Adatbányászati technikák 2. házi feladat

Általános szabályok

- Az első két feladatot a Weka Explorerben kell megoldani. Itt a lépésekről képernyőképeket kell készítenetek, és ezt beilleszteni a dokumentumba. A képeket úgy méretezzétek, hogy a lényegi rész látható legyen rajta. Az algoritmusok beállításáról mindig készíts képet.
- A harmadik feladatban Java kódot kell írni.
- A szöveges megoldásokat a képekkel pdf formátumban küldjétek el.
- A magyarázatokat olyan részletességgel kell megadni, hogy egy, az adatbányászattal alapszinten tisztában levő ember megértse.
- A harmadik feladathoz írt programnál csak a forráskódot (.java file-ok) kell elküldeni.
- A kész feladatokat a `kabodil@cs.bme.hu` e-mail címre küldjétek. A levél tárgya kezdődjön "[adatbányászati technikák]"-kal. A levélbe mindenképp írjátok bele a nevetek és neptun kódotok. Lehetőleg egy tömörített (pl. zip) file-ként.
- A beadási határidő a bemutatás előtti vasárnap éjfél, azaz páratlan héten járóknak május 8-a, páros héten járóknak pedig május 15-e.
- Az utolsó gyakorlaton a házit meg kell védeni. Az első két feladatnál ez a küldött eredmény reprodukcióját, a harmadiknál a forrás értelmezését és kisebb módosítását jelenti. Mindegyik feladatnál előfordulhatnak a témához kapcsolódó más kérdések is.
- A feladatok értelmezésével kapcsolatban szintén a fenti e-mail címen kérdezhettek.

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

1. Töltsd be az adathalmazt a Weka Explorerbe! Melyik attribútum fölösleges? Miért? Hagyd el az adathalmazból! (1 pont)
2. Az adathalmaz cukorbetegségről szól. Mi a nagyobb hiba, egészséges embert cukorbetegnek nyilvánítani, vagy cukorbeteg embert egészségesnek? Miért? (0,5 pont)
3. Az attribútumokat végignézve, ha csak egyre végezhetsz egy bináris döntést, melyik atribútumra milyen döntést végeznél, és miért? (Képpel támaszd alá.) (1 pont)
4. Készíts J48 fát alapbeállítások mellett, az utolsó attribútumot osztálycímkének használva. Nézd meg a készült fát (készíts róla képet is). (1 pont)

5. Osztályozd a fa alapján a következő rekordot:

   ```
   769,2,110,80,30,100,30,0.4,35
   ```

   A rekord tartalmazza az első részfeladatban kitörölt attribútumot is, viszont az osztálycímkét nem. Az attribútumok sorrendje megegyezik az adathalmazban levővel. Melyik osztályba sorolná ezt az általad kapott fa? (0,5 pont)

6. Készíts egy másik J48-as fát is. Állítsd át a levelenkénti minimális elemek számát 5-re, és válassz egy tetszőleges véletlen seedet. Melyik fa lett jobb, és miért? (1 pont)

7. 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, tetszőleges seeddel, és hasonlítsd össze az előző kettővel. (1 pont)

2. feladat A feladatban a Weka programcsomag Explorer felhasználói felületén kell klaszterezési feladatokat elvégezni. A használt adathalmaz a `http://cs.bme.hu/~kabodil/2sp2glob.arff` címen érhető el.

1. Töltsd be az adathalmazt a Weka Explorerbe. Nézd meg koordinátarendszerben ábrázolva az adatokat. (Készíts róla képet is.) Hány klaszter található rajta? (1 pont)

2. Futtasd az egyszerű K-Means algoritmust annyi klaszterre, amennyit gondolsz, hogy van az adathalmazban. Majd futtasd annyira, amennyit valószínűleg meg is talál. Mi a különbség oka? (A klaszterezések eredményéről készíts képet.) (2 pont)

3. Futtasd a DBSCAN algoritmust az adathalmazon. Az ε-t változtatva melyik a legjobb általad elért klaszterezés? Szerinted van-e jobb? (Ne csak a legjobbról készíts képet, hanem arról is, hogy miért gondolod azt a legjobbnak.) (1 pont)

3. feladat A feladatban Java nyelven a Weka API-t felhasználva kell neurális hálóval kapcsolatos feladatokat megoldani. A használt adathalmaz a `http://cs.bme.hu/~kabodil/ionosphere.arff` címen érhető el.

1. Olvasd be az adathalmazt Weka-val használható formátumba, és állítsd be az osztályváltozót az utolsó attribútumra. (0,5 pont)

2. Hozz létre egy multilayer perceptront az osztályozási feladat megoldására. A neurális háló tulajdonságai legyenek a következők:

   - Két rejtett réteg, 50-50 neuronnal.
   - Legalább 10 iterációs lépés. (Ha a géped elég erős, és van rá időd, inkább több.)

   Használj 10-szeres keresztvalidációt! Írasd ki a confusion mátrixot, valamint az F-Measure-t és az AUC-ot (ROC alatti terület) mindkét osztályra. (1,5 pont)

3. Hozz létre egy J48 fát az osztályozási feladat megoldására. A fa tulajdonságai legyenek a következők:

   - Levelenként minimum 5 elem.
   - Ne legyen pruning.

Használj 10-szeres keresztvalidációt! Írasd ki itt is azokat a statisztikákat, amiket a neurális hálónál. (1 pont)

4. A két osztályozó közül melyik adott jobb eredményt? Mi alapján gondolod ezt? (0,5 pont)

5. Töröld ki az osztályváltozót egy szűrővel, és az így keletkezett új adathalmazon futtasd az egyszerű K-Means klaszterezőt. A k legyen az eredeti adathalmazban levő osztályok száma. Az eredeti adathalmazon értékeld ki az így kapott klasztereket. Hogy viszonyul ez az eredmény a két osztályozó által kapott eredményhez? (1,5 pont)