CME594 Introduction to Data Science

Instructor: Professor S. Derrible, 2071 ERF, derrible@uic.edu
Office hours: open door policy

Hours: Thursday: 5:00 – 7:30

Location: SH 103

Summary: This course introduces students to techniques of complexity science and machine learning with a focus on data analysis. One new technique is covered every week, including: scaling laws, principal component analysis, hierarchical clustering, decision tree learning, neural networks, network science, agent-based modeling and text mining. The main assessment is a final paper where the students are asked to pick any data set (preferably from their own research) and apply one or multiple techniques from the course. No programming experience is required, but the course includes Python coding.

Objectives: This course aims to provide students with introductory knowledge of several data science techniques that can be used for data analysis. The material learned should then be useful in the student’s own research. More specifically, at the end of this course, students should be able to:
1. explain the main concepts behind all the techniques covered
2. identify the type of technique preferable to use depending on the type of data to analyze
3. use the various Python libraries learned to be able to apply these techniques
4. apply rigorously one or multiple of these techniques learned in their own research

Textbook: No textbook is required, but the following books may be useful:


Software: # Python 2.7.xx: https://www.python.org/downloads/
or simply install
# Anaconda – python 2.7 (recommended package that includes Python and most recommended libraries; sometimes the 32bit version works better even for 64bit computers): https://www.continuum.io/downloads

# NetLogo (sometimes the 32bit version works better even for 64bit computers): https://ccl.northwestern.edu/netlogo/

**Tentative Grading Policy:**  
Attendance, participation, behavior (15%)  
Homework (25%)  
Report and Presentation of Data Science Technique (15%)  
Abstract (5%)  
Presentation (5%)  
Final Paper (35%)  

Work submitted late may receive a penalty.

**Plagiarism:**  
Plagiarism is a serious offense and it will not be tolerated; see university policy. All reviews, papers and any other submitted material will be run through a plagiarism tool.

**Attendance Policy:**  
All students are required to attend the lectures and be on time. If at any moment a student is to be absent, he/she should have discussed it prior with the instructor.

**Professional Conduct:**  
Students are always expected to conduct themselves with the utmost respect towards the instructor and their fellow students. Cellphones are to be turned off.

**Class Schedule and Readings**

**Week 1: Python Installation and Tutorial**

- Install python and recommended libraries (see software section above)

**Readings:**

  
or
  

**Week 2: Scaling Laws, Zipf’s Law, and Regression Analysis**

- Install scikit-learn in python.
Readings:

Supplementary Readings:
- Cristelli, M., Batty, M., Pietronero, L., 2012, “There is More than a Power Law in Zipf”, Scientific Reports 2(812)

Week 3: Principal Component Analysis

Readings
- Smith, L. I., 2002, “A tutorial on Principal Components Analysis”, Notes for Course COSC453

Supplementary Readings:

Week 4: Introduction to Basic Probability for Data Mining

Readings:
- Han, J., Kamber, M., Pei, J., 2011, “Chap. 6 Mining Frequent Patterns, Associations, and Correlations: Basic Concepts and Methods”, in Data Mining: Concepts and Techniques, Elsevier Science.

Supplementary Readings:
- Han, J., Kamber, M., Pei, J., 2011, “Chap. 4 Data Warehousing and Online Analytical Processing”, in Data Mining: Concepts and Techniques, Elsevier Science.
Week 5: Introduction to Machine Learning and k-Nearest Neighbor Algorithm

Readings:

Supplementary Readings:

Week 6: Clustering Analysis

Readings:
- scikit-learn, 2016, “2.3 Clustering”, scikit-learn.org (accessed Feb. 15, 2016)

Supplementary Readings:

Week 7: Support Vector Machine

Readings
• Berwick, R., 2009, “An Idiot’s guide to Support vector machines (SVMs)”, Notes for Course CAP 6412 (Advanced Computer Vision)

Supplementary Readings:

Week 8: Decision Tree Learning and Random Forests

Readings

Supplementary Readings:

Week 9: Neural Networks and Deep Learning

Readings

Supplementary Readings:

Week 10: Network Science
• Install the Python library networkx.

Readings:
• Barabási, A.-L., 2014, “Chap. 2 Graph Theory” in Network Science, Creative Commons: CC BY-NC-SA 2.0. PDF V26, 05.09.2014

Supplementary Readings:

Popular Books:

Week 11: No Class (spring break)

• No readings

Week 12: Abstract Presentation

• No readings / Abstract Assignment Presentation

Week 13: Text Mining

• Install the nltk library, nltk data, the TextBlob library, and the gensim library.

Readings:

Supplementary Readings:

Week 14: Fisher Information

• Download Fisher Information library at http://csun.uic.edu/codes/fisher.html

Readings

Supplementary Readings:

Week 15: Agent-Based Modeling

• Install NetLogo (see software section above)

Readings:
• School of Informatics (U. Edinburgh), 2010, “Cellular Automata and Agent models for ecosystems”, Slides for course Computational Methods for Global Change Research 2009-2010

Supplementary Readings:

Week 16: Network-based Frequency Analysis

Readings
• Derrible, S., and Ahmad, N., 2015, “Network-Based and Binless Frequency Analyses”, PLoS ONE, 10(11): e0142108

Week 17: Final Presentation and Paper Deadline

• No readings