Adatelemzés és adatbányászati módszerek

Tárgyfelelős: dr. Kovács László
Szak: Mérnök informatikus mesterszak
Kód: GEIAL526M
Évfolyam: II

Hét	Elmélet	Gyakorlat
1.	OLAP/DW alapfogalmak Statisztikai alapok áttekintése. Eloszlások és alap statisztikai tételek.	Adatelemzési rutinok készítése Java-ban
2	Python programozás alapjai. A nyelvi elemek áttekintése. Függvények, struktúrák, vezérlési szerkezetek. OOP alapú programozás.	Adatelemzési rutinok készítése Python-ban
3	Statisztikai tesztek áttekintése. T-teszt, F-teszt, ANOVA. Statisztikai csomagok Python-ban.	Adatelemzési rutinok készítése Python-ban
4	DW architektúra; MD modell, csillag és hópehely modellek	MD séma tervezése
5	MDX nyelv áttekintése. Műveleti algebra és parancsnyelv. Tupple; set, measures, dimensions, MDX függvények. Saiku interfész.	MDX programozása, Saiku
6	ETL folyamat architektúrája, transzformációs lépések, adattisztítás; DW fizikai rétege	MDX programozása, Saiku
7.	Adatbányászat alapfogalmai;Adatbányászat áttekintése, adatelemzés lépései, DM eszközök áttekintése: Python, rapidMiner	Adatelemzési rutinok készítése Python-ban
8.	Klaszterezési technikák, k_means és HAC módszerek, eszközök Python-ban	Klaszterezés Python-ban
9.	Osztályozás alapeszközei, Bayes-alapú osztályozók, döntési fák, eszközök Python-ban	Osztályozás Python-ban
10.	Osztályozás további eszközei: neurális háló alapú osztályozás, eszközök Python-ban, Keras	Osztályozás Python-ban
11..	Asszociatív szabályok feltárása, kivülálló elemek elemzése, eszközök Python-ban.	Piaci kosár elemzés Python-ban
12.	Előrejelzési módszerek, trend és véletlen elemek, periódikus elemek, eszközök Python-ban.	Előrejelzés Python-ban

Kötelezõ irodalom

- Tantárgyi honlap: moodle.iit.uni-miskolc.hu

Ajánlott irodalom

- M.J. Zaki, W. Meira Jr.: Data Mining and Analysis (Fundamental Concepts and Algorithms)
- http://www.dataminingbook.info/pmwiki.php

A tárgy lezárásának módja: aláírás, vizsgajegy
Évközi számonkérés: több egyedi heti szintű kis feladat
Aláírás megszerzés feltételei:
- legalább 7 gyakorlaton való részvétel a szorgalmi időszakban és
- az egyéni feladatok sikeres megvédése.
Pótlás módjai:
- Az egyéni feladatok pótlása az utolsó szorgalmi héten történik
- Az gyakorlati számonkérés pótlása a szorgalmi időszakban és a vizsgaidőszakban egyszer, a kijelölt napon.

Vizsga formája: írásbeli és szóbeli Az írásbeli rész legalább elégséges teljesítése után következik a szóbeli rész. Az írásbelin elméleti vagy gyakorlati feladatok szerepelhetnek. Az írásbeli és szóbeli rész értékelése: 0%-50% : elégtelen 50%-62% : elégséges 62%-75% : közepes 75%-88% : jó 88%-100% : jeles. Az eredő teljesítmény a 0.667*írásbeli + 0.333*szóbeli képlettel kerül meghatározásra, melyhez jegy a megadott táblázat szerint rendelődik. Elégtelen írásbeli elégtelen vizsgajegyet jelent. A szóbelin a megjelenés kötelező.

Minta vizsga kérdéssor

1. Lineáris regresszió jelentése, működése; Korreláció jelentése, képlete.
2. MD algebrai műveletek és megvalósításuk PE-ben
3. K-means és QTC klasztering és algoritmusuk
4. A back-propagation NN módszer célja és a működése (formulák)

Vizsga javítókulcs

1. Lineáris regresszió jelentése, működése; Korreláció jelentése, képlete.(10 pont)
   - regresszió: közelítő függvény meghatározása
   - paraméterérték optimalizálása
   - lineáris regresszió: paraméterekben lineáris
   - eltérés négyzet minimalizálása
   - lineáris egyenletrendszer megoldása
   - korreláció: együttmozgás mérése, képlete
   - értéke -1, 1
   - ha független, akkor zérus
   - ha nem zérus, akkor nem független

2. MD algebrai műveletek és megvalósításuk PE-ben (10 pont)
   - MD adatmodell: adatkocka, nincs join
   - hierarchikus dimenzió
   - slice and dice: szelekció és projekció
   - drill down: részletesebb szintre áttérés
   - roll up : összegzőbb szintre lépés
   - fold: aggregáció
   - PE: LIMIT TO és KEEP
   - PE: TOTAL()

3. K-means és QTC klaszterezés és algoritmusuk (10 pont)
   - Klaszterezés jelentése
   - K-means: K középpont tetszőleges kijelölése
   - a középpontokhoz a legközelebbi elemek meghatározása
   - minden csoporthoz az új középpont kijelölése
   - addig mozgás, amíg van eltérés
   - célfüggvény: távolságösszeg
   - klaszterszám meghatározás problémája
   - QTC: klaszteren belüli távolság minimalizálása
   - határérték a testvérek távolságára
   - klaszter felbontás algoritmusa
   - többértelműség kezelése

4. A back-propagation NN módszer célja és a működése (formulák) (10 pont)
   - neurális háló mint függvény közelítés
   - neurális háló paraméterei az él súlyok
   - optimális paraméterű háló keresése
   - gradiens módszer alkalmazása
   - belső függvény változója szerinti deriválás
   - láncszabály
   - hibaérték a súlyok függvényében
   - sigmoid függvény deriválása
   - külső élek súlya szerinti gradiens
   - belső élek súlya szerinti gradiens
   - tanítási ciklusok (epoch)

- jóság (accuracy) mérése