Labortor 4

Versiunea 2018-03-123

Urmariti documentul pentru cerintele ce se vor adauga. Tema se va preda la laborator.

Reprezentarea grafica a datelor

Preluati un fisier de date numerice dintr-unul din repository-urile date in cursul 4. Reprezentati valorile din setul de date prin:

1. Histograme pentru fiecare atribut
2. Boxplots
3. Scatter plotts pentru perechi de atribute
4. Matrix scatter plots
5. Vizualizare de matrice de corelatie
6. Coordonate paralele
7. Alte tipuri de grafice

Se pot urmari exemplele din prezentarea ./Exemple/curs3_IDM.pdf. Pentru reprezentare se poate folosi Matplotlib sau Seaborn (https://seaborn.pydata.org/).

Missing value imputation

Implementati tehnici pentru umplerea valorilor lipsa dintr-un set de date. Implementati:

1. Mean imputation (https://www.iriseekhout.com/missing-data/missing-data-methods/imputation-methods/) - o valoare lipsa se inlocuieste cu media valorilor cunoscute
2. Utilizarea valorii celei mai frecvente, pentru date catogoriale
3. Median value imputation - valorie lipsa se umplu cu mediana
4. kNN pentru regresie, cu diferite valori ale lui k
5. Alte metode de regresie, in care valorile complete sunt folosite pentru prezicerea valorilor lipsa
6. Algoritmul expectation maximization (https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm)

Aplicati aceste metode pentru Pandas Datframes. Scrieti functii care implementeaza metodele mentionate. Puteti modele de regresie, EM etc. puteti folosi bibliteca sklearn.

Resurse:

1. How to Handle Missing Data (https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4)
2. Comparison of Data Imputation Techniques and their Impact (https://arxiv.org/ftp/arxiv/papers/0812/0812.1539.pdf)
3. Missing-data imputation (http://www.stat.columbia.edu/~gelman/arm/missing.pdf)
4. Missing data imputation using statistical and machine learning methods in a real breast cancer problem (https://www.sciencedirect.com/science/article/pii/S0933365710000679)

Laboratorul 4 se va prezenta cel tarziu in 20 aprilie.