Adatbányászat 2018

Gyakorlati követelmények

A gyakorlatot kétféle módon lehet teljesíteni: a) egy projekt megvalósítása RapidMinerben, b) egy nem standard adatbányászati algoritmus leprogramozása (pl. egy új node készítése RapidMinerben).

A RapidMiner (RM) projekt tartalmazza az alábbi elemeket:

- Egy adatbeolvasási rész, ahol egy külső adatforrásból (lehet egy az RM-nél eltérő típusú fájl, akár egy egy URL címen) származnak az adatok.
- Feltáró adatelemzés: az alapadatállomány vizsgálata.
- Metadatok módosítása beállítása: a felügyelt változó kijelölése, mérési skálák megváltoztatása stb.
- Egy adat előfeldolgozó, módosító rész: esetleges hiányzó adatok pótlása, változók transzformációja stb.
- Több felügyelt vagy nem nem-felügyelt operátor alkalmazása.
- Az alkalmazott operátorok kalibrálása, azaz egyes beállítható paraméterek közül az optimális megkeresése pl. loop alkalmazásával, a RM programozási elemeinek használata
- A kapott modellek vizsgálata megfelelő eszközökkel, a legjobb modell megtalálása.

Adatforrások:

- https://github.com/caesar0301/awesome-public-datasets
- https://grouplens.org/datasets/
- http://snap.stanford.edu/data/
- http://mobblog.cs.ucl.ac.uk/datasets/
- http://www.face-rec.org/databases/

Kérem, hogy a standard UCI Machine Learning adatállományokat ne használják.

Debrecen, 2018.04.24.

Dr. Ispány Márton