ZIMSKA ŠOLA 2012/2013

Predmet Odkrivanje znanja v podatkih, Podatkovno rudarjenje

13.2.2013

Sestavil: doc. dr. Matej Mertik

NAVODILA ZA IZDELAVO SEMINARSKE NALOGE

1. Preberi zbirko iz ustreznega formata (UCI Machine learning repository)
2. Počisti manjkajoče vrednosti
3. Poišči morebitne ubežnike (outlayer)
4. Preglej podatke z različnimi vizualizacijskimi metodami
5. Uporabi dve primerni metodi za izgradnjo modela in primerjaj rezultate med njima (KNN, SVM, Naive Bayes, Odločitveno drevo)
6. Izgradi model nad podatki in ga testiraj s pomočjo operatorja CrossValidation
7. Predstavi rezultate in hipoteze predikcijskega modela nad podatki

V nalogi uporabi orodje RapidMiner 5.3. V nalogi podaj sheme procesov, ki si jih ustvaril v orodju in jih komentiraj. Za vsak operator gradnje modela (KNN, SVM, Naive Bayes, Odločitveno drevo) podaj opis vseh nastavitev, ki si jih uporabil med delom in predstavi njegove optimalne vrednosti s katerimi si prišel do rezultatov.

Naloga naj vsebuje naslednje poglavja:

- Opis problema in podatkovne zbirke
- Opis procesa priprave podatkovne zbirke (priprava in vizualizacija, čiščenje, detekcija outlayerjev)
- Proces rudarjenja (opis nastavitev za metode, optimalna nastavitev in metoda)
- Predstavitev rezultatov modela

Pri pripravi podatkov se lahko glede na rezultate odločiš za zajemanje in filtriranje le dela podatkovne zbirke, v tem primeru utemelji s katerimi razmisleki si prišel do tega dela zbirke. Uporabno v primeru, kjer ni mogoče doseči napovedi modela nad 70%.

*Nalogo oddaj na učilico v PDF formatu skladno s pravili izdelave seminarskih nalog na fakulteti. Za prvi rok izpita je potrebno naloge oddati do **27.2.2013**, zagovor nalog za prvi rok bo predvidoma **28.2.2013**. O uri boste obveščeni naknadno. Naloge, ki jih boste oddali po prvem roku bodo ocenjene sproti, datum zagovorov drugega in tretjega roka bosta objavljena v letnem semestru predvidoma maja in septembra.*

***

UCI Repository: http://archive.ics.uci.edu/ml/

Podatkovne zbirke, ki pridejo v poštev:

- Yeast, Forest Fire, Wine Quality, One-hundred plant species leaves, Car Evaluation Data Set, Human Activity Recognition Using Smartphones Data Set, Internet Advertisements Data Set, Adult Data Set, Seeds, Letter Recognition, Mechanical Analysis Data Set, Poker Hand Data Set, MAGIC Gamma Telescope Data Set, Balance scale
- Mogoče je izbrati tudi lastno zbirko, vendar naj obsega preko 300 vzorcev z vsaj 12 atributi (na primer analiza vaših finančnih podatkov, telefonskih pogovorov...)