Mašinsko učenje 2024

Zadatak 4

Sadržaj

- Zadatak 3 - Rekapitulacija
- Zadatak 4

Zadatak 3 - Rekapitulacija

Zadatak 3 - Rekapitulacija

- Procenat uspešnosti: **88%** (30/34).
- Najveće preklapanje izvornih kodova prema alatu za detekciju plagijata: **15%**.
- Najbolji rezultati po terminima:

Termin	Tim	Micro F1
Ponedeljak - G4	**Bivuja**	**0.766**
Utorak - G5	**Placeholder**	**0.757**
Utorak - G3	**StudentVentures**	**0.778**
Četvrtak - G2	**tim15_24**	**0.77**
Petak - G1	**Tehno trube**	**0.764**

Zadatak 3 - Rekapitulacija

- Dobre stvari (na nivou generacije):
  - Pretprocesiranje
  - Vektorizacija
  - Propratni izveštaji.

- Stvari koje mogu biti bolje (na nivou generacije):
  - Optimizacija hiperparametara modela.

Zadatak 4

Zadatak 4

- Klasifikacija:
  - Na osnovu dostupnih informacija o igricama, odrediti njihov žanr (**Genre**):
    - **Action**
    - **Adventure**
    - **Racing**
    - **Platform**.
  - Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije **makro f1 mera** (eng. *macro f1 score*) veća od 0.30.
  - Zadatak se rešava upotrebom ansambla klasifikatora.
  - Rok za izradu zadatka je **22.05.2024. u 23:59h**.

Zadatak 4

- Klasifikacija:
  - Instalirane biblioteke za Zadatak 4:
    - **NumPy**
    - **Pandas**
    - **SciPy**
    - **scikit-learn**.
  - Sledeći termin vežbi (odbrana Zadatka 4 i predstavljanje Zadatka 5) je u nedelji **27.05. - 31.05.2024.**

Zadatak 4

- Atributi:
  - **Gaming_Platform** - platforma za igranje:
    - **PC**
    - **N64**
    - **PS2**
    - **PS4**
    - **…**
  - **YoR** - godina izdavanja igrice
  - **Sales_NA** - količina prodatih igrica u Americi (u milionima)
  - **Sales_EU** - količina prodatih igrica u Evropi (u milionima)
  - **Sales_JP** - količina prodatih igrica u Japanu (u milionima)
  - **Other_Sales** - količina prodatih igrica u ostatku sveta (u milionima).

Zadatak 4

- Koncepti vezani za Zadatak 4:
  - Nedostajuće vrednosti
  - Redukcija dimenzionalnosti
  - Ansambli klasifikatora
  - Metrika

Zadatak 4

- Nedostajuće vrednosti:
  - Trening skup podataka sadrži nedostajuće vrednosti (u pitanju su prazne ćelije).
  - Testni skup podataka **ne** sadrži nedostajuće vrednosti.
  - Rad sa nedostajućim vrednostima:
    - Uklanjanje torki koje sadrže nedostajuće vrednosti
    - Zamena nedostajućih vrednosti nekom statistikom
    - Popunjavanje nedostajućih vrednosti na osnovu najbližih suseda
    - …
  - Više o popunjavanju nedostajućih vrednosti u scikit-learn biblioteci možete pročitati ovde.

Zadatak 4

- Nedostajuće vrednosti:
  - Popunjavanje nedostajućih vrednosti uz pomoć scikit-learn:
    - Klase
    - Upotreba:

```
imputer = SimpleImputer()
X_train = imputer.fit_transform(train)
```

ili

```
imputer = SimpleImputer()
imputer.fit(train)
X_train = imputer.transform(train)
```

Zadatak 4

- Redukcija dimenzionalnosti:
  - Kreirati novi podskup obeležja koji dobro sumarizuje polazna obeležja
  - Dobar skup obeležja je onaj koji je relevantan za ciljnu funkciju
  - *Principal Component Analysis (PCA)*:
    i. Konstruisati mali broj linearnih obeležja koji sumarizuju ulazne podatke
    ii. Zadržati što više informacija u podacima.

Zadatak 4

- Redukcija dimenzionalnosti:
  - *PCA* se može implementirati samostalno, a može se iskoristiti i implementacija iz scikit-learn biblioteke:
    - *PCA*
    - *KernelPCA*

```
pca = PCA()
X_train = pca.fit_transform(train)
X_test = pca.transform(test)
```

ili

```
pca = PCA()
pca.fit(train)
X_train = pca.transform(train)
X_test = pca.transform(test)
```

Zadatak 4

- Ansambli klasifikatora:
  - Zadatak se **mora** rešiti upotrebom neke od metoda ansambla:
    - *Bagging*
    - *Boosting*
    - *Stacking*
    - *Voting*.

  - Metode ansambla u scikit-learn.
  - Napomena: *Random Forest* se može koristiti za izradu zadatka, ali se mora znati objasniti kako taj model radi.

Zadatak 4

- Metrika:
  - Kao meru performansi modela u ovom zadatku imamo makro f1 meru (eng. *macro f1 score*).
  - **macro f1 score** - računa metrike za svaku labelu i pronalazi njihovu neponderisanu srednju vrednost:
    - *sklearn.metrics.f1_score(y_true, y_pred, average='macro')*

  - Prilikom treninga, od pomoći može biti i *classification_report*.

Zadatak 4

- Saveti za rešavanje zadatka:
  - Podsetiti se gradiva sa predavanja
  - Uraditi eksplorativnu analizu podataka
  - Isprobati više tehnika za rad sa nedostajućim vrednostima
  - Isprobati redukciju dimenzionalnosti
  - Isprobati više modela i analizirati njihovo ponašanje po klasama.