Adatbányászati technikák 2. házi feladat

Általános szabályok

- Az első feladatot a Weka Explorerben oldjad meg. A lépésekről készíts képernyőképeket, és illeszd a dokumentumba. A képeket úgy méretezd, hogy a lényegi rész látható legyen rajta. Az algoritmusok beállításáról mindig készíts képet.
- A második és harmadik feladatban Python kódot kell írnod.
- A szöveges megoldásokat a képekkel pdf formátumban küldd el.
- A magyarázatokat olyan részletességgel add meg, hogy egy, az adatbányászattal alapszinten tisztában levő ember megértse.
- A második és harmadik feladathoz írt programnál a .py file-okat vagy Jupyter notebookot küldd el.
- A kész feladatokat a `kabodil@cs.bme.hu` e-mail címre küldd. A levél tárgya kezdődjön "[adatbányászati technikák]"-kal. A levélbe mindenképp írd bele a neved és neptun kódod. A megoldásaidat egy tömörített (pl. zip) file-ként küldd.
- A beadási határidő a bemutatás előtti vasárnap éjfél, azaz páratlan héten járóknak május 6-a, páros héten járóknak pedig május 13-a.
- Az utolsó gyakorlaton a házit meg kell védeni. Ez a megoldás értelmezését, és annak kisebb módosítását jelenti. Mindegyik feladatnál előfordulhatnak a témához kapcsolódó más kérdések is.
- A feladatok értelmezésével kapcsolatban szintén a fenti e-mail címen kérdezhetsz.

1. feladat A feladatban a Weka programcsomag Explorer felhasználói felületén kell osztályozással kapcsolatos feladatokat elvégezni. A használt adathalmaz a `http://cs.bme.hu/~kabodil/adatbanyaszat/hf/weka/diabetes.arff` címen érhető el.

1. Töltsd be az adathalmazt a Weka Explorerbe! Melyik attribútum fölösleges? Miért? Hagyd el az adathalmazból! (0,5 pont)
2. Az adathalmaz cukorbetegségről szól. Mi a nagyobb hiba, egészséges embert cukorbetegnek nyilvánítani, vagy cukorbeteg embert egészségesnek? Miért? (0,5 pont)
3. Készíts egy J48 fát. Állítsd át a levelenkénti minimális elemek számát 10-re. Osztálycímkének az utolsó attribútumot használd. Nézd meg a készült fát (és készíts róla képet) (1 pont)
4. Készíts mesterséges neurális hálót, ami két rejtett réteget tartalmaz, 50-50 neuronnal. A tanítás 400 iterációig tartson. Hasonlítsd össze az így kapott eredményt az előbbi döntési fával. (1 pont)

5. Osztályozd az adatokat Adaboosttal. Az Adaboost által használt osztályozó Decision Stump legyen. Az iterációk száma legyen 30. Hasonlítsd össze ezt az eredményt is a korábbiakkal. (1 pont)

6. Diszkretizáld a „mass" attribútumot öt kategóriára, amik egyenlő szélességűek, és a „skin" attribútumot szintén öt kategóriára, de ezekben a példányszámok legyenek egyenlőek. Készíts erről is J48 fát alapbeállítások mellett, és hasonlítsd össze az előző osztályozókkal. (1 pont)

2. feladat A feladatban egy map és egy reduce föggvényt kell megírni az órán is használt *mapreduce.py*-hoz. Az adatok és a python file-ok a `http://cs.bme.hu/~kabodil/adatbanyaszat/hf/mapreduce/mapreduce.zip` címről tölthetők le.

(Az adatok forrása: `http://archivio-meteo.distile.it/tabelle-dati-archivio-meteo/`.)

Írj egy *map* és egy *reduce* függvényt, ami a *data_2010.txt*, ..., *data_2015.txt* fileokban tárolt dátum - hőmérséklet párokból hónap - átlaghőmérséklet párokat számol. Vagyis a feladat a sokéves átlag havi hőmérsékletek meghatározása. (Eredményül 12 kulcs-érték pár kell, hónaponként egy.) Csak a my_map.py és a my_reduce.py file-okba írj. (5 pont)

3. feladat A feladatban az órai sklearn-t használó programot kell kiegészíteni egy saját osztályozóval. A bővítendő forrás és a felhasznált adatok a `http://cs.bme.hu/~kabodil/adatbanyaszat/hf/sklearn/sklearn.zip` címről tölthetők le.

Készíts egy olyan osztályozót, ami az sklearn-ben található beépített osztályozókat használja, és azok eredményeit kombinálja egy saját jóslattá. Az eredmények kombinálása lehet egyszerű többségi szavazás alapján, vagy akár tetszőleges súlyozással. A saját osztályozód legalább három beépített osztályozó eredményeit használja fel. Az osztályozót írja ki a kapott eredmények értékelését (sklearn.metrics.accuracy_score és sklearn.metrics.classification_report). (5 pont)