Mašinsko učenje 2018

Zadatak 2

Sadržaj

Zadatak 1 - Rekapitulacija

Zadatak 1 - Rekapitulacija

- Procenat uspešnosti: **83%** (24/29)
- Najbolji rezultat po terminima:
  - Ponedeljak:
    - **code10** (RMSE:68.91)
  - Utorak 1:
    - **tim1** (RMSE: 68.57)
  - Utorak 2:
    - **tartufi** (RMSE:68.51)
  - Sreda:
    - **tim9** (RMSE: 68.15)
- Maksimalno preklapanje po MOSS-u: **10%**

Zadatak 2

Zadatak 2

- Višestruka regresija:
  - Prediktovati platu (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 29000.
  - Rok: **11.04.2018. u 23:59h** (sledeće vežbe u nedelji 16.04.2018.).
  - Dostupne biblioteke: **Numpy, Pandas i SciPy.**

Zadatak 2

- Atributi:
  - **rank** - zvanje:
    - **Prof** - redovni profesor
    - **AssocProf** - vanredni profesor
    - **AsstProf** - docent
  - **discipline** - disciplina istraživanja:
    - **A** - theoretical
    - **B** - applied
  - **yrs.since.phd** - godina proteklo od doktoriranja
  - **yrs.service** - godina “radnog staža”
  - **sex** - pol:
    - **Female** - žensko
    - **Male** - muško.

Zadatak 2

- Kategorički podaci:
  - Label Encoding - konvertovanje kategoričkih podataka u broj iz opsega [0, num_class-1].
  - One Hot Encoding - konvertovanje svake klase u novu kolonu i pridruživanje vrednosti 1 ili 0 (True ili False)
  - Custom Binary Encoding - kombinacija Label Encoding-a i One Hot Encoding-a kako bi se kreirala dodatna kolona od značaja.

Višestruka regresija

- Višestruka linearna regresija y=h(x1, x2, … xd).
- Regularizacija:
  - Ridge Regression (L2):
    - GD
    - Closed form
  - Lasso Regression (L1):
    - GD
    - Coordinate Descent
  - Elastic Net:
    - Linearna kombinacija L1 i L2.
- Neparametarski pristup:
  - Nearest Neighbor
  - Kernel Regression.