Matematične osnove za umetno inteligenco

Programski odbor izbirnega tekmovanja iz umetne inteligence

1 Uvod

Pri reševanju nalog iz umetne inteligence pogosto uporabljamo matematične pojme iz algebre, verjetnosti, statistike in analize. Namen tega gradiva ni podati popolnega pregleda teh področij, temveč predstaviti matematična orodja, ki so potrebna za razumevanje metod in formul pri umetni inteligenci.

Gradivo je razdeljeno na tri dele: najprej je podan kratek pregled ključnih formul, nato sledijo podrobnejše razlage posameznih matematičnih pojmov s primeri, na koncu pa so zbrane naloge za utrjevanje skupaj z rešitvami.

Poudarek je na razumevanju pomena formul in njihovi uporabi. Pri večini nalog ni pomembno, da formule izpeljemo, temveč da razumemo, kaj posamezne količine pomenijo in kako jih izračunamo.

2 Kratek pregled ključnih formul

Pojem	Formula
Aritmetična sredina	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Varianca	$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$
Standardizacija	$z = \frac{x - \mu}{\sigma}$
Evklidska razdalja	$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$
Skalarni produkt	$\mathbf{x} \cdot \mathbf{y} = \sum_i x_i y_i$
Pogojna verjetnost	$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$
Bayesov izrek	$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$
Pričakovana vrednost	$E[X] = \sum_x xP(X = x)$
MSE	$\frac{1}{n}\sum_i (y_i - \hat{y}_i)^2$
MAE	$\frac{1}{n}\sum_i \lvert y_i - \hat{y}_i \rvert$
Sigmoidna funkcija	$\sigma(z) = \frac{1}{1 + e^{-z}}$
Entropija	$H = -\sum_i p_i \log_2 p_i$
Gradientni spust	$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla L$
Precision	$\frac{TP}{TP + FP}$
Recall	$\frac{TP}{TP + FN}$
F_1	$2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$
Diskontirani povratek	$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

3 Osnovni matematični zapis

3.1 Spremenljivke, indeksi in vektorji

Posamezen podatek pogosto označimo z x. Če imamo več podatkov, jih lahko označimo

$$x_1, x_2, \ldots, x_n. \tag{1}$$

Indeks i pove, kateri element opazujemo. Na primer, če imamo

$$x_1 = 3, \qquad x_2 = 7, \qquad x_3 = 2, \tag{2}$$

potem je $x_2 = 7$.

Pri strojnem učenju posamezen primer pogosto vsebuje več značilk. Zapišemo ga kot vektor

$$\mathbf{x} = (x_1, x_2, \ldots, x_d), \tag{3}$$

kjer je d število značilk.

Če hišo opisujemo s površino, številom sob in starostjo, lahko na primer zapišemo

$$\mathbf{x} = (120, 4, 15). \tag{4}$$

To pomeni: površina je 120, hiša ima 4 sobe in je stara 15 let.

3.2 Vsota

Znak $\sum$ uporabljamo za krajši zapis vsote. Na primer

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4. \tag{5}$$

Če so

$$x_1 = 2, \quad x_2 = 5, \quad x_3 = 1, \quad x_4 = 4, \tag{6}$$

potem je

$$\sum_{i=1}^{4} x_i = 2 + 5 + 1 + 4 = 12. \tag{7}$$

Tak zapis se pogosto pojavlja v funkcijah izgube. Povprečna kvadratna napaka je na primer

$$\mathrm{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{8}$$

Pri tem je y_i dejanska vrednost, $\hat{y}_i$ napoved modela, n pa število primerov.

3.3 Absolutna vrednost

Absolutna vrednost predstavlja oddaljenost števila od ničle:

$$|5| = 5, \qquad |-5| = 5. \tag{9}$$

Zato je absolutna napaka napovedi

$$|y - \hat{y}|. \tag{10}$$

Na primer, če je prava vrednost $y = 10$, model pa napove $\hat{y} = 7$, je absolutna napaka

$$|10 - 7| = 3. \tag{11}$$

4 Funkcije

Funkcija je predpis, ki vhodni vrednosti priredi izhodno vrednost. Zapis

$$y = f(x) \tag{12}$$

pomeni, da vrednost y dobimo tako, da funkcijo f uporabimo na vhodu x.

Na primer

$$f(x) = 2x + 1. \tag{13}$$

Za $x = 3$ dobimo

$$f(3) = 2 \cdot 3 + 1 = 7. \tag{14}$$

V strojnem učenju lahko model razumemo kot funkcijo, ki vhodnim značilkam priredi napoved:

$$\hat{y} = f(\mathbf{x}). \tag{15}$$

4.1 Linearna funkcija

Preprost linearni model ene značilke ima obliko

$$\hat{y} = wx + b, \tag{16}$$

kjer je w utež, b pa prosti člen.

Če je

$$\hat{y} = 2x + 3, \tag{17}$$

potem pri $x = 4$ dobimo

$$\hat{y} = 2 \cdot 4 + 3 = 11. \tag{18}$$

Pri več značilkah lahko linearni model zapišemo

$$\hat{y} = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b. \tag{19}$$

Vsaka utež w_j določa, kako močno posamezna značilka vpliva na napoved.

4.2 Potence in koreni

Pri strojnem učenju pogosto srečamo kvadrate:

$$x^2 = x \cdot x. \tag{20}$$

Na primer

$$(-3)^2 = 9. \tag{21}$$

Kvadratni koren je obratna operacija kvadriranja:

$$\sqrt{25} = 5. \tag{22}$$

Kvadrati in koreni se pojavijo na primer pri evklidski razdalji in kvadratni napaki.

4.3 Eksponentna funkcija

Eksponentna funkcija ima obliko

$$f(x) = a^x, \tag{23}$$

kjer je $a > 0$.

V strojnem učenju se pogosto pojavlja število

$$e \approx 2.718. \tag{24}$$

Pomembna je zato funkcija

$$e^x. \tag{25}$$

Na primer logistična regresija uporablja sigmoidno funkcijo

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{26}$$

Njena vrednost je vedno med 0 in 1, zato jo lahko uporabimo za predstavitev verjetnosti pozitivnega razreda.

Velja na primer

$$\sigma(0) = \frac{1}{1 + e^0} = \frac{1}{2}. \tag{27}$$

4.4 Logaritem

Logaritem je obratna operacija potenciranja. Zapis

$$\log_a x = y \tag{28}$$

pomeni

$$a^y = x. \tag{29}$$

Na primer

$$\log_2 8 = 3, \tag{30}$$

ker je

$$2^3 = 8. \tag{31}$$

Koristna pravila so

$$\log(ab) = \log a + \log b, \tag{32}$$

$$\log\left(\frac{a}{b}\right) = \log a - \log b, \tag{33}$$

in

$$\log(a^k) = k \log a. \tag{34}$$

V strojnem učenju se logaritem pojavlja na primer pri logaritemski transformaciji, entropiji in logaritemski funkciji izgube.

Primer: logaritemska transformacija. Če primerjamo vrednosti

$$10, \qquad 100, \qquad 1000, \tag{35}$$

dobimo pri desetiškem logaritmu

$$1, \qquad 2, \qquad 3. \tag{36}$$

Logaritem torej močno stisne velik razpon vrednosti. Zato je lahko koristen pri značilkah z nekaj zelo velikimi vrednostmi.

5 Vektorji

Vektor je urejen seznam števil. Na primer

$$\mathbf{x} = (2, 3, 5). \tag{37}$$

V strojnem učenju vektor pogosto predstavlja en primer, njegove komponente pa posamezne značilke.

5.1 Seštevanje vektorjev

Vektorja seštejemo po komponentah:

$$(1, 2) + (3, 4) = (4, 6). \tag{38}$$

Podobno velja

$$(5, 7) - (2, 3) = (3, 4). \tag{39}$$

5.2 Množenje vektorja s skalarjem

Skalar je običajno število. Vektor s skalarjem pomnožimo tako, da z njim pomnožimo vsako komponento:

$$3(1, 2) = (3, 6). \tag{40}$$

5.3 Skalarni produkt

Za vektorja

$$\mathbf{x} = (x_1, \ldots, x_d), \qquad \mathbf{w} = (w_1, \ldots, w_d) \tag{41}$$

je skalarni produkt

$$\mathbf{w} \cdot \mathbf{x} = \sum_{j=1}^{d} w_j x_j. \tag{42}$$

Za

$$\mathbf{x} = (2, 3), \qquad \mathbf{w} = (4, 1) \tag{43}$$

dobimo

$$\mathbf{w} \cdot \mathbf{x} = 4 \cdot 2 + 1 \cdot 3 = 11. \tag{44}$$

Linearni model lahko zato krajše zapišemo kot

$$\hat{y} = \mathbf{w} \cdot \mathbf{x} + b. \tag{45}$$

5.4 Dolžina oziroma norma vektorja

Evklidska norma vektorja je

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_d^2}. \tag{46}$$

Za

$$\mathbf{x} = (3, 4) \tag{47}$$

je

$$\|\mathbf{x}\|_2 = \sqrt{3^2 + 4^2} = 5. \tag{48}$$

5.5 Evklidska razdalja

Razdalja med vektorjema $\mathbf{x}$ in $\mathbf{y}$ je

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2. \tag{49}$$

V dveh dimenzijah:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}. \tag{50}$$

Primer. Naj bo

$$\mathbf{x} = (1, 2), \qquad \mathbf{y} = (4, 6). \tag{51}$$

Potem je

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(1 - 4)^2 + (2 - 6)^2} = \sqrt{9 + 16} = 5. \tag{52}$$

Razdalje so pomembne pri algoritmih, kot je k-NN. Če imajo značilke zelo različna merila, lahko značilka z velikimi številskimi vrednostmi prevlada pri izračunu razdalje. Zato je pri takih algoritmih pogosto potrebno skaliranje značilk.

5.6 Kosinusna podobnost

Podobnost dveh vektorjev lahko merimo tudi s kosinusno podobnostjo:

$$\text{cos_sim}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}. \tag{53}$$

Vrednost 1 pomeni, da vektorja kažeta v isto smer, vrednost 0 pa, da sta pravokotna.

Ta mera je uporabna predvsem takrat, ko nas bolj kot velikost vektorjev zanima njihova smer oziroma vzorec vrednosti.

6 Matrike

Matrika je pravokotna tabela števil. Na primer

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{pmatrix}. \tag{54}$$

Ta matrika ima 3 vrstice in 2 stolpca, zato pravimo, da ima dimenzijo

$$3 \times 2. \tag{55}$$

Pri strojnem učenju lahko podatkovno množico predstavimo kot matriko

$$X = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1d} \\ x_{21} & x_{22} & \cdots & x_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nd} \end{pmatrix}, \tag{56}$$

kjer vrstice predstavljajo primere, stolpci pa značilke.

6.1 Množenje matrike z vektorjem

Naj bo

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}, \qquad \mathbf{x} = \begin{pmatrix} 5 \\ 6 \end{pmatrix}. \tag{57}$$

Potem je

$$A\mathbf{x} = \begin{pmatrix} 1 \cdot 5 + 2 \cdot 6 \\ 3 \cdot 5 + 4 \cdot 6 \end{pmatrix} = \begin{pmatrix} 17 \\ 39 \end{pmatrix}. \tag{58}$$

Za osnovno razumevanje strojnega učenja je predvsem pomembno razumeti, da lahko z matričnim zapisom hkrati izvedemo iste linearne operacije nad več podatki.

7 Povprečje, varianca in standardni odklon

7.1 Aritmetična sredina

Za vrednosti

$$x_1, x_2, \ldots, x_n \tag{59}$$

je aritmetična sredina

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i. \tag{60}$$

Za podatke

$$2, \quad 4, \quad 6, \quad 8 \tag{61}$$

je

$$\bar{x} = \frac{2+4+6+8}{4} = 5. \tag{62}$$

7.2 Mediana

Mediana je srednja vrednost urejenih podatkov. Pri

$$1, \quad 3, \quad 4, \quad 8, \quad 20 \tag{63}$$

je mediana 4.

Če je število vrednosti sodo, vzamemo povprečje srednjih dveh.

Mediana je manj občutljiva na ekstremne vrednosti kot aritmetična sredina.

Na primer pri podatkih

$$2, \quad 3, \quad 3, \quad 4, \quad 100 \tag{64}$$

je mediana 3, aritmetična sredina pa

$$\frac{112}{5} = 22.4. \tag{65}$$

7.3 Varianca

Varianca meri, kako močno so vrednosti razpršene okoli povprečja.

Za populacijo jo lahko zapišemo kot

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{66}$$

Večja varianca pomeni večjo razpršenost.

Primer. Za podatke

$$2, \quad 4, \quad 6 \tag{67}$$

je povprečje

$$\bar{x} = 4. \tag{68}$$

Odkloni od povprečja so

$$-2, \quad 0, \quad 2. \tag{69}$$

Njihovi kvadrati so

$$4, \quad 0, \quad 4. \tag{70}$$

Varianca je zato

$$\sigma^2 = \frac{4+0+4}{3} = \frac{8}{3}. \tag{71}$$

7.4 Standardni odklon

Standardni odklon je kvadratni koren variance:

$$\sigma = \sqrt{\sigma^2}. \tag{72}$$

Prednost standardnega odklona je, da je izražen v isti merski enoti kot prvotni podatki.

7.5 Standardizacija

Pri standardizaciji vrednost x pretvorimo v

$$z = \frac{x - \mu}{\sigma}, \tag{73}$$

kjer je μ povprečje, σ pa standardni odklon.

Če je

$$\mu = 100, \qquad \sigma = 20, \qquad x = 140, \tag{74}$$

dobimo

$$z = \frac{140 - 100}{20} = 2. \tag{75}$$

Vrednost $z = 2$ pomeni, da je x za dva standardna odklona nad povprečjem.

8 Osnove verjetnosti

Verjetnost dogodka A označimo z

$$P(A). \tag{76}$$

Vedno velja

$$0 \leq P(A) \leq 1. \tag{77}$$

Verjetnost 0 pomeni nemogoč dogodek, verjetnost 1 pa gotov dogodek.

8.1 Komplement dogodka

Če A^c pomeni, da se dogodek A ne zgodi, velja

$$P(A^c) = 1 - P(A). \tag{78}$$

Če je verjetnost dežja

$$P(A) = 0.3, \tag{79}$$

je verjetnost, da ne bo deževalo,

$$P(A^c) = 0.7. \tag{80}$$

8.2 Presek dogodkov

Zapis

$$P(A \cap B) \tag{81}$$

pomeni verjetnost, da se zgodita oba dogodka A in B.

8.3 Pogojna verjetnost

Pogojna verjetnost

$$P(A \mid B) \tag{82}$$

pomeni verjetnost dogodka A, če že vemo, da se je zgodil dogodek B.

Velja

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \qquad P(B) > 0. \tag{83}$$

Primer. V razredu je 20 dijakov. 12 jih programira v Pythonu, med temi pa jih 5 sodeluje na tekmovanju iz UI.

Če naključno izberemo dijaka in vemo, da programira v Pythonu, je verjetnost, da sodeluje na tekmovanju,

$$P(\text{UI} \mid \text{Python}) = \frac{5}{12}. \tag{84}$$

Pomembno je, da

$$P(A \mid B) \tag{85}$$

na splošno ni enako

$$P(B \mid A). \tag{86}$$

8.4 Neodvisnost

Dogodka A in B sta neodvisna, če informacija o enem ne spremeni verjetnosti drugega.

Takrat velja

$$P(A \cap B) = P(A)P(B). \tag{87}$$

Ekvivalentno, če je $P(B) > 0$,

$$P(A \mid B) = P(A). \tag{88}$$

8.5 Bayesov izrek

Bayesov izrek povezuje pogojni verjetnosti v nasprotnih smereh:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}. \tag{89}$$

Pri klasifikaciji lahko A predstavlja razred, B pa opažene značilke.

Primer. Naj bo 1 % elektronskih sporočil določene vrste neželenih:

$$P(S) = 0.01. \tag{90}$$

Neka značilka se pojavi pri 80 % neželenih sporočil:

$$P(F \mid S) = 0.8, \tag{91}$$

v celotni množici sporočil pa pri 5 % sporočil:

$$P(F) = 0.05. \tag{92}$$

Po Bayesovem izreku je

$$P(S \mid F) = \frac{P(F \mid S)P(S)}{P(F)} = \frac{0.8 \cdot 0.01}{0.05} = 0.16. \tag{93}$$

Čeprav je značilka pogosta pri neželenih sporočilih, je verjetnost, da je sporočilo neželeno ob opaženi značilki, v tem primeru le 16 %.

9 Slučajne spremenljivke in pričakovana vrednost

Slučajna spremenljivka X številčno opiše izid slučajnega poskusa.

Pri metu običajne kocke lahko na primer X predstavlja število pik.

Za diskretno slučajno spremenljivko je pričakovana vrednost

$$E[X] = \sum_x xP(X = x). \tag{94}$$

Pri poštenem kovancu lahko definiramo

$$X = \begin{cases} 1, & \text{če pade glava,} \\ 0, & \text{če pade cifra.} \end{cases} \tag{95}$$

Potem je

$$E[X] = 1 \cdot 0.5 + 0 \cdot 0.5 = 0.5. \tag{96}$$

Pričakovana vrednost ne pomeni nujno vrednosti, ki jo bomo dejansko opazili v posameznem poskusu. Predstavlja dolgoročno povprečje ob velikem številu ponovitev.

10 Porazdelitve

Porazdelitev opisuje, katere vrednosti lahko slučajna spremenljivka zavzame in kako verjetne so.

Pri strojnem učenju pogosto govorimo o porazdelitvi vhodov

$$P(X), \tag{97}$$

porazdelitvi ciljnih vrednosti

$$P(Y) \tag{98}$$

in pogojni porazdelitvi

$$P(Y \mid X). \tag{99}$$

Zapis

$$P(Y \mid X) \tag{100}$$

opisuje, kako je ciljna spremenljivka Y povezana z vhodnimi značilkami X.

Če se pri uporabi modela spremeni

$$P_{\text{train}}(X) \neq P_{\text{deploy}}(X), \tag{101}$$

se je spremenila porazdelitev vhodnih podatkov.

Če pa se spremeni

$$P_{\text{train}}(Y \mid X) \neq P_{\text{deploy}}(Y \mid X), \tag{102}$$

se je spremenila povezava med vhodi in ciljem.

11 Odvod

Odvod opisuje, kako hitro se funkcija spreminja, ko nekoliko spremenimo njen vhod.

Odvod funkcije f označimo

$$f'(x) \tag{103}$$

ali

$$\frac{df}{dx}. \tag{104}$$

Če je odvod v neki točki pozitiven, funkcija tam lokalno narašča. Če je negativen, funkcija lokalno pada.

Za osnovne funkcije je koristno poznati

$$(c)' = 0, \tag{105}$$

$$(x)' = 1, \tag{106}$$

$$(x^n)' = nx^{n-1}, \tag{107}$$

$$(e^x)' = e^x, \tag{108}$$

$$(\ln x)' = \frac{1}{x}. \tag{109}$$

Veljata tudi pravili

$$(f+g)' = f' + g' \tag{110}$$

in

$$(cf)' = cf'. \tag{111}$$

11.1 Primer

Naj bo

$$f(x) = x^2. \tag{112}$$

Potem je

$$f'(x) = 2x. \tag{113}$$

Pri $x = 3$ je

$$f'(3) = 6. \tag{114}$$

Pozitivni odvod pomeni, da funkcija v okolici te točke narašča, če povečujemo x.

12 Funkcije več spremenljivk in parcialni odvodi

V strojnem učenju je funkcija izgube pogosto odvisna od več parametrov:

$$L = L(w_1, w_2, \ldots, w_d). \tag{115}$$

Takšno funkcijo imenujemo funkcija več spremenljivk.

Na primer

$$f(x, y) = x^2 + 2y^2. \tag{116}$$

12.1 Parcialni odvod

Pri parcialnem odvajanju spreminjamo eno spremenljivko, druge pa obravnavamo kot konstante.

Za

$$f(x, y) = x^2 + 2y^2 \tag{117}$$

je parcialni odvod po x

$$\frac{\partial f}{\partial x} = 2x, \tag{118}$$

po y pa

$$\frac{\partial f}{\partial y} = 4y. \tag{119}$$

Parcialni odvod nam torej pove, kako občutljiva je funkcijska vrednost na majhno spremembo ene izmed spremenljivk, če ostale ohranimo nespremenjene.

13 Gradient

Gradient združi vse parcialne odvode funkcije v en vektor.

Za funkcijo

$$f(x_1, \ldots, x_d) \tag{120}$$

je gradient

$$\nabla f = \left(\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_d} \right). \tag{121}$$

Za

$$f(x, y) = x^2 + 2y^2 \tag{122}$$

je

$$\nabla f(x, y) = (2x, 4y). \tag{123}$$

V točki $(1, 2)$ dobimo

$$\nabla f(1, 2) = (2, 8). \tag{124}$$

Gradient kaže v smeri najhitrejšega naraščanja funkcije. Če želimo funkcijo zmanjševati, se zato premikamo v nasprotni smeri gradienta.

14 Gradientni spust

Pri učenju modela pogosto želimo poiskati parametre, pri katerih je funkcija izgube čim manjša.

Gradientni spust (angl. *gradient descent*) uporablja posodobitev

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \nabla L(\boldsymbol{\theta}), \tag{125}$$

kjer je

- $\boldsymbol{\theta}$ vektor parametrov modela,
- L funkcija izgube,
- ∇L gradient funkcije izgube,
- $\alpha > 0$ učna stopnja (angl. *learning rate*).

Primer z enim parametrom. Naj bo

$$L(w) = (w - 3)^2. \tag{126}$$

Minimum je pri $w = 3$, vendar predpostavimo, da tega ne poznamo.

Odvod je

$$L'(w) = 2(w - 3). \tag{127}$$

Začnimo pri

$$w = 5. \tag{128}$$

Takrat je

$$L'(5) = 4. \tag{129}$$

Če je učna stopnja

$$\alpha = 0.1, \tag{130}$$

dobimo po enem koraku

$$w_{\text{novo}} = 5 - 0.1 \cdot 4 = 4.6. \tag{131}$$

Premaknili smo se iz 5 proti minimumu pri 3.

Če je učna stopnja prevelika, lahko koraki minimum preskakujejo. Če je premajhna, je učenje lahko zelo počasno.

15 Funkcije izgube

Funkcija izgube meri, kako slaba je napoved modela. Manjša vrednost praviloma pomeni boljšo napoved.

15.1 Povprečna absolutna napaka

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|. \tag{132}$$

15.2 Povprečna kvadratna napaka

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2. \tag{133}$$

Ker napake kvadriramo, velike napake pri MSE vplivajo močneje kot pri MAE.

Primer. Naj bodo dejanske vrednosti

$$(2, 4, 8) \tag{134}$$

in napovedi

$$(3, 4, 5). \tag{135}$$

Napake so

$$-1, \quad 0, \quad 3. \tag{136}$$

Zato je

$$\text{MAE} = \frac{1 + 0 + 3}{3} = \frac{4}{3}, \tag{137}$$

medtem ko je

$$\text{MSE} = \frac{1^2 + 0^2 + 3^2}{3} = \frac{10}{3}. \tag{138}$$

16 Logistična regresija in sigmoidna funkcija

Pri binarni klasifikaciji želimo napovedati verjetnost

$$P(Y = 1 \mid \mathbf{x}). \tag{139}$$

Logistična regresija najprej izračuna linearno kombinacijo

$$z = \mathbf{w} \cdot \mathbf{x} + b \tag{140}$$

in jo pretvori s sigmoidno funkcijo

$$\sigma(z) = \frac{1}{1 + e^{-z}}. \tag{141}$$

Ker velja

$$0 < \sigma(z) < 1, \tag{142}$$

lahko rezultat interpretiramo kot verjetnost.

Če je $z = 0$, je

$$\sigma(0) = 0.5. \tag{143}$$

Če je z zelo velik in pozitiven, je rezultat blizu 1. Če je z zelo negativen, je rezultat blizu 0.

17 Logaritemska izguba

Pri binarni klasifikaciji se pogosto uporablja izguba

$$L = -\left[y \ln(p) + (1 - y) \ln(1 - p)\right], \tag{144}$$

kjer je $y \in \{0, 1\}$ pravi razred, p pa napovedana verjetnost razreda 1.

Če je $y = 1$, se formula poenostavi v

$$L = -\ln(p). \tag{145}$$

Če model pravilnemu razredu pripiše veliko verjetnost, je izguba majhna.

Na primer

$$-\ln(0.9) \approx 0.105, \tag{146}$$

medtem ko je

$$-\ln(0.1) \approx 2.303. \tag{147}$$

Samozavestna napačna napoved je zato močno kaznovana.

18 Entropija in informacijski dobiček

Odločitvena drevesa potrebujejo mero, ki pove, kako mešana je množica primerov.

Za razrede z verjetnostmi $p_1, \ldots, p_k$ je entropija

$$H = -\sum_{i=1}^{k} p_i \log_2 p_i. \tag{148}$$

Pri dveh razredih:

$$H = -p \log_2 p - (1-p) \log_2 (1-p). \tag{149}$$

Če vsi primeri pripadajo istemu razredu, je entropija

$$H = 0. \tag{150}$$

Če sta pri dveh razredih oba enako zastopana,

$$p = 0.5, \tag{151}$$

je

$$H = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) = 1. \tag{152}$$

Takšna množica je zelo mešana.

Informacijski dobiček primerja entropijo pred razdelitvijo in po njej:

$$IG = H(\text{pred}) - H(\text{po}). \tag{153}$$

Dobra razdelitev močno zmanjša entropijo in ima zato velik informacijski dobiček.

19 Metrike pri klasifikaciji

Pri binarni klasifikaciji razlikujemo:

- TP – pravilno napovedani pozitivni primeri,
- TN – pravilno napovedani negativni primeri,
- FP – napačno napovedani pozitivni primeri,
- FN – napačno napovedani negativni primeri.

Točnost je

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{154}$$

Natančnost pozitivnih napovedi (angl. *precision*) je

$$\text{precision} = \frac{TP}{TP + FP}. \tag{155}$$

Priklic (angl. *recall*) je

$$\text{recall} = \frac{TP}{TP + FN}. \tag{156}$$

Mera F_1 je harmonično povprečje precision in recall:

$$F_1 = 2\frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \tag{157}$$

Primer. Naj bo

$$TP = 40, \qquad FP = 10, \qquad FN = 20, \qquad TN = 130. \tag{158}$$

Potem je

$$\text{precision} = \frac{40}{50} = 0.8, \tag{159}$$

$$\text{recall} = \frac{40}{60} \approx 0.667. \tag{160}$$

Zato je

$$F_1 \approx 2\frac{0.8 \cdot 0.667}{0.8 + 0.667} \approx 0.727. \tag{161}$$

20 Matematika spodbujevanega učenja

Pri spodbujevanem učenju agent v posameznem stanju izbere dejanje in prejme nagrado. Prihodnje nagrade pogosto vrednotimo manj kot takojšnje.

20.1 Diskontirana vsota nagrad

Povratek lahko zapišemo

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots, \tag{162}$$

kjer je

$$0 \leq \gamma \leq 1 \tag{163}$$

diskontni faktor.

Če je γ majhen, agent daje večji poudarek takojšnjim nagradam. Če je blizu 1, so pomembne tudi oddaljene prihodnje nagrade.

Primer. Naj bodo naslednje tri nagrade

$$10, \qquad 5, \qquad 2 \tag{164}$$

in

$$\gamma = 0.9. \tag{165}$$

Potem je

$$G = 10 + 0.9 \cdot 5 + 0.9^2 \cdot 2 = 10 + 4.5 + 1.62 = 16.12. \tag{166}$$

20.2 Vrednost stanja

Vrednost stanja lahko intuitivno razumemo kot pričakovano prihodnjo diskontirano nagrado, če začnemo v tem stanju in sledimo določeni politiki:

$$V^{\pi}(s) = E_{\pi}[G_t \mid S_t = s]. \tag{167}$$

20.3 Q-vrednost

Q-vrednost opisuje kakovost dejanja a v stanju s:

$$Q(s, a). \tag{168}$$

Večja vrednost pomeni, da glede na pričakovane prihodnje nagrade ocenjujemo dejanje kot boljše.

20.4 Posodobitev pri Q-learningu

Ena izmed osrednjih formul Q-learninga je

$$Q(s, a) \leftarrow Q(s, a) + \alpha \left[r + \gamma \max_{a'} Q(s', a') - Q(s, a) \right], \tag{169}$$

kjer je

- s trenutno stanje,
- a izbrano dejanje,
- r prejeta nagrada,
- s' naslednje stanje,
- α učna stopnja,
- γ diskontni faktor.

Izraz

$$r + \gamma \max_{a'} Q(s', a') \tag{170}$$

je nova ciljna ocena.

Razlika

$$r + \gamma \max_{a'} Q(s', a') - Q(s, a) \tag{171}$$

pove, za koliko se trenutna ocena razlikuje od nove informacije.

Primer. Naj bo

$$Q(s, a) = 4, \qquad r = 2, \qquad \gamma = 0.9, \tag{172}$$

najboljša Q-vrednost v naslednjem stanju pa

$$\max_{a'} Q(s', a') = 6. \tag{173}$$

Če je

$$\alpha = 0.5, \tag{174}$$

je cilj

$$2 + 0.9 \cdot 6 = 7.4. \tag{175}$$

Nova vrednost je

$$Q(s, a) \leftarrow 4 + 0.5(7.4 - 4) \tag{176}$$

$$= 5.7. \tag{177}$$

21 Naloge za utrjevanje

1. **Vsote**

   Naj bodo

   $$x_1 = 2, \qquad x_2 = 4, \qquad x_3 = 6, \qquad x_4 = 8. \tag{178}$$

   Izračunajte

   $$\frac{1}{4}\sum_{i=1}^{4} x_i. \tag{179}$$

   Rešitev.

   $$\frac{1}{4}(2+4+6+8) = 5. \tag{180}$$

2. **Linearni model**

   Model ima obliko

   $$\hat{y} = 2x_1 - 3x_2 + 5. \tag{181}$$

   Izračunajte napoved za

   $$x_1 = 4, \qquad x_2 = 2. \tag{182}$$

   Rešitev.

   $$\hat{y} = 2 \cdot 4 - 3 \cdot 2 + 5 = 7. \tag{183}$$

3. **Evklidska razdalja**

   Dana sta primera

   $$\mathbf{x} = (2, 3), \qquad \mathbf{y} = (5, 7). \tag{184}$$

   Izračunajte evklidsko razdaljo.

   Rešitev.

   $$d(\mathbf{x}, \mathbf{y}) = \sqrt{(2-5)^2 + (3-7)^2} = \sqrt{9+16} = 5. \tag{185}$$

4. **Skaliranje in razdalje**

   Osebo opisujemo z značilkama starost in letni dohodek. Dana sta primera

   $$A = (20, 30000), \qquad B = (30, 50000). \tag{186}$$

   Katera značilka bo brez skaliranja močneje vplivala na evklidsko razdaljo? Pojasnite.

   Rešitev.

   Razlika v starosti je

   $$30 - 20 = 10, \tag{187}$$

   razlika v dohodku pa

   $$50000 - 30000 = 20000. \tag{188}$$

   Ker je razlika pri dohodku numerično mnogo večja, bo dohodek skoraj v celoti določal razdaljo. To je razlog, da pri metodah, ki temeljijo na razdaljah, pogosto skaliramo značilke.

5. **Standardizacija**

   Značilka ima povprečje

   $$\mu = 50 \tag{189}$$

   in standardni odklon

   $$\sigma = 10. \tag{190}$$

   Standardizirajte vrednost $x = 70$.

   Rešitev.

   $$z = \frac{70 - 50}{10} = 2. \tag{191}$$

   Vrednost je dva standardna odklona nad povprečjem.

6. **Pogojna verjetnost**

   Od 100 sporočil jih je 20 neželenih. Med neželenimi sporočili jih 15 vsebuje besedo “nagrada”.

   Izračunajte

   $$P(\text{nagrada} \mid \text{neželeno}). \tag{192}$$

   Rešitev.

   Med 20 neželenimi sporočili jih 15 vsebuje besedo “nagrada”, zato je

   $$P(\text{nagrada} \mid \text{neželeno}) = \frac{15}{20} = 0.75. \tag{193}$$

7. **Bayesov izrek**

   Naj bo

$$P(A) = 0.2, \qquad P(B \mid A) = 0.7, \qquad P(B) = 0.4. \tag{194}$$

   Izračunajte $P(A \mid B)$.

   Rešitev.

   Po Bayesovem izreku

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} = \frac{0.7 \cdot 0.2}{0.4} = 0.35. \tag{195}$$

8. **Pričakovana vrednost**

   Agent dobi nagrado 10 z verjetnostjo 0.2, nagrado 2 z verjetnostjo 0.5 in nagrado 0 z verjetnostjo 0.3. Izračunajte pričakovano nagrado.

   Rešitev.

$$E[R] = 10 \cdot 0.2 + 2 \cdot 0.5 + 0 \cdot 0.3 = 3. \tag{196}$$

9. **Funkcija izgube**

   Dejanske vrednosti so

$$\mathbf{y} = (2, 4, 6) \tag{197}$$

   in napovedi

$$\hat{\mathbf{y}} = (3, 4, 8). \tag{198}$$

   Izračunajte MSE.

   Rešitev.

$$\text{MSE} = \frac{(2-3)^2 + (4-4)^2 + (6-8)^2}{3} = \frac{1+0+4}{3} = \frac{5}{3}. \tag{199}$$

10. **Odvod**

    Naj bo

$$f(x) = 3x^2 + 2x + 1. \tag{200}$$

    Izračunajte $f'(x)$ in $f'(2)$.

    Rešitev.

$$f'(x) = 6x + 2. \tag{201}$$

Zato je

$$f'(2) = 14. \tag{202}$$

11. **Gradient**

Naj bo

$$L(w_1, w_2) = w_1^2 + 3w_2^2. \tag{203}$$

Izračunajte gradient v točki

$$(w_1, w_2) = (2, 1). \tag{204}$$

Rešitev.

Parcialna odvoda sta

$$\frac{\partial L}{\partial w_1} = 2w_1, \qquad \frac{\partial L}{\partial w_2} = 6w_2. \tag{205}$$

Zato je

$$\nabla L(2, 1) = (4, 6). \tag{206}$$

12. **Gradientni spust**

Naj bo

$$L(w) = (w - 5)^2. \tag{207}$$

Začnemo pri $w = 7$ in uporabimo učno stopnjo $\alpha = 0.1$.

Izračunajte en korak gradientnega spusta.

Rešitev.

Odvod je

$$L'(w) = 2(w - 5). \tag{208}$$

Pri $w = 7$ je

$$L'(7) = 4. \tag{209}$$

Nova vrednost je

$$w \leftarrow 7 - 0.1 \cdot 4 = 6.6. \tag{210}$$

Parameter se je premaknil proti minimumu pri $w = 5$.

13. **Entropija**

Izračunajte entropijo množice, v kateri je polovica primerov pozitivnih in polovica negativnih.

Rešitev.

Ker je

$$p = 0.5, \tag{211}$$

dobimo

$$H = -0.5 \log_2(0.5) - 0.5 \log_2(0.5) \tag{212}$$
$$= 1. \tag{213}$$

Takšna binarna množica ima največjo entropijo.

14. **Precision in recall**

Model ima rezultate

$$TP = 30, \qquad FP = 10, \qquad FN = 20, \qquad TN = 40. \tag{214}$$

Izračunajte precision in recall.

Rešitev.

$$\text{precision} = \frac{30}{30 + 10} = 0.75. \tag{215}$$

$$\text{recall} = \frac{30}{30 + 20} = 0.6. \tag{216}$$

15. **Diskontirane nagrade**

Agent prejme zaporedne nagrade

$$5, \qquad 4, \qquad 3 \tag{217}$$

in uporablja diskontni faktor

$$\gamma = 0.5. \tag{218}$$

Izračunajte diskontirani povratek.

Rešitev.

$$G = 5 + 0.5 \cdot 4 + 0.5^2 \cdot 3 = 5 + 2 + 0.75 = 7.75. \tag{219}$$

16. **Q-learning**

Naj bo

$$Q(s,a) = 2, \qquad r = 3, \qquad \gamma = 0.8, \qquad \alpha = 0.5, \tag{220}$$

in

$$\max_{a'} Q(s',a') = 5. \tag{221}$$

Izračunajte novo vrednost $Q(s,a)$.

Rešitev.

Ciljna vrednost je

$$r + \gamma \max_{a'} Q(s',a') = 3 + 0.8 \cdot 5 = 7. \tag{222}$$

Zato je

$$Q(s,a) \leftarrow 2 + 0.5(7 - 2) \tag{223}$$
$$= 4.5. \tag{224}$$