Dates: March 10 – May 21, 2014
Times: Monday & Wednesday 6:30 – 9:30 pm
Location: 580 Howard Street (GA Lofts)
Instructor: Rob Hall – robhall.ga@gmail.com
Experts-in-Residence (TA's):
- Francesco Mosconi – f@mosconi.me
- Dow Street – dowstreet@prodigy.net
Office Hours: Wednesdays after class sessions with Francesco. Additional office hours to be scheduled. To schedule a personal appointment, please email the instructor team.

COURSE DESCRIPTION

This course is a practical approach to the knowledge and skills required to excel in the field of data science. Through various case studies, real-world examples and guest speakers, students will be exposed to the basics of data science, fundamental modeling techniques, and various other tools to make predictions and decisions about data. Students will gain practical computational experience by running machine learning algorithms and learning how to choose the best and most representative data models to make predictions. Students will be using Python throughout this course.

COURSE MATERIALS

Students are required to bring a laptop to class everyday. Please come to the first class with Continuum Anaconda (http://continuum.io/downloads) installed, as detailed in the Pre-Work document.

COMPLETION REQUIREMENTS

In order to receive a General Assembly Certificate in Data Science, upon completion of the course, students must:

- Complete and submit 80% of all course assignments (homework, labs, quizzes). Students will receive feedback from instructors on their assignments on a timely basis. Students who miss more than 20% of assignments will not be eligible for the course certificate.
- Complete and submit the course final project, earning a satisfactory grade by completing all functional and technical requirements on the project rubric, including delivering a presentation.

Assignments, milestones and feedback throughout the course are designed to prepare students to deliver a quality course project.

COURSE OUTLINE

The weekly schedules for lecture content, lab content, and homework assignments are subject to change according to the needs & desires of the class. Dates for guest speakers may change according to speaker availability.

UNIT 1: DATA SCIENCE OVERVIEW / THE BASICS

LESSON 1: INTRODUCTION TO DATA SCIENCE (3/10)

- Overview of data science
- Describe the data mining workflow and the key traits of a successful data scientist
- Review of basic UNIX command-line tools
- Lab: Introduction to the iPython Notebook and the command line interface

LESSON 2: PYTHON & VERSION CONTROL W/ GIT (3/12)

- Introduce Python and its usefulness for data analysis tasks
- Introduce Git, Github, version control workflow
- Lab: Numpy, array slicing, & intro to Pandas

HW & Project Milestones

HW1 Assigned (Due 3/17 via Schoology)

LESSON 3: WORKING WITH SEMI-STRUCTURED DATA (3/17)

- Lab: Install Git and setup class Github repos (carryover from Lesson 2)
- Introduce web APIs, REST, JSON
- Access data from REST APIs and parse it using Python & JSON
- Lab: Accessing web APIs (Github, Twitter) and parsing the response data

UNIT 2: MACHINE LEARNING FUNDAMENTALS & WORKING WITH DATA

LESSON 4: INTRODUCTION TO MACHINE LEARNING & KNN CLASSIFICATION (3/19)

- Explain the concepts and applications of supervised & unsupervised learning techniques
- Describe categorical and continuous feature spaces, including examples and techniques for each
- Discuss the purpose of machine learning and the interpretation of predictive modeling results

HW & Project Milestones

HW1 Due via Schoology

HW2 Assigned (Due by 11:59PM Sunday, March 30)

- Understand the kNN classification algorithm, its intuition and implementation.
- Minimize prediction error using training & test sets. Optimize predictive performance using cross-validation.
- Lab: Visualization with matplotlib & Implementing kNN classification using scikit-learn

LESSON 5: REGRESSION AND REGULARIZATION (3/24)

- Explain the concepts of regression models, including their assumptions and applications
- Discuss the motivation for regularization techniques and their use
- Implement a regularized fit
- Lab: Regression using statsmodels & Pandas, Regularization using sklearn

LESSON 6: DIMENSIONALITY REDUCTION (3/26)

- Problems with high dimensional data
- Application of dimensionality reduction techniques.
- Principal component analysis to explore high dimensional data

HW & Project Milestones

HW3 assigned: Final Project Elevator Pitch (Due in class Wed. 4/2)

HW2: KNN CLASSIFICATION - DUE SUNDAY 3/30 BY 11:59PM VIA GITHUB

LESSON 7: LOGISTIC REGRESSION (3/31)

- Introduce the concepts of logistic regression and its relation to other regression models
- Describe the applications of logistic regression to classification problems and probability estimation
- ROC curves for evaluating binary classifiers
- Lab: Implementing logistic regression using sklearn

HW & Project Milestones

HW4 / "Midterm" Assigned: Logistic Regression (Due Sunday 4/13 by 11:59PM)

NOTE: HW4 will be graded on a 0-100 scale & is required to receive a Letter of Completion in the course.

LESSON 8: DATABASE TECHNOLOGIES, STRUCTURED DATA, & INTRO TO STRUCTURED QUERY LANGUAGE (SQL) (4/2)

- Introduce relational theory and the benefits and limitations of a normalized database
- Compare SQL to NoSQL databases
- Lab: Build a relational database from raw data using SQL, execute SQL statements from within Python

HW & Project Milestones

HW3: Final Project Elevator Pitch due & presented to the group in class

HW5: Final Project Proposal Assigned (Due 4/16 by Noon)

HW3: FINAL PROJECT ELEVATOR PITCH – DUE WEDNESDAY 4/2 IN CLASS

LESSON 9: UNSUPERVISED CLUSTERING WITH K-MEANS (4/7)

- Clustering as a form of data exploration
- The importance of the distance function and scale normalization in cluster formation
- Implement a k-means clustering algorithm

LESSON 10: UNSTRUCTURED DATA, MAP-REDUCE, & TEXT MINING (4/9)

- Describe the concepts of parallel computing and applications to problems in big data
- Introduce the map-reduce framework and popular implementations including Hadoop
- Lab: Implementing example map-reduce task
- Intro to text mining (if time allows)

HW4 / MIDTERM: LOGISTIC REGRESSION - DUE SUNDAY 4/13 BY 11:59PM VIA GITHUB

UNIT 3: MORE ADVANCED ML TECHNIQUES

LESSON 11: RECOMMENDATION SYSTEMS (4/14)

- Explain the use of recommendation systems, and discuss several familiar examples
- Understand the underlying concepts, including collaborative & content-based filtering
- Implement a recommendation system

HW5: FINAL PROJECT "FORMAL" PROPOSAL - DUE WED 4/16 BY 12:00 NOON

- Proposals to include: data source, problem to be solved, modeling tools, etc.

LESSON 12: DECISION TREES AND RANDOM FORESTS (4/16)

- Describe the use of decision trees for classification tasks
- Create a random forest model for ensemble classification
- Lab: Decision trees and random forests in scikit-learn
- Demo of commercial tool for decision trees (if time allows)

HW & Project Milestones

HW5: Final Project Proposal due by 12:00PM (NOON) on Wed. 4/16

HW6: Project Milestone assigned (Due 4/30 by Noon)

LESSON 13: NON-LINEAR CLASSIFICATION TECHNIQUES & SUPPORT VECTOR MACHINES (4/21)

- Describe the motivation for non-linear classification techniques, as well as the conceptual basis for their use
- Understand the advantages and disadvantages of black box models
- Implement a non-linear classifier & compare results with linear classification

UNIT 4: OFF THE LAPTOP, INTO THE WORLD

There is a wide range of material we could cover in this unit. We may change the topics depending on the progress and interests of the class.

LESSON 14: SCALING UP – PARALLEL & DISTRIBUTED COMPUTING (TENTATIVE) (4/23)

- Scaling up may include iPython.parallel, StarCluster, Hadoop, and/or a demo of a commercially available data science platform-as-a-service. Exact content TBD.

LESSON 15: PUTTING YOUR MODEL INTO PRODUCTION (TENTATIVE) (4/28)

- Once you have done your analysis and trained your model, it's time to put it into production. Subtopics may include working on an AWS node, using emacs, etc.

HW6: FINAL PROJECT MILESTONE - DUE WED 4/30 BY 12:00 NOON

- Github repo live, including README, pointer to dataset to be used, and at least one visualization
- HW7 Assigned: Homework 7 will be to provide peer feedback on the project milestone in your small groups (to be assigned). That feedback will be due Wed 5/7.

LESSON 16: GUEST SPEAKER – MATT SUNDQUIST, COO PLOT.LY (4/30)

- Plot.ly provides a very cool set of APIs for visualization. We will install and do a quick ramp-up with plot.ly before Matt comes to speak.

LESSON 17: FURTHER EXPLORATIONS (5/5)

- Specific topic(s) to be selected closer to the date. Potential topics include Bayesian classification, ensemble techniques, grid search & parameter selection, social network analysis, and others.

UNIT 6: FINAL PROJECT & ADD'L GUEST SPEAKERS

HW7: FINAL PROJECT SMALL-GROUP PEER FEEDBACK - DUE WED 5/7 BY 12:00 NOON

- Github repo live, including README, pointer to dataset to be used, and at least one visualization

LESSON 18: GUEST SPEAKER / FINAL PROJECT WORKING SESSION (5/7)

HW & Project Milestones

We will break into small groups and review peer feedback in class

LESSON 19: GUEST SPEAKER / FINAL PROJECT WORKING SESSION (5/12)

LESSON 20: FINAL PROJECT WORKING SESSION / WHERE TO GO NEXT (5/14)

- Review of concepts and examples from preceding weeks
- Discussion of resources & tools for further study

LESSON 21: FINAL PROJECT PRESENTATIONS (5/19)

LESSON 22: FINAL PROJECT PRESENTATIONS (5/21)