Mašinsko učenje 2020

Zadatak 2

Sadržaj

- Zadatak 1 - Rekapitulacija
- Zadatak 2

Zadatak 1 - Rekapitulacija

Zadatak 1 - Rekapitulacija

- Procenat uspešnosti: **84.85%** (28/33).
- Sva rešenja sa **RMSE ≤ 75.0** se smatraju odličnim.
- Najveće preklapanje izvornih kodova prema alatu za detekciju plagijata: **20%.**

- Najbolji rezultati po terminima:

Termin	Tim	RMSE
Ponedeljak	**mjau_mace**	**68.57**
Utorak 1	**tim8_20**	**68.91**
Utorak 2	**tim20_20**	**68.91**
Četvrtak	**tim7_20** & **tim18_20**	**68.91**
Petak	**pt-pk**	**68.60**

Zadatak 1 - Rekapitulacija

- Dobre stvari (na nivou generacije):
  - Vizualizacija podataka;
  - Rad sa *outlier*-ima;
  - Implementacija algoritama;
  - Računanje metrike (RMSE);
  - Propratni izveštaji.

- Stvari koje mogu biti bolje (na nivou generacije):
  - Rad sa trening skupom podataka.

Zadatak 2

Zadatak 2

- Višestruka regresija:
  - Prediktovati platu (kolona **plata** u dolarima) nastavnog osoblja u SAD na osnovu više atributa.
  - Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije **RMSE (Root Mean Square Error) manji od 28500.**
  - Algoritme mašinskog učenja implementirate sami - **zabranjena upotreba algoritama iz gotovih biblioteka!**
  - Rok za izradu zadatka: **30.04.2020. u 23:59h**.
  - Instalirane biblioteke za Zadatak 2 (verzije date u Uputstvu):
    - **Numpy**
    - **Pandas**
    - **SciPy.**

Zadatak 2

- Atributi (kolone) na osnovu kojih se prediktuje **plata**:
  - **zvanje** - nastavno zvanje:
    - **Prof** - redovni profesor
    - **AssocProf** - vanredni profesor
    - **AsstProf** - docent
  - **oblast** - oblast istraživanja:
    - **A** - teorijska
    - **B** - primenjena
  - **godina_doktor** - broj godina protekao od doktoriranja
  - **godina_iskustva** - broj godina radnog staža u nastavi
  - **pol** - pol:
    - **Female** - ženski
    - **Male** - muški.

Zadatak 2

- Atributi (kolone) **zvanje**, **oblast** i **pol** sadrže kategoričke podatke.
- Neke od tehnika za rad sa kategoričkim podacima su:
  - **Label Encoding** - konvertovanje kategoričkih podataka u broj iz opsega [0, broj_klasa-1], npr.: za kolonu **oblast,** vrednosti **[A, B]** će se konvertovati u vrednosti **[0, 1].**
  - **One Hot Encoding** - konvertovanje svake klase u novu kolonu i pridruživanje vrednosti 1 ili 0 (True ili False), npr.: za kolonu **pol** ćemo dobiti dve binarne kolone **Female** i **Male**.
  - **Custom Binary Encoding** - kombinacija **Label Encoding**-a i **One Hot Encoding**-a kako bi se kreirala dodatna kolona od značaja.

Zadatak 2

- Gradivo za Zadatak 2 obuhvata kompletno gradivo od početka semestra zaključno sa prošlonedeljnim predavanjem **Metod maksimalne verodostojnosti**.
- Naglašeni koncepti za Zadatak 2:
  - Višestruka linearna regresija;
  - Regularizacija (Ridge, Lasso, Elastic Net...);
  - Neparametarski pristupi (Nearest Neighbor, Kernel Regression…);
  - Rad sa kategoričkim podacima.

Zadatak 2

- Gradivo obrađeno na predavanjima je dovoljno kako bi se zadatak uspešno uradio.
- Dodatnim istraživanjem (pod)oblasti i problema moguće je ostvariti bolje rezultate.
- Ohrabruje se dodatno istraživanje i primena istraženog, uz (jedino) ograničenje da (novi, istraženi) algoritmi moraju biti algoritmi višestruke regresije. Ne postoje ograničenja što se tiče tehnika za obradu podataka i rad sa trening skupom. Pošaljite asistentu e-mail ukoliko niste sigurni da li nešto sme ili ne sme da se iskoristi za izradu zadatka.
- Svaki tim može najviše dva puta *submit*-ovati svoje rešenje. Platforma kao konačno rešenje tima uvek uzima rešenje sa boljim ostvarenim rezultatom.