Uputstvo za izradu seminarskog rada

Seminarski rad se radi u timovima od najviše 3 člana (istu temu može da radi najviše 2 tima). Rad treba da sadrži sledeće delove: teorijski uvod, primenu na konkretnim podacima, zaključak i spisak korišćene literature. Rok za izradu seminarskog rada je kraj ispitnih rokova. Moguće je osvojiti maksimalno 50 poena za napisani rad i maksimalno 50 poena za odbranu rada. Svi članovi tima dobijaju isti broj poena za rad, dok su poeni za odbranu individualni.

Prijavljivanje teme i slanje radova se vrši na mejl srbakoski998@gmail.com. Prilikom prijave potrebno je navesti članove tima, naziv teme, kratak plan obrade teme i predlog skupa podataka na kojem ćete primeniti izabranu metodu (npr. na sajtu Kaggle možete pronaći veliki broj skupova podataka). Rad se radi u R-u ili Python-u. Radove treba poslati u pdf ili html formatu, gde treba da budu obuhvaćeni i teorija i kodovi. Prilikom slanja rada poželjno je da predložite i dva termina za odbranu koja vam najviše odgovaraju. Odbrana treba da traje maksimum 20 minuta.

U nastavku je dat spisak tema, ali možete da predložite i neku temu van spiska.

TEME ZA SEMINARSKI RAD:

- **Regresione metode**

  Regresione metode omogućavaju modelovanje odnosa izmedju promenljivih i predvidjanje neprekidnih vrednosti (na primer predvidjanje cena akcija ili vrednosti imovine). Medju najčešće korišćenim regresionim metodama (biće obradjene na kursevima u četvrtoj godini) 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 takodje 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**

  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. Takodje, 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**

  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. K-means predstavlja osnovu za mnoge naprednije metode klasterovanja. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) ne zahteva broj klastera unapred i dobro detektuje nepravilne oblike i šum u podacima. Sporiji je od K-meansa na velikim datasetovima, ali bolje radi kada klasteri imaju različite gustine. Hijerarhijsko klasterovanje bolje radi na manjim skupovima podataka, ali je računski skuplje od K-means i DBSCAN algoritama.

  1. K-means algoritam link1 link2 link3
  2. Hijerarhijsko klasterovanje link1 link2 link3
  3. DBSCAN link1 link2