Curs **Data Mining**, Master BDTS, an 1, sem 1
An univ. 2020-2021
Curs: Florentin Ipate
Laborator: Alexandru Tache

I. Obiective: asimilarea conceptelor si tehnicilor principale de Data Mining si familiarizarea cu limbajul R

II. Cerinte

Notarea se va face pe baza de proiect curs (65%) si proiect laborator (35%). Proiectele se vor efectua in echipe (dimensiunea echipei pentru fiecare proiect de curs este specificata mai jos).

Tema proiectului de curs va fi aleasa de comun acord cu cadrele didactice de la curs si laborator, dupa formarea echipei. Tema proiectului de laborator va fi stabilita de cadrul didactic de laborator, in functie de tema proiectului de curs.

Prezentarea proiectului de curs se va face la curs, conform programarii stabilite cu cele doua cadre didactice, iar prezentarea proiectului de laborator se va face la laborator, conform programarii stabilite de cadrul didactic de la laborator.

Prezentarea proiectului de curs va fi sub forma de slide-uri, insotite de demo-uri atunci cand este cazul. La prezentare este necesara prezenta intregii echipe, fiecare student descriind principala sa contributie la proiect. Atunci cand echipa considera ca nu toti membrii echipei au avut o contributie egala la realizarea proiectului, se va indica, in procente, contributia estimata a fiecaruia.

III. Teme proiecte curs

A. Statistical Techniques for DM

1. Applied statistical testing for DM
Bibliografie: [Davies: ch 17-19]
Marime echipa: 3 studenti
Durata prezentare: aprox. 22 min (3 x 7.5 min)

Scurta descriere: Tema cuprinde cateva metode statistice de baza, folosite - spre exemplu - in A/B testare (https://en.wikipedia.org/wiki/A/B_testing) dar nu numai. Conceptele de baza vizate sunt

- distributia de esantionare si teorema limita centrala;
- intervale de incredere;
- teste de ipoteza pentru medii, proportii si variabile categorice;
- testele ANOVA si Kruskal – Wallis.

Cerinta: Raportul va prezenta, sintetic, conceptele principale, acestea fiind exemplificate cu linii relevante de cod. Studentii vor proba intelegerea acestor concepte prin rezolvarea exercitiilor capitolelor respective.

2. **Linear regression modelling**
   Bibliografie: [Davies: ch 20-22]
   Marime echipa: 3 studenti
   Durata prezentare: aprox. 22 min (3 x 7.5 min)

Scurta descriere: Tema vizeaza intelegerea conceptului fundamental (aflat la intersectia dintre statistica clasica si machine learning) de regresie liniara. Sunt avute in vedere notiunile de

- regresie liniara simpla (ex: venit = coef1*varsta)
- regresie liniara multipla (ex: venit = coef1*varsta + coef2*nivel_de_ed + coef3*sex+...)
- conditii initiale (verificarea asumptiilor), selectia si diagnoza modelului

Cerinta: Raportul va prezenta, sintetic, conceptele principale, acestea fiind exemplificate cu linii relevante de cod. Studentii vor proba intelegerea acestor concepte prin rezolvarea exercitiilor capitolelor respective.

3. **Linear regression from scratch (tema optionala)**
   Bibliografie: [Ghatak: ch 4]
   Marime echipa: 2 studenti
   Durata prezentare: 15 min (2 x 7.5 min)

Proiectul presupune implementarea eficienta de la zero (fara biblioteci) a unor algoritmi de regresie liniara. Detalii suplimentare privind acest proiect si modul sau de prezentare vor fi stabilite de echipa de comun acord cu cele doua cadre didactice.

B. Exploratory Data Analysis (EDA)

Scurta descriere
„Analiza exploratorie a datelor (EDA) reprezintă o abordare a analizei seturilor de date pentru a rezuma principalele caracteristici ale acestora, adesea prin metode vizuale. Un model statistic poate fi folosit sau nu, dar în primul rând EDA este pentru a vedea ce ne pot spune datele dincolo de sarcina de modelare formală sau ipoteza de testare.” (https://en.wikipedia.org/wiki/Exploratory_data_analysis).

Grupajul de teme de mai jos are in vedere analiza exploratorie a datelor prin metode grafice. Sunt prezentate urmatoarele notiuni:

- variabile continue, relatii de dependenta si asociere, date continue multivariate (tema 4)
- date categorice (discrete) univariate si multivariate, valori lipsa si outlieri (tema 5)

4. **Graphical EDA I: continuous data**
   Bibliografie: [Unwin: ch 3, 5, 6]
   Marime echipa: 3 studenti
   Durata prezentare: aprox. 22 min (3 x 7.5 min)

Raportul va prezenta, sintetic, conceptele si tehnicile principale, acestea fiind exemplificate cu linii relevante de cod.

5. **Graphical EDA II: categorical data & quality control**
   Bibliografie: [Unwin: ch 4, 7, 9]
   Marime echipa: 3 studenti
   Durata prezentare: aprox. 22 min (3 x 7.5 min)

Raportul va prezenta, sintetic, conceptele si tehnicile principale, acestea fiind exemplificate cu linii relevante de cod.

C. Dimensionality Reduction Techniques

Scurta descriere

„Large data sets containing multiple samples and variables are collected everyday by researchers in various fields. Discovering knowledge from these data requires specific techniques for analysing data sets containing multiple variables. Multivariate analysis (MVA) refers to a set of techniques used for analysing a data set containing more than one variable.
Among these techniques, there are:

- cluster analysis for identifying groups of observations with similar profile according to a specific criterion (*vezi temele 9 si 10*);
- principal component methods (*temele 6 – 8*)" (*[KassambaraPCM], pag 5*)

"The type of principal component methods to use depends on variable types contained in the data set. This practical guide will describe the following methods:

i. *Principal Component Analysis (PCA)*, which is one of the most popular multivariate analysis methods. The goal of PCA is to summarize the information contained in a continuous (i.e., quantitative) multivariate data by reducing the dimensionality of the data without losing important information.

ii. *Correspondence Analysis (CA)*, which is an extension of the principal component analysis for analysing a large contingency table formed by two *qualitative variables* (or categorical data).

iii. *Multiple Correspondence Analysis (MCA)*, which is an adaptation of CA to a data table containing more than two categorical variables.

iv. *Factor Analysis of Mixed Data (FAMD)*, dedicated to analysing a data set containing both quantitative and qualitative variables.

v. *Multiple Factor Analysis (MFA)*, dedicated to analyse data sets, in which variables are organized into groups (qualitative and/or quantitative variables)." (*[KassambaraPCM], pag 6*)

6. **Principal components methods I: PCA**
   Bibliografie: [KassambaraPCM: ch 4; link1, link2, link3]
   Marime echipa: 2 studenti
   Durata prezentare: 15 min (2 x 7.5 min)

Raportul va prezenta metodele descrise, acestea fiind exemplificate cu linii relevante de cod.

7. **Principal components methods II: (multiple) correspondence analysis**
   Bibliografie: [KassambaraPCM: ch 5, 6]
   Marime echipa: 2 studenti
   Durata prezentare: 15 min (2 x 7.5 min)

Raportul va prezenta metodele descrise, acestea fiind exemplificate cu linii relevante de cod.

8. **Principal components methods III: factor analysis & HCPC**
   Bibliografie: [KassambaraPCM: ch 7, 8, 9]
   Marime echipa: 2 studenti
   Durata prezentare: 15 min (2 x 7.5 min)

Raportul va prezenta metodele descrise, acestea fiind exemplificate cu linii relevante de cod.

D. Unsupervised Machine Learning

Scurta descriere

„Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense) to each other than to those in other groups (clusters). […] Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their understanding of what constitutes a cluster and how to efficiently find them." (https://en.wikipedia.org/wiki/Cluster_analysis)

Temele vizeaza intelegerea principalilor algoritmi de clusterizare (tema 9) si a principalelor metode de validare a unei clusterizari (tema 10).

9. **Clustering methods I: main clustering algorithms**
   Bibliografie: [KassambaraCA: ch 3-7, 16, 19]
   Marime echipa: 3 studenti
   Durata prezentare: aprox. 22 min (3 x 7.5 min)

Raportul va prezenta, sintetic, conceptele si tehnicile principale, acestea fiind exemplificate cu linii relevante de cod.

10. **Clustering methods II: validation techniques**
    Bibliografie: [KassambaraCA: ch 11-15]
    Marime echipa: 2 studenti
    Durata prezentare: 15 min (2 x 7.5 min)

Raportul va prezenta, sintetic, conceptele si tehnicile principale, acestea fiind exemplificate cu linii relevante de cod.

E. Supervised Machine Learning

Scurta descriere

„Supervised learning is a method used to enable machines to classify objects, problems or situations based on related data fed into the machines. Machines are fed with data such as characteristics, patterns, dimensions, color and height of objects, people or situations repetitively until the machines are able to perform accurate classifications. Supervised learning is a popular technology or concept that is applied to real-life scenarios. Supervised learning is used to provide product recommendations, segment customers based on customer data, diagnose disease based on previous symptoms and perform many other tasks." (https://www.techopedia.com/definition/30389/supervised-learning)

Temele vizeaza intelegerea principalilor algoritmi de invatare supervizata: k nearest neighbours (kNN), naive Bayes, regresia logistica, metoda discriminantului liniar, arbori de decizie, SVM si altii.

11. **Classification tehniques I: kNN & naive Bayes**
    Bibliografie: [Lantz: ch 3, 4]
    Marime echipa: 2 studenti
    Durata prezentare: 15 min (2 x 7.5 min)

Raportul va prezenta tehnicile descrise, acestea fiind exemplificate cu linii relevante de cod.

12. **Classification tehniques II: logistic regression, LDA & QDA**
    Bibliografie: [JamesEtAl: ch 4]
    Marime echipa: 2 studenti
    Durata prezentare: 15 min (2 x 7.5 min)

Raportul va prezenta tehnicile descrise, acestea fiind exemplificate cu linii relevante de cod.

13. **Classification & regression tehniques: trees, SVMs and neural networks**
    Bibliografie: [Lantz: ch 5, 7]
    Marime echipa: 3 studenti
    Durata prezentare: aprox. 22 min (3 x 7.5 min)

Raportul va prezenta tehnicile descrise, acestea fiind exemplificate cu linii relevante de cod.

F. Time Series Analysis

Scurta descriere

„A time series is a series of data points indexed (or listed or graphed) in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-time data. Examples of time series are heights of ocean tides, counts of sunspots, and the daily closing value of the *Dow Jones Industrial Average*." (https://en.wikipedia.org/wiki/Time_series)

14. **Time series II: decompositions & exponential smoothing**
    Bibliografie: [Hyndman, Athanasopoulos: ch 6, 7]
    Marime echipa: 2 studenti
    Durata prezentare: 15 min (2 x 7.5 min)

Raportul va prezenta, sintetic, conceptele si tehnicile principale, acestea fiind exemplificate cu linii relevante de cod. Studentii vor proba intelegerea acestor concepte prin rezolvarea exercitiilor capitolelor respective.

15. **Time series III: ARIMA models**
    Bibliografie: [Hyndman, Athanasopoulos: 8]
    Marime echipa: 2 studenti
    Durata prezentare: 15 min (2 x 7.5 min)

Raportul va prezenta, sintetic, conceptele si tehnicile principale, acestea fiind exemplificate cu linii relevante de cod. Studentii vor proba intelegerea acestor concepte prin rezolvarea exercitiilor capitolelor respective.

G. Network Analysis / Graph Mining

Scurta descriere

"Network science is a broad approach to research and scholarship that uses a relational lens to study and understand biological, physical, social, and informational systems. The primary tool for network scientists is network analysis, which is a set of methods that are used to

(1) visualize networks

(2) describe specific characteristics of overall network structure as well as details about the individual nodes, ties, and subgroups within the networks, and

(3) build mathematical and statistical models of network structures and dynamics.

Because the core question of network science is about relationships, most of the methods used in network analysis are quite distinct from the traditional statistical tools", having a modern flavour of their own. (*[Luke] A User's Guide to Network Analysis in R, pag 3*)

16. **Graph mining II: analysis & modeling**

Bibliografie: [Luke: ch 7-10]

Marime echipa: 2 studenti

Durata prezentare: 15 min (2 x 7.5 min)

Raportul va prezenta, sintetic, conceptele si tehnicile principale, acestea fiind exemplificate cu linii relevante de cod.

H. Frequent pattern mining

Scurta descriere

„Frequent patterns are itemsets, subsequences, or substructures that appear in a data set with frequency no less than a user-specified threshold. For example, a set of items, such as milk and bread, that appear frequently together in a transaction data set, is a frequent itemset. [...] Finding frequent patterns plays an essential role in mining associations, correlations, and many other interesting relationships among data. Moreover, it helps in data indexing, classification, clustering, and other data mining tasks as well. Frequent pattern mining is an important data mining task and a focused theme in data mining research." (http://www.kdd.org/kdd2016/topics/view/frequent-pattern-mining)

Tema vizeaza intelegerea principalelor tehnici din aceasta arie:

- apriori / eclat, folosit traditional in analiza cosului de cumparaturi (market basket analysis);
- filtrarea colaborativa, metoda clasica folosita in domeniul sistemelor de recomadare.

17. **Association rules & collaborative filtering**

Bibliografie: [Chapman, Feit: ch 12], [Shmueli: ch 14]

Marime echipa: 2 studenti

Durata prezentare: 15 min (2 x 7.5 min)

Raportul va prezenta, sintetic, conceptele si tehnicile principale, acestea fiind exemplificate cu linii relevante de cod.

IV. Bibliografie

[Davies] The Book of R
[Ghatak] ML with R
[Unwin] Graphical Data Analysis with R
[KassambaraPCM] Practical Guide to Principal Component Methods
[link1] http://www.sthda.com/english/articles/31-principal-component-methods-in-r- practical-guide/112-pca-principal-component-analysis-essentials/
[link2] http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/118-principal-component-analysis-in-r-prcomp-vs-princomp/
[link3] http://www.sthda.com/english/articles/31-principal-component-methods-in-r-practical-guide/119-pca-in-r-using-ade4-quick-scripts/
[Lantz] Machine Learning with R
[JamesEtAl] Introduction to Statistical Learning with R
[KassambaraCA] Practical Guide to Cluster Analysis in R
[Hyndman, Athanasopoulos] Forecasting - principles and practice, https://otexts.org/fpp2/
[Luke] A User's Guide to Network Analysis in R
[Chapman, Feit] R for Marketing and Research Analytics
[Shmueli] DM for Business Analytics

V. **Textbook pentru curs**: Tan, Steinbach, Karpatne, Kumarhttps. Introduction to Data Mining, 2nd Edition by://www-users.cs.umn.edu/~kumar001/dmbook/index.php