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: TBD

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:

Han, J., Kamber, M., Pei, J., 2011, "Data Mining: Concepts and Techniques", Elsevier Science.
Murphy, K., 2012, "Machine Learning: A Probabilistic Perspective", MIT Press, Cambridge, MA.
Batty, M., 2007, "Cities and Complexity", MIT Press, Cambridge, MA.

Software: # Python 2.7.xx: https://www.python.org/downloads/
Libraries: NumPy, SciPy, Pandas, igraph, SciKit learn (for Windows, see: http://www.lfd.uci.edu/~gohlke/pythonlibs/)
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/

Grading Policy: Attendance, participation, behavior (15%)
Homework (25%)
In Class Technique Presentation and Application (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: 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:

- Week 1: Python coding tutorial
- Week 2: Scaling Laws, Zipf's Law, and Regression Analysis
- Week 3: Principal Component Analysis
- Week 4: Introduction to Machine Learning and Basic Probability for Data Mining
- Week 5: Introduction to Scikit-Learn and k-Nearest Neighbor Algorithm
- Week 6: Clustering Analysis
- Week 7: Support Vector Machine
- Week 8: Decision Tree Learning and Random Forests
- Week 9: Neural Networks and Deep Learning
- Week 10: Fisher Information
- Week 11: No Class (spring break)
- Week 12: Abstract Presentation
- Week 13: Text Mining
- Week 14: Network Science
- Week 15: Agent-Based Modeling
- Week 16: Network-Based Frequency Analyses
- Week 17: Final Presentation and Paper Deadline