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:

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.
Barabási, A-L., 2014, "Network Science", Creative Commons: CC BY-NC-SA 2.0. PDF V26, 05.09.2014
Batty, M., 2013, "The New Science of Cities", 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/

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:

- Learnpython.org: https://www.learnpython.org/ (Learn the Basics and Data Science Tutorials) (accessed Jan. 5, 2017)
  or
- A Byte of Python: https://python.swaroopch.com/ (accessed Jan. 5, 2017)

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

- Install scikit-learn in python.

Readings:

- West, G., 2011, "The surprising math of cities and corporations", TED Talk (accessed Jan. 11, 2016)
- Arcaute, E., et al., 2014, "Constructing cities, deconstructing scaling laws", *Journal of the Royal Society Interface*, 12(102):2014.0745
- Saichev, A., Malevergne, Y., Sornette, D., 2010, "Introduction", in *Theory of Zipf's Law and Beyond*. Ed. Saichev, A., Malevergne, Y., Sornette, D., pp.1-7, Springer Berlin Heidelberg.

Supplementary Readings:

- Bettencourt, LMA, et al., 2007, "Growth, innovation, scaling, and the pace of life in cities", *Proceedings of the National Academy of Science (PNAS)*, 104(17):7301-7306
- Bettencourt, LMA, 2013, "The Origins of Scaling in Cities", *Science*, 340(6139):1438-1441
- 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:

- scikit-learn, 2016, "2.5. Decomposing signals in components (matrix factorization problems)", scikit-learn.org (accessed Jan. 20, 2017)

Week 4: Introduction to Basic Probability for Data Mining

Readings:

- Han, J., Kamber, M., Pei, J., 2011, "Chap. 1 Introduction", in *Data Mining: Concepts and Techniques*, Elsevier Science.
- Han, J., Kamber, M., Pei, J., 2011, "Chap. 2 Getting to Know Your Data", in *Data Mining: Concepts and Techniques*, Elsevier Science.
- 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. 3 Data Preprocessing", in *Data Mining: Concepts and Techniques*, Elsevier Science.
- 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:

- pythonprogramming (video), 2016, "Intro to Machine Learning with Scikit Learn and Python", pythonprogramming.net (accessed Jan. 29, 2017)
- Han, J., Kamber, M., and Pei, J., 2011, "Section 9.5.1 k-Nearest-Neighbor Classifiers", in Chap. 9 Classification: Advanced Methods in *Data Mining: Concepts and Techniques*, Elsevier Science.
- Markham, K., 2015, "scikit-learn video #4: Model training and prediction with K-nearest neighbors", The official blog of kaggle.com (accessed Jan. 29, 2017)
- Markham, K., 2015, "scikit-learn video #5: Choosing a machine learning model", The official blog of kaggle.com, accessed Jan. 29, 2017
- scikit-learn, 2016, "1.6. Nearest Neighbors", scikit-learn.org (accessed Jan. 29, 2017)

Supplementary Readings:

- Sehn Korting, T. (video), 2014, "How kNN algorithm works", YouTube.com (accessed Jan. 29, 2017)
- Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U., 1999, "When Is "Nearest Neighbor" Meaningful?", in *Database Theory — ICDT'99*, Proceedings of 7th International Conference, Jerusalem, Israel, January 10-12, Springer.

Week 6: Clustering Analysis

Readings:

- Han, J., Kamber, M., and Pei, J., 2011, "Chap. 10 Cluster Analysis: Basic Concepts and Method" in *Data Mining: Concepts and Techniques*, Elsevier Science.
- scikit-learn, 2016, "2.3 Clustering", scikit-learn.org (accessed Feb. 15, 2016)

Supplementary Readings:

- Tan, P-N., Steinbach, M., Kumar, V., 2006, "Chapter 8. Cluster Analysis: Basic Concepts and Algorithms", in *Introduction to Data Mining*, Pearson, pdf of chapter and slide accessible at http://www-users.cs.umn.edu/~kumar/dmbook/index.php (accessed Feb. 15, 2016)

Week 7: Support Vector Machine

Readings

- Han, J., Kamber, M., and Pei, J., 2011, "Section 9.3 Support Vector Machine", in Chap. 9 Classification: Advanced Methods in *Data Mining: Concepts and Techniques*, Elsevier Science.

- Berwick, R., 2009, "An Idiot's guide to Support vector machines (SVMs)", Notes for Course CAP 6412 (Advanced Computer Vision)
- Udiprod, 2007, "SVM with polynomial kernel visualization", YouTube.com (accessed Feb. 20, 201

Supplementary Readings:

- scikit-learn, 2016, "1.4. Support Vector Machines", scikit-learn.org (accessed Apr. 7, 2016)
- Sehn Korting, T. (video), 2014, "How SVM (Support Vector Machine) algorithm works", YouTube.com (accessed Feb. 20, 2017)

Week 8: Decision Tree Learning and Random Forests

Readings

- Han, J., Kamber, M., and Pei, J., 2011, "Chap. 8 Classification: Basic Concepts" in *Data Mining: Concepts and Techniques*, Elsevier Science.
- Catalano, M., Leise, T., and Pfaff, T., 2009, "Measuring Resource Inequality: The Gini Coefficient", *Numeracy*, 2(2), DOI: http://dx.doi.org/10.5038/1936-4660.2.2.4
- Wang, T., 2011, "Information & Entropy", Slides for Comp 595 DM (accessed Feb. 26, 2017)

Supplementary Readings:

- Khanacademy (video), 2016, "Information Theory", khanacademy.org (accessed Feb. 26, 2017)
- scikit-learn, 2016, "1.10. Decision Trees", scikit-learn.org (accessed Feb. 26, 2017)

Week 9: Neural Networks and Deep Learning

Readings

- Han, J., Kamber, M., and Pei, J., 2011, "Section 9.2 Classification by Backpropagation", in Chap. 9 Classification: Advanced Methods in *Data Mining: Concepts and Techniques*, Elsevier Science.
- Welch Labs, 2014, "Neural Networks Demystified", Part 1 to Part 7, YouTube.com (accessed Feb. 26, 2017)

Supplementary Readings:

- scikit-learn, 2016, "1.17. Neural network models (supervised)", scikit-learn.org (accessed Feb. 7, 2017)
- Shiffman, D., 2012, "Chapter 10. Neural Networks" in *The Nature of Code*, Creative Commons Attribution-NonCommercial 3.0 Unported License, ISBN: 0985930802

Week 10: Network Science

- Install the Python library networkx.

Readings:

- Derrible, S., 2017, "Section 5 Network Science" in Chap. 10 Science of Cities and Machine Learning" in *Urban Engineering for Sustainability*, in progress
- Barabási, A-L., 2014, "Chap. 1 Introduction" in *Network Science*, Creative Commons: CC BY-NC-SA 2.0. PDF V26, 05.09.2014
- 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:

- Barabási, A-L., 2014, "Network Science", Creative Commons: CC BY-NC-SA 2.0. PDF V26, 05.09.2014
- Newman, M., 2010, "Networks: An Introduction", Oxford University Press, Oxford, UK.
- Easley, D., Kleinberg, J., 2010, "Networks, Crowds, and Markets: Reasoning About a Highly Connected World", Cambridge University Press, Cambridge, UK.

Popular Books:

- Barabási, A-L., 2003, "Linked: How Everything Is Connected to Everything Else and What It Means for Business, Science, and Everyday Life", Plume, New York, NY
- Christakis, N., Fowler, J., 2011, "Connected: The Surprising Power of Our Social Networks and How They Shape Our Lives -- How Your Friends' Friends' Friends Affect Everything You Feel, Think, and Do", Back Bay Books, Boston, MA.

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:

- Bird, S., Klein, E., Loper, E., 2010, "Chap. 1 Language Processing and Python" in *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*, O'Reilly Media, Inc, Sebastopol, CA

- Doig, C., 2015, "Introduction to Topic Modeling in Python", PyTexas 2015, available at https://chdoig.github.io/pytexas2015-topic-modeling/#/ (accessed April 2 2017)
- Napitupulu, J., 2015, "Text Learning with scikit-learn", available at https://napitupulu-jon.appspot.com/posts/text-learning-ud120.html (accessed April 2 2017)

Supplementary Readings:

- Zhai, C., 2017, "Text Mining and Analytics", coursera course, available at https://www.coursera.org/learn/text-mining (accessed April 2 2017)

Week 14: Fisher Information

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

Readings

- Ahmad, N., et al., 2016, "Using Fisher information to track stability in multivariate systems", *Royal Society Open Science*, 3:160582

Supplementary Readings:

- Karunanithi, A.T., Cabezas, H., Frieden, B.R., Pawlowski, C.W., 2008, "Detection and assessment of ecosystem regime shifts from fisher information", Ecology and Society, 13(1):22
- González-Mejía, A., Vance, L., Eason, T., Cabezas, H., 2015, "Recent developments in the application of Fisher information to sustainable environmental management", in *Assessing and Measuring Environmental Impact and Sustainability*. Ed. by Klemes, Butterworth-Heinemann

---

Week 15: Agent-Based Modeling

- Install NetLogo (see software section above)

Readings:

- Macal, C., North, M.J., 2006, "Tutorial on Agent-based Modeling and Simulation Part 2: How to Model with Agents", Proceedings of the 38th conference on Winter Simulation, pp 73-83, Monterey, CA, December 03-06
- School of Informatics (U. Edinburgh), 2010, "Cellular Automota and Agent models for ecosystems", Slides for course *Computational Methods for Global Change Research 2009-2010*

Supplementary Readings:

- Shanthi, M., Rajan, E.G., 2012, "Agent Based Cellular Automata: A Novel Approach for Modeling Spatiotemporal Growth Processes", *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 1(3):56-61

- Clarke, K.C., 2014, "Cellular Automata and Agent-Based Models" in *Handbook of Regional Science*. Ed. by Fischer, M.M., and Nijkamp, P., pp.1217-1233, Springer Berlin Heidelberg

---

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
