Mašinsko učenje 2020

Zadatak 5

Sadržaj

Zadatak 4 - Rekapitulacija

Zadatak 4 - Rekapitulacija

- Procenat uspešnosti: **63.64%** (21/33).
- Sva rešenja sa **micro f1 score ≥ 0.55** se smatraju odličnim.
- Najveće preklapanje izvornih kodova prema alatu za detekciju plagijata: **8%.**

- Najbolji rezultati po terminima:

Termin	Tim	micro f1
Ponedeljak	**tim6_20**	**0.58**
Utorak 1	**tim8_20**	**0.555**
Utorak 2	**tim15_20**	**0.546**
Četvrtak	**tim7_20**	**0.568**
Petak	**tim22_20**	**0.569**

Zadatak 4 - Rekapitulacija

- Dobre stvari (na nivou generacije):
  - Rad sa nedostajućim vrednostima;
  - Određivanje hiperparametara modela;
  - Propratni izveštaji.

Zadatak 5

Zadatak 5

- Klasterovanje:
  - Klasterovati države na osnovu njihovih karakteristika u klastere koji predstavljaju geografske regione (**region**):
    - **Africa**
    - **Americas**
    - **Asia**
    - **Europe**

Zadatak 5

- Klasterovanje:
  - Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije v mera (eng. *v measure score*) > 0.40.
  - Zadatak se rešava upotrebom Modela Gausovih mešavina (eng. *Gaussian Mixture Model*, *GMM*), tj. algoritmom Očekivanje - maksimizacija (eng. *Expectation-maximization*, *EM*).
  - Rok **31.05.2020. u 12:59h.**
  - Trening skup podataka sadrži nedostajuće vrednosti (prazne ćelije).
  - Instalirane biblioteke za Zadatak 5 (verzije date u Uputstvu):
    - **NumPy**
    - **SciPy**
    - **Pandas**
    - **scikit-learn**.

Zadatak 5

- Atributi:
  - **income** - prihod po glavi stanovnika u $
  - **infant** - smrtnost odojčadi na 1000 živorođenih
  - **oil** - da li je država izvoznik nafte:
    - **yes** - da
    - **no** - ne.

Zadatak 5

- Trening skup podataka sadrži nedostajuće vrednosti (u pitanju su prazne ćelije).
- Testni skup podataka **ne** sadrži nedostajuće vrednosti.

- Zadatak se **mora** rešiti upotrebom GMM, tj. EM algoritma.
- Algoritam možete implementirati samostalno, a možete iskoristiti i implementaciju scikit-learn biblioteke.

Zadatak 5

- Kod ovog zadatka, evaluaciju klasterovanja ćemo zasnivati na poznavanju *ground truth* labela klastera.
- Koristićemo **v meru** (eng. *v measure score*), koja se zasniva na intuitivnim metrikama zasnovanim na uslovnoj analizi entropije:
  - **homogenost** (eng. *homogeneity*) - svaki klaster sadrži članove samo jedne grupe/klastera;
  - **potpunost** (eng. *completeness*) - svi članovi iste grupe/klastera su dodeljeni istom klasteru.
- **v mera** predstavlja harmonijsku sredinu homogenosti i potpunosti:
  - *sklearn.metrics.v_measure_score(labels_true, labels_pred)*