Uputstvo za izradu projekta

Projekat se radi u timovima od najviše **3 člana**. Svaku temu može da radi samo jedna ekipa. Dovoljno je da **jedan član tima prijavi temu** tako što će u mejlu navesti:

- ime i prezime svih članova tima,
- naziv tima,
- temu koju će obrađivati.

Prijava teme i sva pitanja u vezi sa projektom šalju se na mejl: nikola.vucic@matf.bg.ac.rs.

Napomena: Studenti mogu predložiti i temu van spiska, uz kratko obrazloženje zašto je tema relevantna i ukratko šta planiraju da rade.

Udruživanje timova

Postoji mogućnost da se dva tima udruže i međusobno uporede konačne rezultate. Na primer, oba tima uzimaju istu bazu podataka i primenjuju različite metode (npr. različite algoritme klasterovanja), a zatim analiziraju u čemu se rezultati podudaraju, a u čemu razilaze i zašto. U tom slučaju svaki tim i dalje brani svoj deo samostalno i dobija poene nezavisno, ali se od njih očekuje i zajednički deo u kome komentarišu razlike i sličnosti između primenjenih pristupa.

Literatura

Svaki projekat treba da sadrži i **spisak sve korišćene literature** – knjige, naučni radovi, blogovi, dokumentacija biblioteka i slično. Literatura treba da bude navedena na kraju predatog materijala, u standardnom formatu (autor, naslov, godina, izvor/link).

Obaveze i bodovanje

Branjenje projekta sastoji se iz dve obaveze:

1. **Predispitna obaveza** (40 poena od 100)
   Cilj predispitne obaveze je da studenti objasne matematičku pozadinu za temu koju su izabrali i ispričaju motivaciju – zašto je važno to znati i kako se koristi u praksi. Predispitna obaveza se može braniti od **4. maja do 19. juna** u terminu vežbi, uz najavu na mejl bar **2 dana ranije**.

2. **Ispitna obaveza** (60 poena od 100)
   Cilj ispitne obaveze je pronaći bazu podataka na internetu koja zadovoljava normalne uslove (da može da se radi sa njom u smislu dimenzije), najpre je srediti (očistiti i pripremiti), potom analizirati podatke i na njoj implementirati glavnu ideju projekta.

   - **Predrok:** branjenje je moguće od **29. maja do 19. juna** u terminu vežbi, u trajanju od oko 20 minuta.
   - **Ispitni rok:** termin odbrane je po dogovoru u okviru ispitnog roka.

Projekat treba predati u jednom od sledećih formata:

- **Jupyter Notebook** sa komentarima i teorijskom pozadinom ispisanom u Markdown ćelijama, ili
- **Python kod** (standardni `.py` fajl) sa propratnim tekstom u PDF formatu.

Predlog tema za projektni rad

Pored svake teme navedeno je maksimalni **broj članova** koji mogu da rade tu temu. Teme sa većom teorijskom dubinom ili više podtema dozvoljavaju veće timove.

Regresione metode (max. 2 člana)

Regresione metode omogućavaju modelovanje odnosa između promenljivih i predviđanje neprekidnih vrednosti (na primer predviđanje cena akcija ili vrednosti imovine). Među najčešće korišćenim regresionim metodama su linearna regresija, polinomijalna regresija, Ridge i Lasso regresija, kao i naprednije metode poput Random Forest, XGBoost i neuronskih mreža, koje su posebno korisne za složene i nelinearne odnose u podacima. Dve takođe vrlo značajne metode su Nadaraja-Votson regresija i Bajesovska regresija. Nadaraja-Votson metoda je klasičan primer kernel-based regresione metode i predstavlja osnovu za mnoge modernije metode. Bajesovska linearna regresija je zanimljiva zbog svoje probabilističke prirode, jer se koeficijenti modela posmatraju kao slučajne veličine sa odgovarajućim raspodelama, umesto kao fiksne vrednosti. Zahvaljujući tome, predikcija modela nije samo tačkasta procena, već cela distribucija mogućih ishoda, što omogućava kvantifikaciju nesigurnosti u prognozama i bolju interpretaciju rezultata, posebno kod malih skupova podataka.

1. Bajesovska linearna regresija link1 link2 link3
2. Nadaraja-Votson regresija link1 link2

Redukcija dimenzije podataka (max. 3 člana)

Redukcija dimenzije podataka omogućava smanjenje broja promenljivih u skupu podataka, ali tako da se očuva što više informacija početnog skupa. Ove tehnike su korisne kada radimo s visokodimenzionim podacima (a to se vrlo često dešava), gde preveliki broj promenljivih može dovesti do prevelike vremenske složenosti, odnosno do usporenog treniranja modela mašinskog učenja. Takođe, ove tehnike se često koriste i u obradi slike,

analizi genetskih podataka i vizualizaciji kompleksnih skupova podataka. PCA (Principal Component Analysis) je klasična tehnika redukcije dimenzija, i dalje se koristi, ali se najčešće kombinuje sa modernijim metodama. UMAP (Uniform Manifold Approximation and Projection) i t-SNE (t-distributed Stochastic Neighbor Embedding) su novije metode za redukciju dimenzija koje omogućavaju vizualizaciju visokodimenzionalnih podataka. UMAP se zasniva na algebarskoj topologiji i teoriji mnogostrukosti, dok je t-SNE probabilistička metoda zasnovana na Studentovoj raspodeli. UMAP je računski brži i skalabilniji od t-SNE metode, pa je sve češće u upotrebi.

1. PCA link1 link2 link3
2. t-SNE link1 link2 link3
3. UMAP link1 link2 link3

Klasterovanje podataka (max. 3 člana)

Klasterovanje je grupisanje podataka prema njihovim sličnostima. Najčešće se koristi u analizi klijenata i segmentaciji tržišta. K-means je jedan od najbržih algoritama, efikasan za velike skupove podataka, ali loše prepoznaje klastere nepravilnog oblika i zahteva unapred zadat broj klastera. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ne zahteva broj klastera unapred i dobro detektuje nepravilne oblike i šum u podacima. Hijerarhijsko klasterovanje bolje radi na manjim skupovima podataka, ali je računski skuplje od K-means i DBSCAN algoritama. GMM (Gaussian Mixture Models) je probabilistički pristup klasterovanju koji svaki klaster modeluje Gausovom raspodelom i dodeljuje svaku tačku klasteru sa određenom verovatnoćom – za razliku od K-means koji koristi "čvrstuđodelu. Ovo je posebno pogodna tema za udruživanje timova, gde različiti timovi mogu primeniti različite algoritme na istu bazu podataka i uporediti rezultate.

1. K-means algoritam link1 link2 link3
2. Hijerarhijsko klasterovanje link1 link2 link3
3. DBSCAN link1 link2
4. GMM (Gaussian Mixture Models) link1 link2 link3

Skrejpovanje podataka (max. 2 člana)

Skrejpovanje podataka (web scraping) je tehnika automatskog prikupljanja podataka sa veb-stranica. Koristi se kada podaci nisu dostupni putem zvaničnog API-ja ili u strukturiranom formatu, a neophodni su za dalju analizu. Biblioteke poput `BeautifulSoup` i `Scrapy` u Pythonu omogućavaju parsiranje HTML sadržaja i ekstrakciju željenih informacija. Važno je poštovati uslove korišćenja sajta i zakonske propise koji regulisŭ automatsko prikupljanje podataka. Za efikasno čuvanje i razmenu prikupljenih podataka preporučuje se Parquet format – kolonarni format koji za razliku od CSV-a podržava kompresiju, brže čitanje kolona i kompatibilnost sa alatima poput Apache Spark-a i Pandas-a, što ga čini odličnim izborom za veće skupove podataka.

1. Web scraping sa BeautifulSoup link1 link2 link3
2. Web scraping sa Scrapy link1 link2 link3

Klasifikacione metode (max. 3 člana)

Klasifikacija je jedan od osnovnih zadataka nadgledanog mašinskog učenja, gde je cilj na osnovu ulaznih obeležja predvideti kojoj klasi pripada dati primer. Primene su izuzetno široke – od detekcije spama i dijagnoze bolesti do prepoznavanja slika i kreditnog skoringa. Matematička pozadina varira u zavisnosti od metode: logistička regresija se zasniva na sigmoid funkciji i maksimizaciji verodostojnosti, SVM traži hiperravan koja maksimizuje marginu između klasa korišćenjem konveksne optimizacije, dok KNN klasifikuje tačke na osnovu blizine u prostoru obeležja. Stabla odlučivanja i ansamblovske metode (Random Forest, Gradient Boosting) kombinuju više slabih klasifikatora u jedan jak, što ih čini posebno robusnim na šum u podacima. Za evaluaciju klasifikatora koriste se metrike poput tačnosti (accuracy), preciznosti, odziva, F1 skora i ROC krive.

1. Logistička regresija (1 član) link1 link2 link3
2. SVM (Support Vector Machine) link1 link2 link3
3. KNN (K-Nearest Neighbors) (1 član) link1 link2 link3
4. Stabla odlučivanja i Random Forest link1 link2 link3
5. Gradient Boosting (XGBoost, LightGBM) link1 link2 link3
6. Evaluacione metrike za klasifikaciju (1 član) link1 link2 link3

Preporučivački sistemi (max. 3 člana)

Preporučivački sistemi automatski predlažu sadržaje korisnicima na osnovu njihovih prethodnih interakcija ili sličnosti sa drugim korisnicima. Sveprisutni su u svakodnevnom životu – Netflix, Spotify, Amazon i YouTube ih intenzivno koriste. Postoje dva osnovna pristupa: kolaborativno filtriranje, koje se zasniva na matrici interakcija korisnik–stavka i pokušava da popuni nedostajuće vrednosti (faktorizacija matrica, SVD), i filtriranje zasnovano na sadržaju, koje kreira profil korisnika na osnovu karakteristika stavki koje mu se dopadaju. Hibridni pristupi kombinuju oba. Poseban izazov je problem hladnog starta – šta preporučiti novom korisniku ili za novu stavku o kojoj nema dovoljno podataka.

1. Kolaborativno filtriranje link1 link2 link3
2. Filtriranje zasnovano na sadržaju link1 link2 link3
3. Faktorizacija matrica (SVD, NMF) link1 link2 link3

Otkrivanje anomalija (max. 2 člana)

Otkrivanje anomalija (anomaly detection) podrazumeva identifikaciju tačaka u skupu podataka koje se značajno razlikuju od očekivanog ponašanja. Primena je široka: detekcija prevara u finansijskim transakcijama, otkrivanje kvarova u industrijskim sistemima, sajber-bezbednost i medicinska dijagnostika. Metode se dele na statističke (zasnovane na pretpostavci o raspodeli podataka), metode zasnovane na blizini (LOF, DBSCAN) i metode zasnovane na izolaciji (Isolation Forest). Autoenkoderske neuronske mreže su moderno rešenje koje uči kompaktnu reprezentaciju normalnih podataka, pa anomalije prepoznaje po visokoj grešci rekonstrukcije. Poseban izazov je nebalansiranost podataka – anomalije su po definiciji retke.

1. Isolation Forest link1 link2 link3
2. Local Outlier Factor (LOF) link1 link2 link3
3. Autoenkoderi za detekciju anomalija (može i 3 člana) link1 link2 link3

Bajesovske mreže i grafički modeli (max. 3 člana)

Bajesovske mreže su probabilistički grafički modeli koji predstavljaju skup slučajnih promenljivih i njihove uslovne zavisnosti pomoću usmerenog acikličkog grafa (DAG). Svaki čvor u grafu odgovara jednoj promenljivoj, a grana između dva čvora označava uslovnu zavisnost. Ova reprezentacija omogućava efikasno zaključivanje i učenje iz podataka. Primene su raznovrsne: medicinska dijagnoza (sistem ekspert koji zaključuje o bolesti na osnovu simptoma), NLP, robotika i analiza rizika. Markovljeve mreže (neusmereni grafički modeli) su srodna klasa modela koja se često koristi u obradi slike i prostornoj statistici. Algoritmi zaključivanja poput Belief Propagation i Variacionog zaključivanja omogućavaju efikasno računanje marginalnih raspodela i aposteriorne verovatnoće.

1. Bajesovske mreže – osnove link1 link2 link3
2. Učenje strukture Bajesovskih mreža link1 link2 link3
3. Markovljeve mreže i Markovljeva polja link1 link2 link3
4. Belief Propagation i zaključivanje link1 link2 link3