INTRODUCTION TO DATA SCIENCE (LABS)

13/10/2020

WFAiS UJ, Informatyka Stosowana

I stopień studiów

http://th-www.if.uj.edu.pl/~erichter/dydaktyka/Dydaktyka2020/DataScience-2020/index.html

Ready, but I can still have few ideas to make course more interesting.

Introduction to Data Science (Wprowadzenie do analityki danych)

Wydział Fizyki, Astronomii i Informatyki Stosowanej,

Uniwersytet Jagielloński w Krakowie

Rok akademicki 2020/2021

Office hours: Tuesday 13:00 - 14:00; office G-0-10.
COVID-19: Please use MSTeams system or email if applicable.

Lectures & Assigments:

Week	Lecture slides	Lab slides	Python scripts with assignments	Datasets	Tutorials
13.10.2020	Introduction Data exploration	Introdution lab	assignment-0-python assignment-0-numpy assignment-0-numpy-matplotlib assignment-0-pandas assignment-1	kc_house_data.csv.zip info on kc_house	HowToStart DS_cheatsheet_numpy.pdf DS_cheatsheet_matplotlib.pdf DS_cheatsheet_jupyter_notebook.pdf DS_cheatsheet_pandas.pdf
20.10.2020	Regression-Primer Regression-Advanced-I		assignment-2		S.Rashki python-machine-learning-book rashka_ch04.ipynb rashka_ch10.ipynb
27.10.2020	Regression-Advanced-II		assignment-3		numpy tutorial matrix algebra tutorial
3.11.2020	Regression-Advanced-III Classification-Primer				
10.11.2020	Classification-Advanced-I		assignment-4	data:amazon_baby.csv.zip	scikit-learn: LogisticRegression
17.11.2020	Classification-Advanced-II				
24.11.2020	Classification-Advanced-III Clustering-Primer		assignment-5	data:people_wiki.csv.zip	
1.12.2020	Clustering&Retrieval-Advanced-I				
8.12.2020	Clustering&Retrieval-Advanced-II				
15.12.2020	Clustering&Retrieval-Advanced-III Recommending-System-Primer		Its time to start developing your personal Data Science project: Select one from the list below or create your own: Proj_NYcrime.txt Proj_PolandClimateChange.txt Proj_TrollTweets.txt Proj_VoteCast.txt Proj_Classification.txt Proj_Clustering.txt Proj_Regression.txt	lending-club-data.csv.zip https://snap.stanford.edu/data/amazon/ https://mlnce.umass.edu/ml/datasets.html https://data.world/	
12.01.2021	Modeling, simulation, Monte Carlo methods		assignment-6		
19.01.2021	Statistical Inference		assignment-7		OpenIntro Statistics Z. Pan at Standford Uni.
26.01.2021	Multivariate Analyses and Artificial Neural Network				

Lectures are based on the materials from Coursera:

Dr. Mine Çetinkaya-Rundel "Data Analysis and Statistical Inference"
C. Guestrin and E. Fox "Machine Learning Specialization"
- Foundation: link
- Regression: link
- Classification: link
- Clustering and Retrieval: link

Related interesting material from Coursera
D. Peng, J. Leek and B. Caffo, "Exploratory Data Analysis"
J. Leskovec, A. Rajaraman and J. Ullman, "Mining Massive Datasets"
B. Caffo, R. D. Peng and J. Leek, "Regression Models"

Data Science applications in physics:
B. Nachman,
"Advanced Machine Learning for Classification, Regression, and Generation in Jet Physics"
M. Stoye,
"ML applications in CMS"
ML techniques in HEP Workshop, Berkeley Laboratory, 11 - 13 December 2018
https://indico.physics.lbl.gov/indico/event/546/

Collection of datasets
http://mlnce.umass.edu/ml/datasets.html
http://faculty.marshall.usc.edu/gareth-james/ISL/data.html
http://snap.stanford.edu/data/amazon/

Useful links:
https://turi.com/download/install-graphlab-create-aws-coursera.html
https://turi.com/download/academic.html
https://github.com/turi-code/SFrame

Clustering
http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.CountVectorizer.html
http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

Boosting
https://turi.com/learn/userguide/supervised-learning/boosted_trees_classifier.html
https://homes.cs.washington.edu/~tqchen/pdf/BoostedTree.pdf
http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html

----

[1] https://class.coursera.org/statistics
[2] http://www.openintro.org/stat/textbook.php
[3] https://class.coursera.org/exdata-003
[4] https://class.coursera.org/mmds
[5] http://www.mmds.org/
[6] http://www.cs.cmu.edu/~awm/tutorials.html

Additional materials
http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf
http://statweb.stanford.edu/~tibs/ElemStatLearn/
http://en.wikipedia.org/wiki/Determining_the_number_of_clusters_in_a_data_set
http://www.youtube.com/watch?v=wQhVWUcXM0A

Link to lectures given in 2017
http://th-www.if.uj.edu.pl/~erichter/dydaktyka/Dydaktyka2017/AiSAD-2017/index.html

Link to lectures given in 2014
http://th-www.if.uj.edu.pl/~erichter/dydaktyka/Dydaktyka2014/AiSAD-2014/index.html

Ostatnia modyfikacja: 7 October 2020

Elżbieta Richter-Wąs

Wstecz

LABS

- Normal times: run in person
  - Time for you to write your code and (for me) to disscuss with each student her/his progress with assignments.
- COVID-19 times: run on-line
  - We will go through content of assigments, present (mostly you) results of analyses and observations, have short oral presentations (please be active)
  - Ocassion to share with everybody problems, exchange snippets of the code or interesting observations.

Assignments, Project, Short Presentations

This is not a course of programming, but you will be expected to write programs.

- **Baseline is python + anaconda libraries.**
- **You can use also R or other Data Science specific programming language/library**

I will not be teaching you programming or helping to debug your code, you are on your own …

For labs you will be graded with:

- **completed assignments: *7* + optional - max 72 scores**
- **personalised project - max 25 scores**
- **short topical presentations - max 25 scores**

Graded will be not (necessarily) quality of the code, but maturity of how you analyse and interpret the data.

To pass the course you need to collect at least 65 scores.

Assignments, Project, Short Presentations

PEGAZ system:

This system we will use to collect your assigments/projects/short presentations

- **I will be sending you back comments**
- **You will see your grades there**

Please don't use email to send me your scripts!

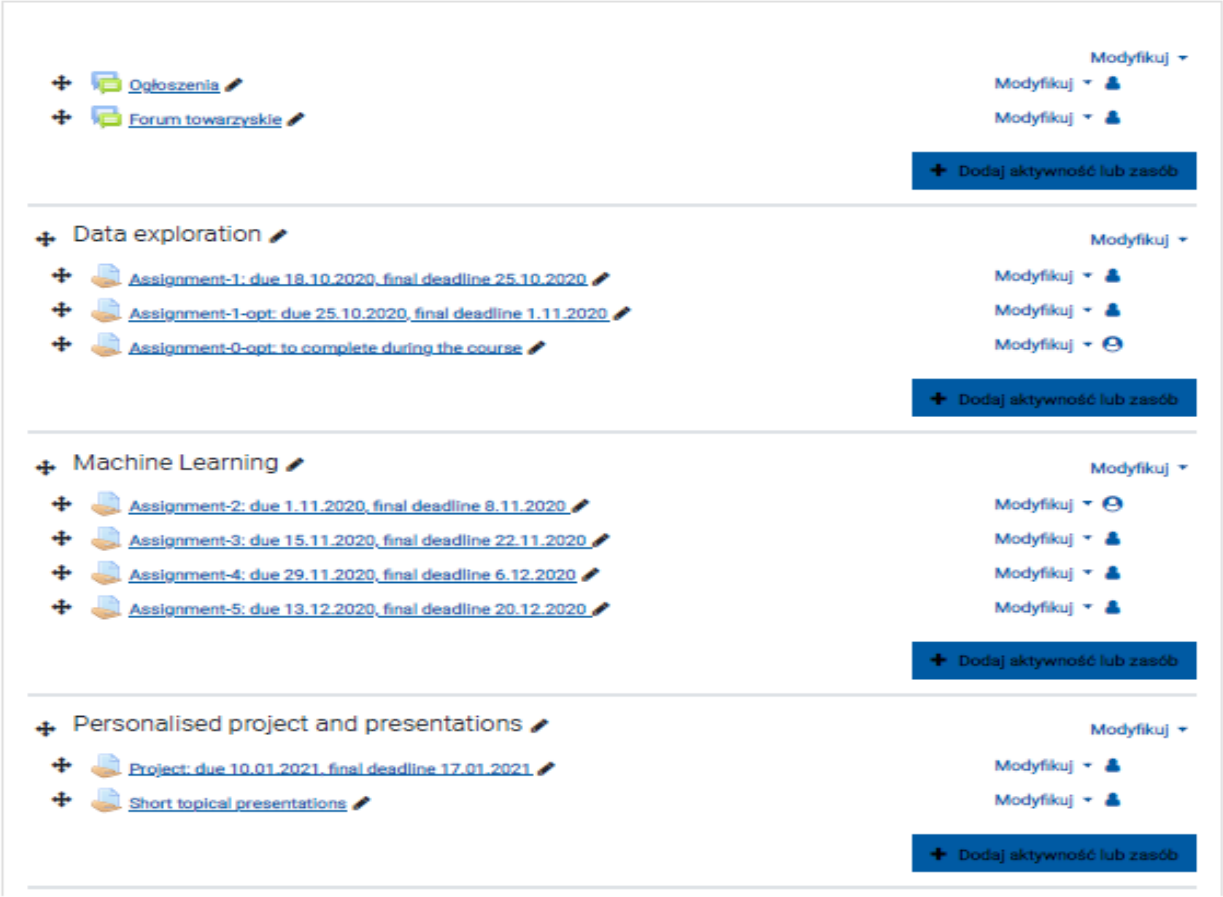
MSTeams:

This system we will use for on-line classes

- **you can use it also for communication among yourself, eg. setting up chats/meetings within a team**
- **for communication with me preferably use emails**

PEGAZ system:

use to submit your assignments/projects/presentations.



PEGAZ system:

use to submit your assignments/projects/presentations.