1. domača naloga pri predmetu ITAP

Anamarija Potokar

30. marec 2025

1. naloga

a)

Pri tej podtočki sem dopolnila funkciji za iskanje koeficientov linearne regresije in napovedovanje koeficientov linearne regresije po definicji. Ciljna spremenljivka je bila določena s 1. in 2. napovedno spremenljivko, zato sem pričakovala vrednosti 1. in 2. koeficienta 1 in 5, vrednosti ostalih koeficientov pa blizu 0. Koeficienti, ki sem jih dobila, so $[-2.00027892 \cdot 10^{-15}, 1.00000000, 5.00000000, -1.12864894 \cdot 10^{-14}, 8.82392934 \cdot 10^{-15}]$, kar je zelo blizu pričakovanjem.

b)

S 5−kratnim prečnim preverjanjem sem preverila natančnost linearne regresije iz točke a), kjer sem ugotovila, da model ni ravno natančen in se koeficienti med različnimi foldi zelo razlikujejo.

c)

Izpisala sem korelacijsko matriko napovednih spremenljivk in ugotovila, da so stolpci 0, 4 in 5 v matriki X zelo korelirani. Koreliranost napovednih spremenljivk sem tudi vizualizirala. Problema bi se lahko poskusili rešiti tako, da bi odstranili stolpca 0 in 4, ker sta najbolj korelirana, in ohranili le stolpec 5.

d)

V tej podtočki sem dopolnila funkcijo, ki poišče koeficiente Tihonova, ki kaznuje velike koeficiente tako, da jim prišteje člen λ in s tem zmanjša njihov vpliv na ciljno spremenljivko.

e)

Ponovno sem s 5−kratnim prečnim preverjanjem preverjala natančnost linearne regresije, kjer sem zdaj koeficiente poiskala z novo funkcijo. To se je za naš primer izkazalo za uspešno, povprečna napaka med različnimi foldi in variabilnost koeficientov sta se namreč močno zmanjšali.

2. naloga

V tej nalogi sem najprej za iskanje optimalnega k standardizirala podatke, saj so posamezne napovedne spremenljivke podane v različnih enotah, kar brez standardizacije lahko negativno vpliva na model. Uporabila sem stratificirano vzorčenje, da so v vsakem foldu razredi ciljne spremenljivke bili približno enako porazdeljeni, ter nastavila parameter shuffle na True, kar prav tako zmanjša nihanje rezultatov med različnimi foldi. Preizkusila sem k od 1 do 30, za vsakega izračunala povprečno točnost 5 foldov in kot optimalno vrednost izbrala k, kjer je bila dosežena najvišja povprečna točnost. S tem modelom sem za najboljšega dobila $k = 8$.
Potem sem obtežila teh najbližjih k točk tako, da bližje, kot je sosed, večjo utež dobi, in to se je izkazalo za uspešno, saj se je povprečna točnost povečala, varianca pa zmanjšala, in novi najboljši k je bil enak 16.
V nadaljevanju sem preverila še, ali so katere izmed napovednih spremenljivk zelo korelirane, in ugotovila, da je korelacija med ravnjo drobnih in grobih trdih delcev zelo visoka (0.9730048883255122), zato sem preverila, ali se model izboljša, če enega izmed teh stolpcev odstranim, vendar pa je bila razlika zelo majhna.
Zdelo se je, da to, kdaj je bila opravljena zadnja meritev, ne vpliva na kvaliteto zraka, zato sem odstranila še zadnji stolpec. V tem primeru se je točnost povečala za 0.46%, varianca pa se je sicer tudi povečala, a je bila sprememba tako majhna, da je zanemarljiva. Torej je bilo v našem primeru bolje, da ne upoštevamo spremenljivke x_{10}, in v tem primeru dobimo, da je optimalen $k = 12$.