Tárgytematika / Course Description

Data Analysis

GKNM_MSTA025

Tárgyfelelős neve / Teacher's name: dr. Takács Gábor

Félév / Semester: 2018/19/1

Beszámolási forma / Assesment: Vizsga

Tárgy heti óraszáma / Teaching hours(week): 4/0/0

Tárgy féléves óraszáma / Teaching hours(sem.): 0/0/0

OKTATÁS CÉLJA / AIM OF THE COURSE

A tárgy célja a számítógépes adatelemzés alapvető módszereinek a bemutatása a gépi tanulás területére történő kitekintéssel. Emellett a tárgy bevezeti a hallgatókat egy konkrét adatelemző eszköz használatába, valós életből vett adathalmazok vizsgálatán keresztül.

TANTÁRGY TARTALMA / DESCRIPTION

- Az adatelemzés, adatbányászat ill. gépi tanulás fogalma, célja, folyamata. Néhány látványosabb alkalmazás.
- A gépi tanulás alapvető feladatai: Osztályozás, regresszió, klaszterezés. Matematikai alapok átismétlése.
- Python programozási alapok: A nyelv jellemzői, egyszerű adattípusok, kollekciók, vezérlési szerkezetek.
- Python programozási alapok: Comprehension-ök, kicsomagolás, haladó indexelés és iterálás, függvények, fájlkezelés.
- A numpy numerikus számítási modul. Tömbök létrehozása, résztömbök, műveletek, broadcasting.
- A pandas adatelemző modul. Series és DataFrame adatszerkezet. CSV fájlok betöltése.
- A K legközelebbi szomszéd algoritmus. Tesztelés a "Római helyszínelők" feladaton. Pontfelhődiagram.
- Lineáris regresszió. Egyváltozós eset. Tesztelés az MLB adathalmazon.
- Többváltozós lineáris regresszió. Megvalósítás a scikit-learn segítségével. Tesztelés a Boston Housing adathalmazon.
- A túltanulás jelensége. Ridge regesszió. Keresztkiértékelés.
- Logisztikus regresszió. Egy- és többváltozós eset. Tesztelés a Wisconsin Breast Cancer adathalmazon.
- Particionáló módszerek: Döntési fák, véletlen erdők, gradient boosting. Tesztelés a Boston Housing adathalmazon.
- Klaszterezés: A K means algoritmus, hierarchikus klaszterező eljárások.
- Összefoglalás. A féléves anyag rendszerezése.

SZÁMONKÉRÉSI ÉS ÉRTÉKELÉSI RENDSZERE / ASSESMENT'S METHOD

A tárgy számítógépes vizsgával zárul, ahol a hallgatóknak egyszerű adatelemzési feladatokat kell megoldaniuk Python nyelven. A vizsgán rendelkezésre álló idő 90 perc. Ponthatárok: 21-24: jeles, 18-20: jó, 15-17: közepes, 12-14: elégséges.

KÖTELEZŐ IRODALOM / OBLIGATORY MATERIAL

- Hastie, R. Tibshirani, J. Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition, Springer, ISBN 978-0387848570, 2009.
- I. Witten, E. Frank, M. Hall, Data Mining: Practical Machine Learning Tools and Techniques, Third Edition, Morgan Kaufmann, ISBN 978-0123748560, 2011.
- Bodon F.: Adatbányászati algoritmusok, online tanulmány, http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf.