Poročilo o nalogi 3

Tim Kmecl, UOZP 2022

Predtekmovanje

Predobdelava podatkov

Podatke iz csv datoteke sem prebral v pandas dataframe. Ciljna spremenljivka y je razlika med časom prihoda in odhoda (trajanje vožnje), za nadaljnjo obdelavo pa ohrani le linijo (ki sicer v preddtekmovanju ni pomembna) in datum/čas odhoda (oba atributa sta uporabljena le za predobdelavo podatkov, ne pa tudi neposredno v regresiji). Doda sledeče 0/1 atribute: ali je dan v tednu eden od prvih štirih dni (torej delovnik brez petka), petek, sobota ali nedelja. Ti atributi bodo torej imeli vrednost 1, če je vožnja na ustrezen dan v tednu, in 0 sicer.

Iz spleta sem pridobil še podatke o šolskih počitnicah tega leta. Dodaten atribut počitnice bo enak 1, če je vožnja v času počitnic, sicer 0. Podobno sem naredil še za praznike, ki so bili takrat med tednom (ne pa tudi v soboto ali nedeljo, saj sem predpostavil, da bo vpliv praznika dosti večji med tednom).

Na koncu sem v one-hot-encoding pretvoril še uro odhoda, in sicer sem razdelil celoten dan na 5-minutne intervale, ter tako dobil še 24*12 atributov. Skupaj je atributov torej 6 + 12*24=294.

Napovedovanje

Za napovedovanje sem uporabil linearno regresijo z L2 regularizacijo. Parameter regularizacije alfa sem poskusil določiti s preizkušanjem različnih vrednosti, pri čemer sem dano učno množico v ta namen naključno razdelil na učno in testno. Najbolje (do reda velikosti natančno) se je obnesel alfa=1. napovedi linearnega modela se na koncu prišteje še čas odhoda in tako dobi čas prihoda, ki se jo shrani kot rezultat.

Rezultat te metode na preddtekmovanju znaša 129 (Ista oddaja na lestivi: 140)

Tekmovanje

Predobdelava

Predobdelava podtkov za tekmovanje je enaka kot za tekmovanje

Napovedovanje

Predobdelane podatke nato model razdeli glede na linijo in za vsako zgradi svoj linearni model z L2 regularizacijo (alfa=1, kot pri predtekmovanju). Poleg tega zgradi še en ''povprečen'' model na vseh podatkih.

Za napovedovanje neznanih časov prihoda model za vsako od linij, za katere ima shranjen linearni model, uporabi pripadajoč model. Če se v teh podatkih slučajno pojavi linija, ki je v učnih podatkih ni bilo, uporabi za napoved povprečni model, zgrajen na vseh učnih podatkih. Na koncu prišteje čase odhoda in tako dobi napoved za čas prihoda.

Rezultat metode na tekmovanju (oddaja na lestvici) znaša 173.