Upute za projekt

Uvod u znanost o podacima - Akademska godina 2023./2024.

Izbor članka

Na početku semestra studenti odabiru jedan od ponuđenih znanstvenih članaka za svoj projektni rad. Na svaki ponuđeni članak može se prijaviti maksimalno 15 studenata. Studenti koji ne izaberu članak do zadanog roka će biti nasumično raspoređeni po člancima na kojima je preostalo slobodnih mjesta.

Ponuđeni članci (i nadležni asistenti) se mogu vidjeti u ovoj tablici. U istoj tablici studenti i odabiru članke tako da ispod željenog članka upišu svoje ime i prezime.

Nakon isteka roka za odabir članka, studenti su dužni napraviti odgovarajući GitHub repozitorij i dodati nadležnog asistenta kao kolaboratora (Settings → Collaborators → Add people).

Rok za odabir članaka: **13.10.2023.**

Projekt će se sastojati od dva dijela, s tri točke provjere. U prvom dijelu, studenti rade individualno na replikaciji rezultata znanstvenog članka. Prvi dio ima dvije točke provjere: pripremu i vizualizaciju podataka, i replikaciju rezultata. Drugi dio projekta se radi u grupi od tri studenta (ili eventualno dva u konzultaciji s asistentom). U ovom dijelu cilj je poboljšati rezultate članka u nekom aspektu. U slučaju eventualnih problema ili nejasnoća tijekom projekta, studenti se uvijek mogu javiti asistentu zaduženom za izabrani članak.

Priprema i vizualizacija podataka

Cilj ovog dijela projekta je upoznati se s podacima. Studenti trebaju pročitati odabrani članak te preuzeti podatke koji su korišteni. Nakon toga potrebno je upoznati se s podacima. Smjernice kako to učiniti su:

- učitati podatke
- provjeriti koji sve tipovi podataka postoje i prikazati deskriptivnu statistiku podataka
- provjeriti postoje li nedostajuće vrijednosti i stršeće vrijednosti
- vizualizirati podatke na nekoliko različitih načina (npr. histogram značajki, linijski dijagrami vremenskih nizova, točkasti dijagrami u ovisnosti o ciljnoj klasi, …)
- …

Studenti nisu ograničeni isključivo na ranije navedene stavke, one služe samo kao smjernice. Tijekom izrade ovog dijela projekta, ne morate se previše obazirati na stvari koje su radili autori u znanstvenom radu, već se trebate samostalno i prema vlastitom nahođenju upoznati s korištenim skupom podataka.

Rok predaje: **10.11.2023**.
Maksimalni broj bodova: **10**

Replikacija rezultata

Za drugu provjeru u prvom dijelu projekta studenti koriste pristupe iz odabranog članka kako bi replicirali prikazane rezultate. Pri samoj implementaciji studenti mogu koristiti već implementirane funkcije iz paketa kao što su *numpy*, *scikit-learn* i sl. Jednom kada su metode implementirane, potrebno ih je pokrenuti na ranije pripremljenim podacima, korektno evaluirati, usporediti s rezultatima iz članka i objasniti eventualne razlike.

Primjerice, ako radite na klasifikacijskom problemu preporučljivo je prikazati:

- vrijednost metrika kao što su točnost, preciznost, odziv, itd.
- AUC/ROC krivulje
- matrice zabune
- . . .

Studenti, naravno, slobodno mogu prikazati rezultate i na alternativne načine koji im se čine interesantni. U slučaju da se izabrani članak bavi specifičnom problematikom, preporučljivo je kontaktirati nadležnog asistenta za savjet.

Rok predaje: **15.12.2023.**
Maksimalni broj bodova: **20**

Poboljšanje rezultata

U drugom dijelu projekta cilj je poboljšati rezultate znanstvenog članka. Drugi dio projekta se radi u grupama od tri studenta (eventualno dva). Neki od načina na koje možete pokušati poboljšati rezultate su:

- ispraviti eventualne nedostatke članka
- proširiti podatke inženjerstvom značajki
- primijeniti neke od metoda/algoritama obrađenih na predavanjima
- optimirati parametre
- kreirati vlastitu metodu
- …

Svako potencijalno poboljšanje mora biti zasebno evaluirano i uspoređeno s originalnim rezultatima članka. U slučaju da poboljšanje nije postignuto, potrebno je komentirati koji je uzrok tome i što bi se još moglo pokušati a da je trenutno izvan vaših mogućnosti.

Na kraju ove faze potrebno je napraviti i kratku video prezentaciju u trajanju od 4 minute u kojoj ćete prezentirati članak i isprobana poboljšanja. Video prezentaciju potrebno je postaviti u GitHub repozitorij.

Nekoliko najboljih poboljšanja rezultata članaka izabrat će asistenti i predložiti za prezentaciju (puštanje snimljene video prezentacije + diskusija) tijekom zadnjeg tjedna nastave.

Rok predaje **19.01.2024.**
Maksimalni broj bodova: **10**

Dodatne napomene

- Sve navedeno radite koristeći Python i Jupyter bilježnice. Konačna verzija bilježnice koju ćete predati (u svakoj točki provjere projekta) mora sadržavati komentare/zaključke svih napravljenih koraka. Bilježnicu mora biti moguće pratiti bez previše čitanja kodova. Bilježnica se predaje postavljanjem na GitHub repozitorij.

- Ako nemate računalo s dovoljno resursa za odraditi sve što je potrebno u sklopu projekta, predlažemo da koristite Google Colab.

- Na kraju svake faze projekta asistenti će sa svakim studentom zasebno proći kroz predanu Jupyter bilježnicu pri čemu će student objasniti implementirani kod, rezultate i zaključke. Na temelju tog ispitivanja i predanog rješenja, asistent će dodijeliti studentu odgovarajući broj bodova.

- U slučaju da student zakasni s predajom rješenja za određenu fazu projekta, taj dio projekta će biti bodovan sa 0 bodova.

- Ako student ne skupi minimalno 25% bodova iz projekta (10 bodova), student nema pravo izlaska na završni ispit (i ispitne rokove).

- Ponavljačima predmeta priznaje se uspješno polaganje projekta u prethodnoj akademskoj godini 2022./2023. (min. 10 prikupljenih bodova). Bodovi će im se prebaciti automatski krajem semestra. U slučaju da ponavljači žele ponoviti projekt, trebaju se zapisati na neki članak i obavijestiti nadležnog asistenta da će ponavljati projekt najkasnije do 13.10.2023.