Adatbányászati technikák
2. Házi feladat

Általános tanácsok a feladatokhoz:

- Először mindig olvassa végig a teljes feladatot, mielőtt nekilátna a megoldásnak!
- Az első két feladatot Weka Explorerben kell megoldania!
- Az első két feladat esetében az alpontoknál található kérdéseket kell megválaszolni, illetve helyenként képernyőképpel alátámasztani a feladatok megoldását.
- Ügyeljen arra, hogy a képernyőképek esetén a méret és a színek beállítása jól látható képet eredményezzen! **A túl kicsi, olvashatatlan képeket nem tudjuk értékelni**!
- Ahol magyarázatot kell adnia valamilyen jelenségre, ott a magyarázatot olyan részletességgel adja meg, hogy egy – az adatbányászat alapfogalmaival tisztában levő, de sem a konkrét adathalmazt sem az alkalmazott szoftvereket nem ismerő – laikus számára is könnyen érthető legyen.
- A harmadik feladatban Java kódot kell írnia, illetve egy rövid magyarázatot adnia a d) feladatban. A Java kódoláshoz az Eclipse IDE használata ajánlott, de nem kötelező.
- A feladatok szöveges megoldását (1-2 illetve 3/d feladatok) **pdf** formátumban kell beadni! A megoldások legyenek igényesen formázva!
- A harmadik feladat esetében a forrásfájlokat (.java kiterjesztésű fájlok) kell elküldeni! Amennyiben több forrásfájlt hoz létre, azokat közös mappába rakja! Ha a megoldását több különböző projektben készíti el, akkor ezek külön mappákba kerüljenek!
- A feladatok beadására e-mailben van lehetőség. A pdf dokumentumot valamint a forrásfájl(ok)at tartalmazó mappá(ka)t csomagolja egy zip állományba (**név_neptun.zip**), majd küldje el e-mailben a bagyibence@gmail.com címre, „**Adatbányászat HF név**” tárggyal (idézőjelek nélkül, a név helyére értelemszerűen a saját nevét kell írnia)!
- A beadási határidő azoknak, akik hivatalosan páratlan héten járnak gyakorlatra: **május 3. 11:59**
- A beadási határidő azoknak, akik hivatalosan páros héten járnak gyakorlatra: **május 10. 11:59**
- A feladatokat az utolsó laboralkalmon (13. illetve 14. hét) meg kell védeni. Ez az 1-2 feladatok esetében azt jelenti, hogy az adott magyarázatokat ki kell tudni fejteni, illetve kérésre az eredményeket reprodukálni kell tudni. A 3. feladat esetében a kód bizonyos részeinek megmagyarázására, illetve minimális módosítására kell készülni.
- A feladatok megoldásában az interneten található anyagok (Weka API, video tutorialok) rengeteg segítséget nyújthatnak.
- A feladat értelmezésével, és az esetleg felmerülő problémákkal kapcsolatban szintén a fenti e-mail címen tudnak segítséget kérni.

1. Feladat (osztályozás, 5 pont)

a) Töltse le az alábbi címen elérhető file-t! Ismerkedjen meg az adatokkal a file tetszőleges text editorban való megvizsgálásával! Mit jelentenek a célváltozó egyes értékei? (0,5 pont)

http://repository.seasr.org/Datasets/UCI/arff/mushroom.arff

b) Töltse be az adatokat a Weka Explorerbe! Melyik attribútum tartalmaz hiányzó értékeket? Az adatok hány százaléka hiányzik ennél az attribútumnál? Hagyjuk el ezt az attribútumot! (0,5 pont)

c) Mit gondol, a „stalk-shape”, vagy az „odor” attribútum segítségével lehet jobban szétválasztani az eseteket? ***Adjon a válaszára vizuális bizonyítékot***, és ***magyarázza meg*** részletesen, hogy miből látható a válasza! (1 pont)

d) Alkalmazzon Naive Bayes osztályozót az adathalmazra! A tesztelési opciók közül azt válassza, amelyik a legmegbízhatóbb eredményt biztosítja! Írja le melyiket választotta, és röviden ***indokolja meg miért***! Végezze el az osztályozást, és adja meg a következő adatokat a teljes adathalmazra (1 pont):

- accuracy
- precision
- recall

e) Egyforma súlyú hiba-e a helytelen osztályozás mindkét lehetséges formája? ***Válaszát indokolja***! Ha nem, melyik a nagyobb probléma? Mennyi a hibás osztályozás költsége, ha a nagyobb hibát ötször akkorának tekintjük, mint a kisebbet (a kisebb hiba legyen egységnyi)? (1 pont)

f) Készítsen döntési fát, amely tökéletesen osztályozza az eseteket! A fában csak bináris elágazások lehetnek, és minden levélelemnek legalább 5 esetet kell tartalmaznia! A döntési fát elkészítő algoritmust és a többi paramétert szabadon megválaszthatja. ***Dokumentálja, hogy milyen osztályozót és paramétereket használt! A döntési fáról mellékeljen jól látható képet!*** (1 pont)

2. Feladat (regresszió, 5 pont)

a) Töltse le az alábbi URL-ről az adathalmazt! Az adatbázis egy amerikai nemzeti park területén kialakult erdőtüzekről tartalmaz információt! A célváltozó az „area” attribútum, amely a leégett terület nagyságát mutatja meg hektárban.

http://archive.ics.uci.edu/ml/machine-learning-databases/forest-fires/forestfires.csv

b) Csak a nap, hőmérséklet, szél, eső, relatív páratartalom (RH) magyarázó változókat tartsa meg! Hozzon létre egy azonosítót az egyes esetekhez! Hogyan tudja ezt megtenni? (0,5 pont)

c) A numerikus magyarázó változókat normalizálja 0 és 1 közé, de a célváltozót ne! Hogyan tudja ezt megtenni? (0,5 pont)

d) Készítsen lineáris regressziós modellt, amellyel előre tudja jelezni, hogy mekkora lesz a leégett terület! ***A teljes adathalmazt tanító adatnak állítsa be!*** Vizsgálja meg a modellt! Milyen problémát tapasztal? Hogyan oldaná meg a problémát a lehető legegyszerűbben? (1 pont)

e) Töltse be újra a letöltött fájlt, és ismét csak b) pontban megadott magyarázó változókat tartsa meg! ***Ezúttal ne hozzon létre azonosítót és ne is normalizáljon!*** Készítsen ismét egy lineáris regressziót az adatokra, de ezúttal alkalmazzon kereszt-validálást! ***Magyarázza meg részletesen az így kapott modell jelentését!*** (2 pont)

f) Készítsen egy újabb regressziós modellt, amelyben ***a célváltozó a relatív páratartalom***! Ezúttal az esetek 95%-át használja fel tanítási céllal és a maradék 5%-ot használja tesztelésre! Figyeljen arra, hogy a százalékos vágás esetében az esetek sorrendje megmaradjon! Hogyan lehet ezt beállítani? A tesztelésre használt esetek közül hányas sorszámúnál tapasztalható a legnagyobb eltérés a valós és az előre jelzett érték között? ***Mutassa be képernyőkép segítségével, hogy ezt hogyan határozta meg!*** (1 pont)

3. Feladat (Weka API használat, 5 pont)

a) Töltse le a következő URL címen található adatbázist egy tetszőleges mappába! Amennyiben szükségesnek tartja, olvassa el az adatbázisról szóló információkat a file elején.

http://repository.seasr.org/Datasets/UCI/arff/zoo.arff

b) ***A következő feladatok mindegyikében Java kódot kell írnia!*** Olvassa be a későbbi Weka használathoz megfelelő formában az adatbázist, és állítsa be a célváltozót (az eredeti adatbázis „type" attribútuma). (0,5 pont)

c) Hozzon létre egy J48-as döntési fát, és állítsa be a következőket:

- A döntési fa legyen bináris!
- A levélelemek legalább 5 esetet tartalmazzanak!
- Kapcsolja be a „Reduced Error Pruning" funkciót!

Értékelje ki a létrehozott modellt 10-fold kereszt-validálással! ***Írassa ki a konzolra a modell teljesítményét összegző Stringet!*** (1,5 pont)

c) ***Írassa ki a konzolra a Confusion mátrixot*** úgy, hogy (a Weka explorerhez hasonló módon) az oszlopok és sorok fejlécében láthatók legyenek a célváltozó értékei! (1 pont)

d) A fenti adathalmazból szűrő segítségével hozzon létre egy új adathalmazt, amely nem tartalmazza a célváltozót! Ezt követően építsen SimpleKMeans klaszterező algoritmust az újonnan létrehozott adatbázisra (a k-t annyira állítsa, ahány lehetséges értéke a célváltozónak volt az eredeti adatbázisban). ***Az eredeti adatbázis segítségével értékelje ki***, hogy mennyire hatékonyan találta meg az algoritmus az eredeti osztályokat! Jobb, vagy rosszabb eredményt kaptunk mint a J48 fa esetében? Ez miből adódhat? (1 pont)

e) Törölje ki az adatbázisból az összes numerikus attribútumot! ***Figyelem, a célváltozó bár számokkal van kódolva, mégsem numerikus típusú!*** Hozzon létre asszociációs szabályokat az Apriori módszer segítségével! A minimális support alsó határát 0,3-ra, a minimális confidence értéket pedig 0,9-re állítsa! ***Írassa ki a konzolra az első 5 asszociációs szabályt!*** (1 pont)