Tárgytematika / Course Description

Adatelemzés

GKNM_MSTM025

Tárgyfelelős neve / Teacher's name: dr. Harmati István

Félév / Semester: 2022/23/2

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 és a gépi tanulás alapvető módszereinek bemutatása. Emellett a tárgy bevezeti a hallgatókat egy konkrét adatelemző eszköz használatába, valós életből vett adathalmazok vizsgálatán keresztül.

TANTÁRGY TARTALMA / DESCRIPTION

- Az adatelemzés 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 csomag. Tömbök létrehozása, résztömbök, műveletek, broadcasting. Egyváltozós lineáris regresszió.
- A K legközelebbi szomszéd algoritmus. Tesztelés a Római Helyszínelők feladaton.
- A pandas adatelemző csomag. Series és DataFrame adatszerkezet. CSV fájlok betöltése. Légszennyezettségi adatok elemzése.
- Többváltozós lineáris regresszió. Tesztelés a Boston Housing adathalmazon. A scikit-learn alapjai. Keresztkiértékelés.
- Logisztikus regresszió. Egy- és többváltozós logisztikus eset. Tanítás Newton-módszerrel. Tesztelés a Wisconsin Breast Cancer adathalmazon.
- A túltanulás jelensége. L1 és L2 regularizáció. Ritka mátrixok. Regularizált ritka logisztikus regresszió tesztelése az SMS Spam adathalmazon.
- Neurális hálózatok. A többrétegű perceptron modell. Tanítás sztochasztikus gradiens módszerrel, tesztelés a Phishing Websites adathalmazon.
- Döntési fák. A döntési tönk és a döntési fa modell, tanítás "brute force" módszerrel, tesztelés a Boston Housing adathalmazon.
- Ensemble módszerek. Véletlen erdő, gradient boosting. Tesztelés a Boston Housing adathalmazon.
- Klaszterezés. A K-means algoritmus. Adatvizualizáció. A t-SNE algoritmus.

SZÁMONKÉRÉSI ÉS ÉRTÉKELÉSI RENDSZERE / ASSESMENT'S METHOD

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.