Mašinsko učenje 2018

Zadatak 6

Sadržaj

Zadatak 5 - Rekapitulacija

Zadatak 5 - Rekapitulacija

- Procenat uspešnosti : **65.5%** (19/29)
- Najbolji rezultati po terminima:
  - Ponedeljak:
    - **code10** (0.62)
  - Utorak 1:
    - **tim6** (0.646)
  - Utorak 2:
    - **nemanja_brzak, ducklifors** (0.626)
  - Sreda:
    - **tim8** (0.66)
- Maksimalno preklapanje po MOSS-u: **16%**.

Zadatak 6

Zadatak 6

- PCA + Klasifikacija:
  - Na osnovu dostupnih informacija o zaposlenima na istočnoj obali SAD, izvršiti predikciju njihove rase (**race**):
    - 1. White
    - 2. Black
    - 3. Asian
    - 4. Other

Zadatak 6

- PCA + Klasifikacija:
  - Zadatak je uspešno urađen ukoliko se na kompletnom testnom skupu podataka dobije mikro f1 mera (eng. *micro f1 score*) > 0.80.
  - **Obavezna upotreba PCA!**
  - Dozvoljena upotreba svih dostupnih klasfikatora/ansambla klasifikatora.
  - Rok: **20.05.2018. u 19:59h.**
  - Dostupne biblioteke: **NumPy, SciPy, Pandas** i **scikit-learn**.
  - Trening skup podataka sadrži nedostajuće vrednosti (prazne ćelije).

Zadatak 6

- Atributi:
  - **year** - godina kada su prikupljane informacije
  - **age** - broj godina (starost) zaposlenih
  - **maritl** - bračni status zaposlenih:
    - **1. Never Married** - nikad venčani
    - **2. Married** - venčani
    - **3. Widowed** - udovice/udovci
    - **4. Divorced** - razvedeni
    - **5. Separated** - rastavljeni

Zadatak 6

- Atributi:
  - **education** - nivo obrazovanja:
    - **1. < HS Grad** - nezavršena srednja škola
    - **2. HS Grad** - završena srednja škola
    - **3. Some College** - nezavršen fakultet
    - **4. College Grad** - završen fakultet
    - **5. Advanced Degree** - MSc, PhD
  - **jobclass** - tip posla:
    - **1. Industrial**
    - **2. Information**

Zadatak 6

- Atributi:
  - **health** - zdravstveno stanje:
    - **1. <=Good**
    - **2. >= Very Good**
  - **health_ins** - da li zaposleni poseduje zdravstveno osiguranje:
    - **1. Yes** - da
    - **2. No** - ne
  - **wage** - godišnja plata (u hiljadama $)

Zadatak 6

- PCA konstruiše mali broj linearnih obeležja koja sumarizuju ulazne podatke (obeležja dobijena pomoću PCA su linearne kombinacije originalnih obeležja).
- PCA ima cilj da zadrži što više informacija u podacima:
  - Identifikuje dominantne dimenzije
  - Odbaci manje dimenzije (šum).
- Svi koraci primenjeni u PCA su nenadgledani.
- Kernel PCA - PCA sa nelinearnim transformacijama obeležja (analogija sa kernelom kod SVM).

Zadatak 6

- scikit-learn:
  - PCA
  - KernelPCA

Zadatak 6

- Korisno:
  - Towards Data Science: PCA for Data Visualization & PCA to Speed-up Machine Learning Algorithms
  - Kaggle: PCA in Glass Classification