Linearna regresija

Kiana Petrič, Nejc Ocepek, Timotej Vičar
Mentor: Žan Hafner Petrovski

Povzetek

V članku se spoznamo z osnovno metodo strojnega učenja, to je z linearno regresijo. Pri tem postopku želimo znanim podatkom prirediti premico, ki se jim čim bolj prilega, za izpeljavo njenega predpisa pa potrebujemo odvod. Prek znanega pojma naklona premice spoznamo njegov geometrijski pomen, ki nam koristi pri razumevanju iskanja ekstremov odvedljivih funkcij – taka pa je tudi funkcija cene regresijske premice, ki v svojem minimumu določi oba koeficienta premice.

1 Uvod

Pri linearni regresiji nas zanima linearna zveza v danih podatkih. Drugače povedano, iščemo premico, ki se podatkom prilega tako, da je vsota vseh razdalj med podatki in premico čim manjša. Za iskanje take premice je ključno poznavanje odvodov in reševanja ekstremalnih problemov.

Tako smo se trije mladi Marsovci pod vodstvom mentorja podali na raziskovanje zgoraj navedenih področij. Začeli smo z delom na odvodih in računskimi operacijami z njimi, kar je opisano v razdelku 2. Z osvojenim znanjem smo raziskavo preselili na naslednjo tematiko, in sicer na ekstremalne probleme. Začeli smo z lažjimi problemi, kot je maksimalna ploščina pravokotnika s konstantnim obsegom, nato pa smo obravnavali tudi nekoliko težje. Nazadnje smo se osredotočili na bistvo te naloge, torej linearno regresijo. Diskusijo smo začeli s samostojnim razmislekom o tem, kako bi lahko definirali premico, ki se dobro prilega podatkom, nato pa uporabili klasični postopek, ki je opisan v razdelku 4.

2 Odvod

2.1 Definicija

Začnimo kar z definicijo odvoda, njen geometrijski pomen bomo razložili kasneje [1].

Definicija 1. *Naj bo funkcija $f : (a-\delta, a+\delta) \to \mathbb{R}$ definirana v okolici točke $a \in \mathbb{R}$ za nek $\delta > 0$. Če limita*

$$\lim_{x \to \alpha} \frac{f(x) - f(a)}{x - a}$$

obstaja, pravimo, da je f ***odvedljiva*** *v točki a in*

$$f'(a) = \lim_{x \to \alpha} \frac{f(x) - f(a)}{x - a}$$

imenujemo ***odvod funkcije*** *f v točki a.*

Opomba 1. *Včasih namesto pisave zgoraj uporabimo*

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h},$$

kjer je $h = x - a$ sprememba argumenta x glede na točko a.

Vidimo, da smo pri definiciji uporabili pojem limite, zato podajmo še definicijo limite funkcije.

Definicija 2. *Naj bo $g : \mathbb{R} \to \mathbb{R}$ funkcija na realnih številih. Potem je L limita funkcije g v točki $x \in \mathbb{R}$, če za vsak $\epsilon > 0$ obstaja $\delta > 0$, da je*

$$|g(x + \delta) - g(x)| < \epsilon.$$

Če to velja, pišemo

$$L = \lim_{\delta \to 0} g(x + \delta).$$

Opazimo, da smo v primeru odvoda za funkcijo g uporabili

$$g(x) = \frac{f(x) - f(a)}{x - a}.$$

2.2 Pravila za računanje z odvodom

V tem razdelku iz definicije odvoda in pravil za računanje z limitami funkcij izpeljemo dve pravili za računanje odvodov, in sicer za odvod vsote in produkta dveh funkcij. V vseh primerih naj bosta funkciji g in d odvedljivi v točki x.

2.2.1 Vsota dveh funkcij

Pokažimo, da za $f(x) = g(x) + d(x)$ velja $f'(x) = g'(x) + d'(x)$. Računamo

$$\begin{aligned} f'(x) &= \lim_{h \to 0} \frac{f(x+h) - f(x)}{h} \\ &= \lim_{h \to 0} \frac{g(x+h) + d(x+h) - (g(x) + d(x))}{h} \\ &= \lim_{h \to 0} \frac{g(x+h) - g(x) + d(x+h) - d(x)}{h} \\ &= g'(x) + d'(x). \end{aligned}$$

Če je torej f vsota nekih funkcij, lahko odvajanje prenesemo kar na te funkcije.

2.2.2 Produkt dveh funkcij

Pokažimo še, da za $f(x) = g(x)d(x)$ velja $f'(x) = g'(x)d(x) + d'(x)g(x)$. Funkcijo f odvajamo po definiciji in dobimo

$$
\begin{aligned}
f'(x) &= \lim_{h\to 0} \frac{g(x+h)d(x+h) - g(x)d(x)}{h} \\
&= \lim_{h\to 0} \frac{g(x+h)d(x+h) - g(x)d(x+h) + g(x)d(x+h) - g(x)d(x)}{h} \\
&= \lim_{h\to 0} \frac{d(x+h)(g(x+h) - g(x))}{h} + \frac{g(x)(d(x+h) - d(x))}{h} \\
&= g'(x)d(x) + d'(x)g(x)
\end{aligned}
$$

Pri tem smo v drugem koraku uporabili trik, kjer smo v števcu prišteli in odšteli $g(x)d(x+h)$. Vidimo tudi, da pri produktu ne moremo kar prenesti odvajanja na faktorje kot pri vsoti.

2.2.3 Kvocient dveh funkcij

Zgolj navedimo še pravilo za odvajanje kvocienta funkcij. Za $f(x) = \frac{g(x)}{d(x)}$ velja

$$f'(x) = \frac{g'(x)d(x) - g(x)d'(x)}{d^2(x)}, \quad d(x) \neq 0.$$

2.3 Potenčna funkcija

Za izpeljavo odvoda potenčne funkcije bomo uporabili pravilo produkta in matematično indukcijo – to je postopek dokazovanja v dveh korakih. V prvem koraku dokazujemo, da trditev velja za osnovni primer oziroma neko začetno vrednost. Pogosto so v formulaciji problema uporabljena naravna števila, zato osnovnemu primeru pripada $n = 0$ ali $n = 1$. Ko dokažemo pravilnost trditve za osnovni primer se premaknemo na drugi korak, v katerem dokazujemo, da če trditev velja za $n-1$, sledi, da velja tudi za n. Z dokazom te izjave pokažemo, da trditev velja za vsak n.

Dokazati želimo, da za potenčno funkcijo $f(x) = x^n$ velja

$$f'(x) = nx^{n-1}.$$

Predpostavljamo torej, da za nek $n \in \mathbb{N}$ trditev velja za $n - 1$, to je

$$(x^{n-1})' = (n-1)x^{n-2}.$$

Pišimo zdaj funkcijo f kot produkt dveh faktorjev

$$f(x) = x \cdot x^{n-1}.$$

Na tem mestu lahko uporabimo pravilo za odvod produkta, torej

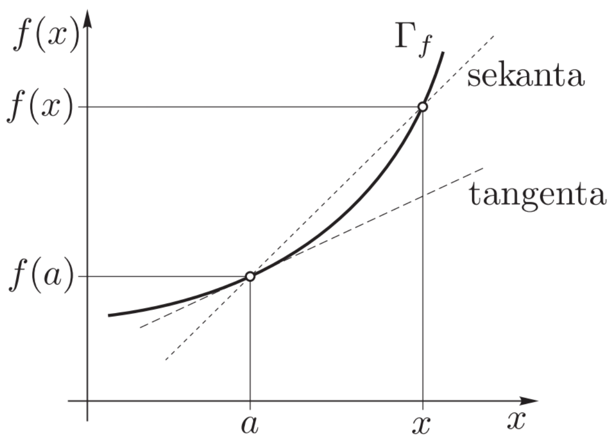
$$f'(x) = 1 \cdot x^{n-1} + x \cdot (x^{n-1})'$$

in z upoštevanjem predpostavke dobimo

$$\begin{aligned} f'(x) &= x^{n-1} + x \cdot (n-1)x^{n-2} \\ &= x^{n-1} + (n-1)x^{n-1} \\ &= nx^{n-1}. \end{aligned}$$

2.4 Geometrijski pomen odvoda

Kot obljubljeno, si oglejmo, kako lahko odvod razumemo tudi geometrijsko [1]. Naj bo funkcija f odvedljiva v točki a in naj točki $(a, f(a))$ in $(x, f(x))$ določata sekanto grafa funkcije f, torej premico skozi ti dve točki. To je prikazano na sliki 1.



Slika 1: Sekanta skozi točki $(a, f(a))$ in $(x, f(x))$ in tangenta na graf funkcije Γ_f v točki $(a, f(a))$.

Če x pošljemo proti a, se točka $((x, f(x))$ na grafu približuje točki $(a, f(a))$. Smerni koeficient sekante

$$\frac{f(x) - f(a)}{x - a}$$

se očitno približuje smernemu koeficientu tangente na graf v točki $(a, f(a))$, doseže pa ga v limiti, to je

$$f'(a) = \lim_{x \to \alpha} \frac{f(x) - f(a)}{x - a}.$$

Premico s smernim koeficientom $f'(a)$, ki gre skozi točko $(a, f(a))$, zato imenujemo tangenta na graf funkcije f v točki $(a, f(a))$. Njeno enačbo v implicitni obliki lahko zapišemo kot

$$y - f(a) = f'(a)(x - a).$$

Če povzamemo, vrednost odvoda funkcije f v točki a je ravno naklon tangente na graf funkcije f v točki a.

3 Ekstremalni problemi

Uporabimo zdaj odvod za reševanje ekstremalnega problema. Recimo, da nas zanima, kakšno mora biti razmerje stranic pravokotnika s podanim obsegom, da bo njegova ploščina največja. Obseg naj bo torej fiksen, najti pa želimo dolžine stranic pravokotnika, da bo ploščina čim večja.

Najprej iz formule za obseg pravokotnika $o = 2a + 2b$ izrazimo dolžino stranice a in dobimo

$$a = \frac{o - 2b}{2}.$$

V formulo za ploščino pravokotnika $p(a, b) = ab$ lahko zdaj namesto a vstavimo zgornji izraz. Dobimo predpis za ploščino, ki je odvisen zgolj od b

$$p(b) = \frac{o - 2b}{2} b = -b^2 + \frac{o}{2} b.$$

Graf funkcije p za primer, ko je $o = 1$, je prikazan na sliki 2, z uporabo pravil za odvajanje pa dobimo

$$p'(b) = -2b + \frac{o}{2}.$$

Pogoj, da v neki točki nastopi ekstrem funkcije, je, da je v tej točki odvod enak 0 – to lahko opazimo tudi na sliki 2. Kandidatne točke za ekstrem torej najdemo, če za b rešimo enačbo

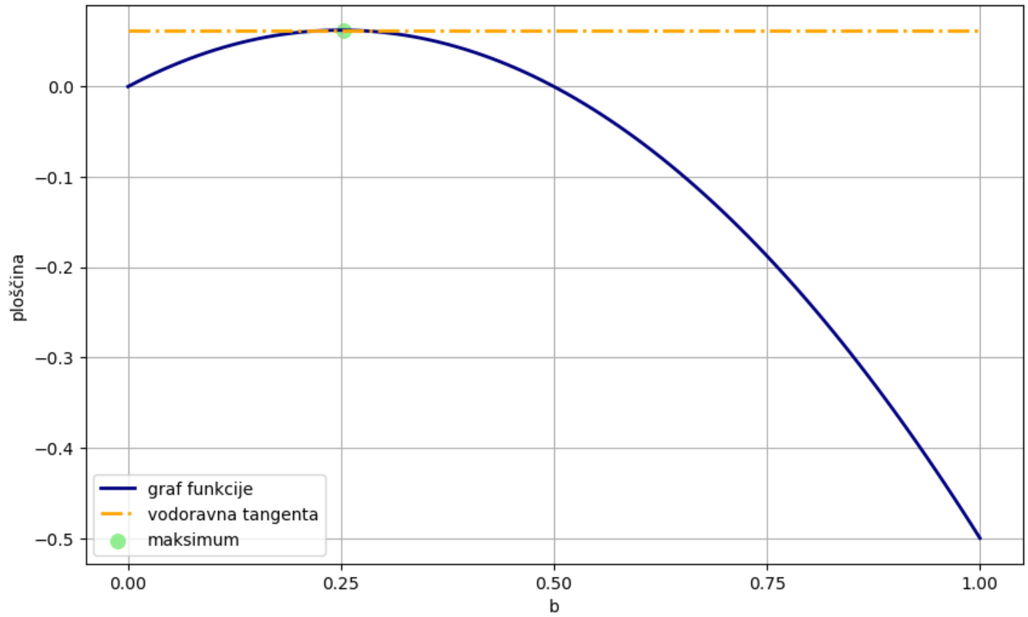
$$0 = -2b + \frac{o}{2}.$$

Izrazimo b

$$b = \frac{o}{4},$$

dobljen rezultat vstavimo še v formulo za obseg in dobimo vrednost za a

$$a = \frac{o}{2} - \frac{o}{4} = \frac{o}{4}.$$



Slika 2: Funkcija za ploščino pravokotnika v odvisnosti od b.

Ugotovili smo, da morata biti za dosego največje ploščine stranici pravokotnika enako dolgi. Z drugimi besedami to pomeni, da je ta pravokotnik kvadrat.

Z enako metodo lahko rešujemo tudi kompleksnejše probleme. V nadaljevanju izpeljemo osnovni primer strojnega učenja, to je linearna regresija. Pri izpeljavi s pridom uporabimo odvod, v bistvu rešimo ekstremalni problem.

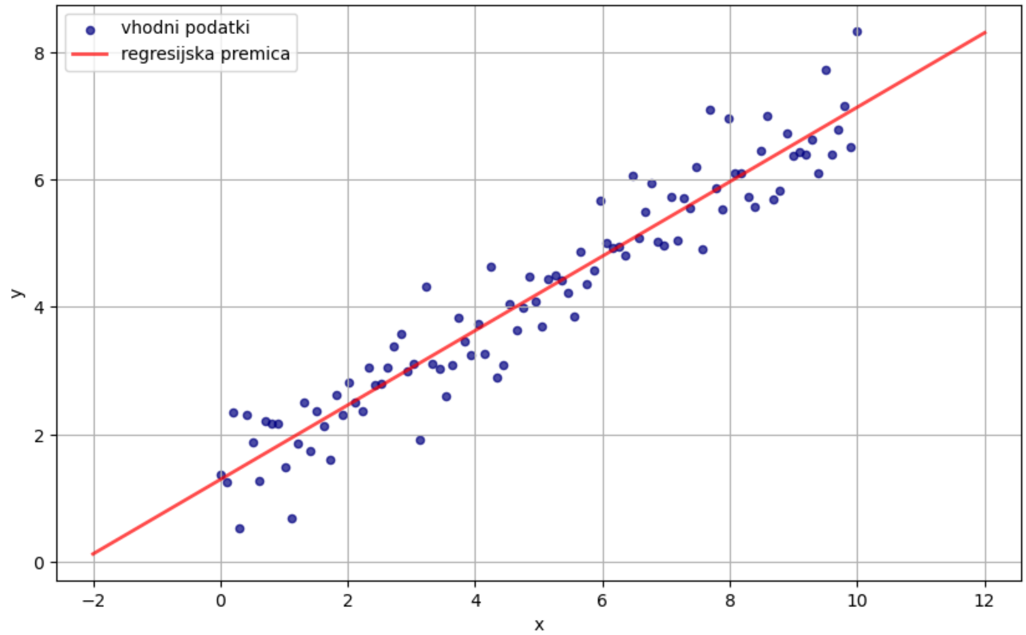
4 Linearna regresija

Ko imamo množico točk $T = \{(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)\}$, lahko med njimi iščemo neko povezavo. V primeru linearne regresije predpostavimo, da je ta povezava premica. Iščemo torej premico

$$f(x) = bx + a,$$

za katero bo za vse $i \in \{1, 2, \ldots, n\}$ vsota razlik med vrednostmi funkcije f v x_i in y_i čim manjša. Ta razlika bi bila lahko negativna – temu se izognemo tako, da razliko kvadriramo in dobimo vsoto, ki je odvisna od a in b:

$$S(a, b) = \sum_{i=1}^{n} (y_i - (bx_i + a))^2.$$



Slika 3: Linearna regresija.

Da bi našli iskano premico, moramo najti ustrezne vrednosti parametrov a in b. Pri iskanju optimalnih vrednosti teh parametrov si pomagamo z odvodi. V tem primeru imamo funkcijo dveh neodvisnih spremenljivk, zato so kandidatne točke za ekstreme tiste točke, v katerih je odvod po obeh spremenljivkah hkrati enak 0. To sicer v splošnem ne velja, je pa naša funkcija dovolj lepa, da se lahko poslužimo te metode. Začnemo torej z odvajanjem po a in dobimo

$$\frac{\partial f}{\partial a}(a,b) = -2\sum_{i=1}^{n}(y_i - bx_i - a).$$

Da bi našli minimum, ta odvod enačimo z 0, torej

$$0 = \sum_{i=1}^{n}(y_i - bx_i - a).$$

Iz dobljene enačbe izrazimo a in dobimo

$$a = \frac{1}{n}\sum_{i=1}^{n} y_i - \frac{b}{n}\sum_{i=1}^{n} x_i,$$

kjer opazimo, da v izrazu nastopata povprečji

$$\overline{y} := \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{in} \quad \overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Vrednost a lahko izrazimo s povprečji $\overline{x}$ in $\overline{y}$

$$a = \overline{y} - b\overline{x}.$$

Nadaljujmo zdaj še z odvajanjem f po b

$$\frac{\partial f}{\partial b}(a,b) = -2\sum_{i=1}^{n} x_i(y_i - bx_i - a).$$

Za kandidatne točke ponovno dobljeni odvod enačimo z 0 in vstavimo prej izraženi a

$$0 = \sum_{i=1}^{n} x_i(y_i - bx_i - \overline{y} + b\overline{x}).$$

Z nekaj računanja lahko izrazimo b

$$b = \frac{\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i(x_i - \overline{x})}$$

in z upoštevanjem $a = \overline{y} - b\overline{x}$ sledi

$$a = \overline{y} - \frac{\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i(x_i - \overline{x})}\overline{x}.$$

Tako dobimo končno obliko premice

$$y = \frac{\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i(x_i - \overline{x})}x + \overline{y} - \frac{\sum_{i=1}^{n} x_i y_i - \overline{y}\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} x_i(x_i - \overline{x})}\overline{x},$$

ki je za dane vhodne podatke narisana na sliki 3.

Opomba 2. *Linearna regresija je osnovni primer strojnega učenja. To je področje algoritmov, ki dane podatke uporabijo tako, da "naučen model" strojnega učenja za neke neznane oziroma nove podatke napove neko lastnost. V primeru linearne regresije smo iz množice točk* $\{(x_1, y_1), ..., (x_n, y_n)\}$ *dobili regresijsko premico, ki je bila naš "naučen model". Za vsako novo vrednost* x *lahko s pomočjo regresijske premice napovemo, kakšna je vrednost* y*, ki ji pripada – na tak način smo za nek neznani* x *na podlagi znanih podatkov napovedali vrednost* y*. Preostale metode strojnega učenja delujejo podobno, le redko pa lahko njihovo optimalno rešitev izrazimo natančno, kot smo to lahko storili v primeru linearne regresije.*

5 Zaključek

V članku smo predstavili pojem odvoda in izpeljali nekaj pravil za računanje z njim. Kot primer smo rešili lažji ekstremalni problem, nato pa isto metodo uporabili še pri izpeljavi regresijske premice, ki služi kot osnovni primer strojnega učenja.

Literatura

[1] J. Globevnik, *Analiza I*, 2016, DMFA - založništvo