Mašinsko učenje 2024

Zadatak 1

Sadržaj

- Podsetnik - praktičan deo
- Zadatak 1
- SciPy Stack
- Uputstva i saveti

Podsetnik - praktičan deo

Podsetnik - praktičan deo

- Praktičan deo predmeta nosi najviše 60* bodova.
- Sastoji se od:
  - 5 domaćih zadataka
  - Predmetnog projekta.
- Akcenat na timskom radu:
  - Svaki član tima mora dati svoj doprinos
  - Bodovi dodeljeni članovima istog tima mogu da se razlikuju.

* u posebnim slučajevima 60 bodova donosi i dodatnih 40.

Podsetnik - praktičan deo

- Opcije:
  - Samo projekat = najviše 25 bodova
  - 2 domaća zadatka (najviše 25) + projekat (najviše 25) = najviše 50 bodova
  - 4+ domaćih zadataka (najviše 35) + projekat (najviše 25) = najviše 60 bodova
  - Nagrada za najuspešnije = najviše 60 bodova.
  - Najuspešniji od najuspešnijih = 100 bodova.

Podsetnik - praktičan deo

- Kriterijumi:
  - Ostvareni rezultati i kako se do njih došlo:
    - Pristup problemima
    - Korišćeni algoritmi
    - Određivanje (hiper)parametara algoritama
    - Rad sa trening skupom podataka.
  - Propratni izveštaji:
    - Sadržaj propratnih izveštaja
    - Usklađenost izveštaja i izvornih kodova rešenja.
  - Diskusija:
    - Prezentovanje rešenja i odgovori na pitanja prilikom prezentovanja.

Podsetnik - praktičan deo

- Raspored domaćih zadataka:
  - 04.03. - 19.03. Jednostruka linearna regresija
  - 25.03. - 16.04. Višestruka regresija
  - 22.04. - 08.05. Klasifikacija
  - 13.05. - 22.05. Ansambl klasifikatora
  - 27.05. - 02.06. Klasterovanje

Zadatak 1

Zadatak 1

- Jednostruka linearna regresija:
  - Upotrebom jednostruke linearne regresije prediktovati **Y** na osnovu **X**.
  - Zadatak je uspešno urađen ukoliko se na testnom skupu podataka dobije **RMSE (Root Mean Square Error)** manji od 180.
  - Algoritmi mašinskog učenja se samostalno implementiraju - **zabranjena upotreba algoritama iz biblioteka**.
  - Rok za izradu zadatka je **19.03.2024. u 23:59h.**
  - Instalirane biblioteke za Zadatak 1:
    - **NumPy**
    - **Pandas**.
  - Sledeći termin vežbi (odbrana Zadatka 1 i predstavljanje Zadatka 2) je u nedelji **25.03. - 05.04.2024.**

Zadatak 1

- Koncepti vezani za Zadatak 1 (podsetiti se gradiva sa predavanja i ranijih predmeta):
  - *Gradient Descent* (*Batch vs Stochastic*)
  - *Normal Equation*
  - Normalizacija
  - *Outlier*-i
  - Rad sa skupom podataka
  - Pravilno računanje metrike

Zadatak 1

- Normalizacija:
  - *min-max* normalizacija:
    - *normalized_data = (data - np.min(data)) / (np.max(data) - np.min(data))*
  - *z-score* normalizacija:
    - *normalized_data = (data - np.mean(data)) / np.std(data)*
  - Koeficijenti se računaju isključivo na trening skupu podataka, a koriste se za normalizaciju i trening i testnog skupa podataka

Zadatak 1

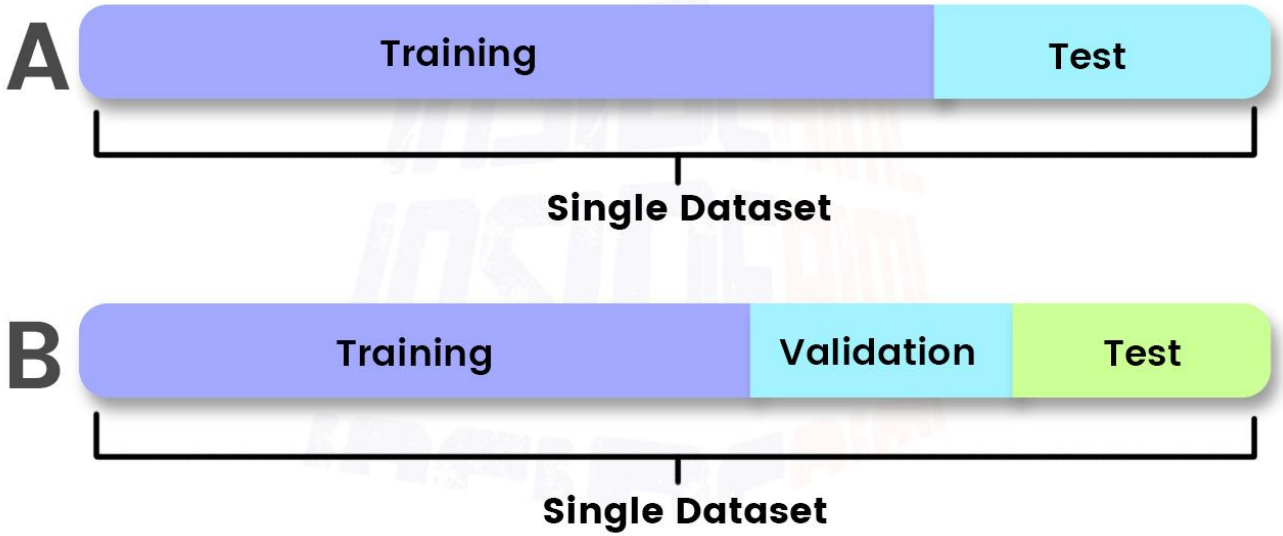
- *Outlier*-i:
  - Podaci koji se značajno razlikuju od ostalih opažanja:
    - Greške u merenju, greške u unosu podataka, neobični opaženi fenomeni,...
  - Mogu imati veliki uticaj na rezultujući model
  - Neke od strategija za rad sa *outlier*-ima:
    - Ignorisanje
      - Ako imaju minimalan uticaj na performanse modela
    - Uklanjanje
      - Ručno
      - Korišćenjem neke od statističkih metoda:
        - *z-score*: udaljenost opservacije od srednje vrednosti izražena u broju standardnih devijacija
      - Na ovom kursu je dozvoljeno uklanjanje podataka (redova i kolona) samo iz trening skupa

Zadatak 1

- *Outlier*-i:
  - Neke od strategija za rad sa *outlier*-ima:
    - Transformacija podataka
      - Npr.: logaritamska transformacija
        - `transformed_data = np.log(data)`
    - Korišćenje robusnih metoda:
      - Koristiti algoritme koji su prirodno otporni(ji) na *outlier*-e.

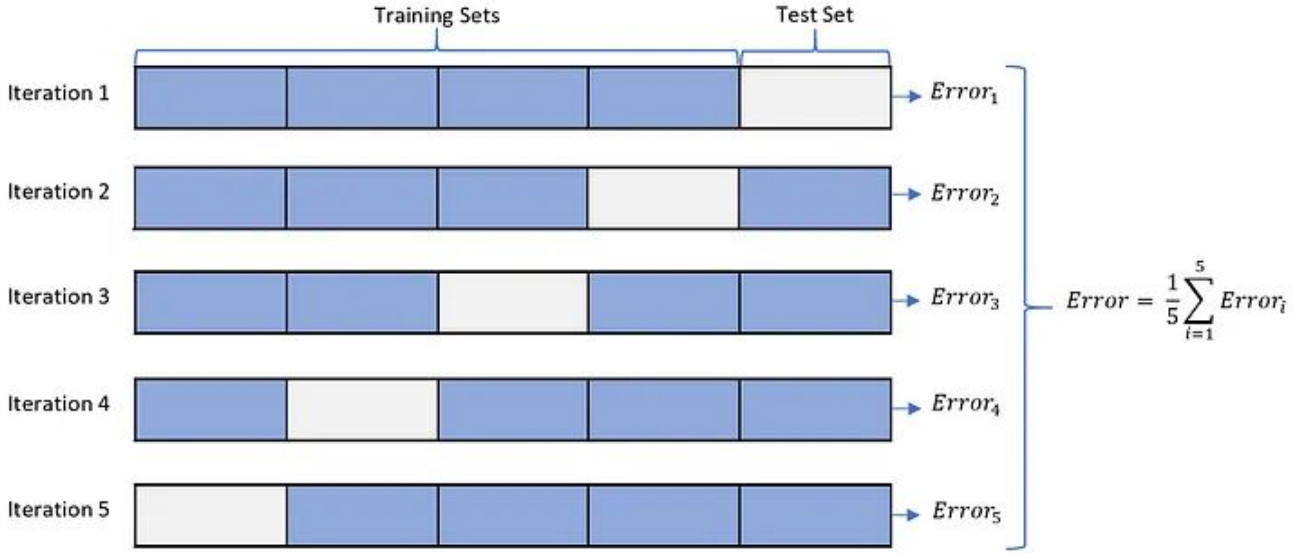
Zadatak 1

- Rad sa skupom podataka:



Zadatak 1

- Rad sa skupom podataka:



Zadatak 1

- Pravilno računanje metrike:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

Zadatak 1

- Pravilno računanje metrike:
  - Ako su podaci prethodno bili transformisani, potrebno ih je vratiti u originalan oblik pre računanja metrike
  - Na primer, ako su podaci normalizovani, pre računanja metrike ih je potrebno denormalizovati:
    - *min-max* denormalizacija:
      - `data = normalized_data * (np.max(data) - np.min(data)) + np.min(data)`
    - *z-score* denormalizacija:
      - `data = normalized_data * np.std(data) + np.mean(data)`

SciPy Stack

Zadatak 1

- Za izradu zadataka koristiti **Python 3.10.x**.
- Preporuka da se prilikom izrade zadataka oslonac bude SciPy Stack i njegove biblioteke:
  - NumPy
  - SciPy
  - Matplotlib
  - Jupyter
  - Pandas.

Zadatak 1

- Za Zadatak 1 na platformi su instalirane biblioteke (verzije date u Uputstvu):
  - **NumPy:**
    - Docs
    - Stanford Tutorial
  - **Pandas:**
    - Docs
    - Tutorial
    - 10 Minutes to Pandas.
- Za potrebe vizualizacije podataka i pisanja propratnog izveštaja od pomoći može biti biblioteka **Matplotlib**:
  - Docs
  - Tutorial.

Uputstva i saveti

Zadatak 1

- Uputstvo za rad sa platformom i pisanje propratnog izveštaja se nalazi u:
  - **Files/Vežbe/Uputstvo.pdf**.
- Saveti za rešavanje zadataka:
  - Podsetiti se gradiva sa predavanja
  - Detaljno pročitati uputstvo za rad sa platformom i pisanje propratnog izveštaja
  - Vizualizacija podataka
  - Isprobati više pristupa - podeliti zaduženja tako da svaki član tima implementira jedan pristup. Nakon toga, zajedno analizirati implementirano i odabrati najbolji pristup koji će se evaluirati na platformi.
  - Ako se radi normalizacija podataka, obratiti pažnju kako će se računati RMSE metrika.

Zadatak 1

- Savet za implementaciju:
  - Metoda **fit(x, y)** za “fitovanje” trening podataka
  - Metoda **predict(x)** za predikciju vrednosti testnog skupa
  - Metoda **calculate_rmse(y_true, y_predict)** za računanje RMSE na osnovu date formule.