

NSU 2022/23 – peta domača naloga

5.6.2023

Na voljo sta dve nalogi, osnovna in napredna. Vsaka je vredna 10 točk. Rešuješ lahko samo osnovno nalogo, ali pa obe.

1 Časovne vrste

Datoteka **okuzeni.csv** vsebuje podatke o številu aktivno okuženih oseb med epidemijo Covid-19 v Sloveniji. Vsaka vrstica predstavlja en dan, vsak stolpec pa eno od slovenskih občin. Tvoj cilj je napovedovanje števila aktivno okuženih v vsaki od slovenskih mestnih občin (Ljubljana, Maribor, Kranj, Koper, Celje, Novo mesto, Velenje, Nova Gorica, Krško, Ptuj, Murska Sobota, Slovenj Gradec).

Razvij model, ki bo na podlagi podatkov N prejšnjih dni napovedoval število okuženih za M prihodnjih dni. Izbira N je tvoja, za M pa se omejimo na en teden ($M=7$) in en mesec ($M=30$).

Model preizkusi na celotnem naboru podatkov. Na primer, če je $N=20$ in $M=7$, je prvi testni primer sekvenca z indeksi na intervalu $[20,27)$, drugi testni primer $[21,28)$, itd. do $[793, 800)$.

Analiziraj tudi, kako se napaka napovedi spreminja z oddaljenostjo napovedi. Na primer, pri $M=7$ nas zanima, kakšna je napaka napovedi za jutri v primerjavi z napako napovedi za pojutrišnjem, ali pa čez cel teden.

Poročilo naj vsebuje:

1. podrobnosti uporabljenih metod strojnega učenja,
2. opis postopkov učenja in testiranja,
3. povprečni testni napaki izbranih modelov ($M=7$, $M=30$),
4. graf primerjave napovedi modela $M=7$ s pravimi vrednostmi za mestno občino z najboljšimi napovedmi in za mestno občino z najslabšimi napovedmi,
5. analizo odvisnosti napake od časovne oddaljenosti napovedi (npr. graf povprečne napake v odvisnosti od oddaljenosti napovedi).

2 Časovne vrste na grafih

Datoteka **sosedi.csv** vsebuje informacije o sosedih posameznih občin v Sloveniji. V datoteki sta dve vrstici; vsako sestavljajo imena občin, ločena s presledki. Par (i -ta občina iz prve vrstice, i -ta občina iz druge vrstice) pomeni, da sta ti občini sosedni. Sestavimo lahko torej graf, v katerem so občine vozlišča, povezane pa so sosednje občine.

Tvoj cilj je preizkusiti, ali lahko z uporabo informacije o sosesčini občin razvijemo model, ki bolje napoveduje število aktivno okuženih v posameznih mestnih občinah od modela, ki si ga razvil.a v prvem delu naloge.

Dodatne informacije lahko izkoristimo na (vsaj) dva načina:

1. uporabljamo samo podatke o številu aktivno okuženih v mestnih občinah ter graf sosesčine,

2. uporabljamo tudi podatke o številu aktivno okuženih v vseh ostalih slovenskih občinah (celotna datoteka **okuzeni.csv**).

Poročilo naj vsebuje:

1. podrobnosti uporabljenih metod strojnega učenja,
2. opis postopkov učenja in testiranja,
3. primerjavo obeh zgoraj predlaganih pristopov,
4. primerjavo uspešnosti metod iz obeh delov domače naloge,
5. komentar o koristnosti informacij o soseščini občin pri napovedovanju števila okuženih.