COMP 7150/8150
Fundamentals of Data Science
Fall 2016

Instructor: Deepak Venugopal

Time and Location: M, W 8:35-10:00am, Fedex Institute of Technology, 227

Email: dvngopal@memphis.edu

Office Hours: M, W 4:00-5:00 PM, Dunn Hall 317

TA: Kazi Iftekar Zaman (kizaman@memphis.edu)

TA Office Hours: MW 4:00 - 5:00 PM, DH 215

Course Textbook: Doing Data Science: Straight Talk from the FrontLine, Cathy O'Neil and Rachel Schutt (O'Reilly)

Other References

- Python for Data Analysis, Wes McKinney (O'Reilly)
- Data Science from Scratch, Joel Grus (O'Reilly)
- Machine Learning, *Tom Mitchell*, McGraw Hill, 1997
- Pattern Recognition and Machine Learning, Chris Bishop, Springer-Verlag, 2006
- Elements of Statistical Learning, Trevor Hastie, Robert Tibshirani and Jerome Friedman, Springer, 2009
- Data Mining: The Textbook, *Charu C Agarwal*, Springer, 2015
- Database Management Systems, *Raghu Ramakrishnan and Johannes Gehrke*

Pre-Requisites

Knowledge of a programming language and descriptive statistics, or equivalent, or permission of instructor.

Learning Objectives

The aim of this course is to introduce students to tools and techniques in all stages of the data science pipeline including, pre-processing, storage, visualization and analysis, in a domain-independent manner. The emphasis will be more on the practical aspects of data science. At the end of this course, students are expected to have the basic skills needed to start working in applications involving data science.

Topics

1. Data Exploration
   - Data cleansing and loading
   - Basic Statistics
   - Visualization
2. Data Storage
   - NoSQL databases
3. Data Processing
   - Dimensionality Reduction
   - Basics of Machine Learning
     - Regression
     - Classification
     - Clustering
     - Evaluation methods
   - Big data processing
     - The MapReduce framework
     - Scalable Machine Learning with Spark
   - Advanced Methods
     - Regularization using Lasso
     - Outlier/Anomaly detection (one-class SVMs)
     - Handling time series data (HMMs)
     - Graph Analytics (Pagerank)
     - Processing streaming data
4. Applications
   - Recommendation Systems
   - Topic Modeling using LDA
   - Sentiment Analysis using Twitter APIs

Evaluation

1. Midterm (date TBD): 20%
2. Homeworks (5): 45% (+1 homework for 8150 students)
3. Project: 35%
   - Groups of 2-4 students
   - You need to first form a team and then submit a proposal of what you intend to work on (kaggle.com is a good source for data). If possible, relating it to your area of research is encouraged.
   - Final submission: paper + poster/demo + code
   - An excellent project is one that is publishable in a good conference/journal.
   - Timeline of project: Team information by 09/02/2016, proposal by 09/30/2016

Grading: $A > 90$, $B > 80$, $C > 65$
Note: $+, -$ grades will be given at the discretion of the instructor. A modified curve may be used for determining the grades at the discretion of the instructor.

Policies

1. No late homeworks or projects will be accepted unless well-documented reasons are presented
2. All homeworks must be individual work. Turnitin will be used to check for plagiarism. Plagiarizing assignments or code sharing is not permitted. If plagiarism or cheating occurs, the student will receive a failing grade on the assignment and (at the instructors discretion) a failing grade in the course. The course instructor may also decide to forward the incident to the University Judicial Affairs Office for further disciplinary action. For further information on U of M code of student conduct and academic discipline procedures, please refer to: http://www.people.memphis.edu/ jaffairs/
3. Regular class attendance is mandatory. There is a strong correlation between regular attendance and obtaining a good grade. Students are responsible for any material and contents of missed lectures.
4. No early or late exams will be given unless under extreme situations.
5. Any grading errors in assignments should be notified within a week to the TA.