Na samom početku potrebno je da definišemo kriterijum optimalnosti i predstavimo ga u kvadratnoj formi, gde je cilj da minimizujemo grešku:

$$F = \sum_{k=1}^{n} (y(x_k) - y_k)^2 . \tag{3}$$

Kako se bavimo problemom linearne regresije, kriterijum optimalnosti sada dobija oblik

$$F = \sum_{k=1}^{n} (a_1 x_k + a_0 - y_k)^2 . \tag{4}$$

Kao što smo ranije napomenuli, cilj je da odredimo optimalne vrednosti parametara a_0 i a_1 tako da je kriterijum optimalnosti (12) u minimumu. Na osnovu potrebnih uslova ekstrema, sledi

$$\begin{aligned}
\frac{\partial F}{\partial a_0} &= \sum_{k=1}^{n} 2\left(a_1 x_k + a_0 - y_k\right) = 0 \\
\frac{\partial F}{\partial a_1} &= \sum_{k=1}^{n} 2\left(a_1 x_k + a_0 - y_k\right) x_k = \sum_{k=1}^{n} 2\left(a_1 x_k^2 + a_0 x_k - y_k x_k\right) = 0 .
\end{aligned} \tag{5}$$

Linearni sistem jednačina (5) može se napisti i u matričnoj formi

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} \tag{6}$$

U nastvaku ćemo se baviti rešavanje matrične jednačine (3) i njegovom daljom primenom u studiji problema regresije. Do rešenja sistema (6) dolazimo primenom Kramerovog pravila [2]. Rešavanjem sistema jednačina i dobijanjem optimalne vrednosti parametara a_0 i a_1, formalno je završen postupak linearne regresije.

[2] Kramerovo pravilo jeste teorema u linearnoj algebri koja daje rešenja sistema linearnih jednačina preko determinanti. Ako sistem jednačina predstavimo u formi množenja matrica: $Ax = c$, gde je kvadratna matrica A regularna a x je vektor kolone promenljivih, onda je $x_i = \frac{drt(A_i)}{det(A)}$. A_i je matrica koja se dobija zamenom i-te kolone iz A vektorom kolone c.

$$\begin{aligned} a_1^* =& \frac{n\sum_{k=1}^n x_k y_k - \sum_{k=1}^n x_k \sum_{k=1}^n y_k}{n\sum_{k=1}^n x_k{}^2 - \left(\sum_{k=1}^n x_k\right)^2} \\ a_0^* =& \frac{\sum_{k=1}^n y_k}{n} - a_1^* \frac{\sum_{k=1}^n x_k}{n} \ , \end{aligned} \tag{7}$$

gde su sa a_0^* i a_1^* obeležene optimlane vrednosti parametara a_0 i a_1 u skladu sa pretpostavljenim kriterijumom optimalnosti (12). Optimalnu vrednost kriterijima, obeležićemo sa F_{LSE} i izračunavamo ga na sledeći način

$$F_{LSE} = \sum_{i=1}^{n} \left(a_1^* x_i + a_0^* - y_i\right)^2 . \tag{8}$$

Linearizacija nelinearnih modela

Logično je pretpostavit da veze između ulaza x_i i izlaza y_i nije uvek linearna, ali postoji široka klasa nelinearnih problema, gde se uz odgovarajuće smene početni problem nelinearne regresije može svesti na linearan.
Ukoliko bismo početni problem želeli da aproksimiramo polinomom drugog reda, odnosno:

$$y = a_2 x^2 + a_1 x + a_0, \tag{9}$$

tada bi kriterijum optimalnosti bio u sledećoj formi:

$$F = \sum_{k=1}^{n} \left(a_2 x_k^2 + a_1 x_k + a_0 - y_k\right)^2 . \tag{10}$$

Lakom matematikom dobijamo uslove ekstrema, izjednačavanjem parcijalnih izvoda sa nulom. Do optimalnih vrednosti parametara a_2, a_1 i a_0 se dolazi takođe Kramerovim pravilom, a to ostavljamo čitatelju za vežbu:

$$\begin{aligned} \frac{\partial F}{\partial a_0} &= \sum_{k=1}^{n} 2\left(a_2 x_k^2 + a_1 x_k + a_0 - y_k\right) = 0 \\ \frac{\partial F}{\partial a_1} &= \sum_{k=1}^{n} 2\left(a_2 x_k^2 + a_1 x_k + a_0 - y_k\right) x_k = 0 \\ \frac{\partial F}{\partial a_2} &= \sum_{k=1}^{n} 2\left(a_1 x_k + a_0 - y_k\right) x_k^2 = 0. \end{aligned} \tag{11}$$

Eksponencijalni model bismo predstavili na sledeći način:

$$y(x) = Ce^{Ax} \ .$$

Ukoliko bismo ovaj model uvrstili u kriterijum optimalnosti (3)

$$F = \sum_{k=1}^{n} (Ce^{Ax_k} - y_k)^2 \ , \tag{12}$$

potrebne uslove ekstrema dobili bismo određivanjem parcijalnih izvoda kriterijuma po svim promenljivim:

$$\begin{aligned} \frac{\partial F}{\partial A} &= \sum_{k=1}^{n} 2\left(Ce^{Ax_k} - y_k\right) Cx_k e^{Ax_k} = 0 \\ \frac{\partial F}{\partial C} &= \sum_{k=1}^{n} 2\left(Ce^{Ax_k} - y_k\right) e^{Ax_k} = 0 \ . \end{aligned} \tag{13}$$

Jasno je da je ovo nemoguće napisati u matričnoj formi, pa zbog toga pribegavamo procesu aproksimacije - linearizaciji.
Linearizacija ovog modela je prirodna i dosta jednostavna. Ako nađemo prirodni logaritam prethodnog izraza dobijamo

$$\ln y = \ln C + Ax \ .$$

sada je očigledno da je novo $Y = \ln y$, a koeficijenti iz jednačina (7) su $a_0 = \ln \alpha$ i $a_1 = \beta$, odnosno

$$Y = a_0 + a_1 x \ .$$

Primenom jednačina (7) lako određujemo optimalne vrednosti parametara linearizovanog modela, odnosno izračunavamo a_0^* i a_1^*, što konačno daje eksponencijalni regresioni model kao

$$y(x) = Ce^{Ax} = \left(e^{a_0^*}\right) e^{a_1^* x} \ .$$

1. Tokom ispita prikupljeni su podaci o studentim, koliko vremena su proveli spremajući ispit i koliko su bodova osvojili, i ti podaci su predstavljeni u tabeli.

SATI	BODOVI
6	82
10	88
2	56
4	64
0	23

Podatke predstaviti:

(a) linearno

(b) eksponencijalno

i proceniti koliko je potrebno vremena da bi se ispit položio.

Rešenje.

(a) Kako je u tekstu zadatka traženo da se podaci predstave linearno, potrebno je da odredimo parametre prave, kako bismo aproksimirali podatke linearno, odnosno potrebno je nađemo optimalno a_0 i a_1, koristeći izraze (7):

$$a_1^* = \frac{5(6*82+10*88+2*56+4*64)-(6+10+2+4)(82+88+56+64+23)}{5(36+100+4+16)-(6+10+2+4)^2} = 6.1284$$
$$a_0^* = \frac{82+88+56+64+23}{5} - 6.1284\frac{6+10+4+2}{5} = 35.6351.$$

Pa bi prava bila oblika $y = 6.1284 + 35.6351$. Na osnovu toga možemo da zaključimo da bi se ispit položio potrebno je 2.5h učenja.

(b) Na samom početku potrebno je da odredimo Y kao $ln(y)$:

$$Y = ln(y) = [4.4067, 4.4773, 4.0254, 4.1589, 3.1355],$$

koje se uvrsti u izraz (7) i dobiju se optimalne vrednosti parametara $a_1* = 0.1183$ i $a_0* = 3.5202$, pa bi na osnovu toga izračunali $C = e^{a_0*} = 33.7927$, i dobili eksponencijalnu krivu $y = 33.7927e^{0.1183x}$