Dates: May 31st – Aug 16, 2014
Times: Saturday 10:00am – 5:00pm
Location: 580 Howard st.
Instructor: Mike Tamir – mntamir@gmail.com
Experts-in-Residence (TA's):
- Frank Taylor - franctaylor.ft@gmail.com
- Zack DeSario - zacharydesario@gmail.com

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.

Schoololgy: https://www.schoology.com/course/113605631/materials
GitHub: https://github.com/mike-tamir/GA_DAT7

COMPLETION REQUIREMENTS

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

- Complete, submit, and pass 80% of all course assignments. 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 FAIR to EXCELLENT 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.

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 1.1: INTRODUCTION TO DATA SCIENCE (5/31)

- Overview of data science
- Describe the data mining work flow 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 1.2: PYTHON & VERSION CONTROL W/ GIT (5/31)

- 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 6/7 and submitted via Schoology)

LESSON 2.1: WORKING WITH SEMI-STRUCTURED DATA (6/7)

- Lab: Install Git and setup class Git hub repos (carry over from Lesson 2)
- Linear Algebra
- 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 2.2: INTRO. TO MACHINE LEARNING & KNN CLASSIFICATION (6/7)

- 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
- 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

HW & Project Milestones

HW1 Due via Schoology

HW2 Assigned (Due by 11.59 PM Sat 6/21 and submitted via Schoology)

LESSON 3.1: REGRESSION AND REGULARIZATION (6/14)

- 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 3.2: DIMENSIONALITY REDUCTION (6/14) • Problems with high dimensional data "curse of dimensionality" and applications • Principal component analysis for high dimensional data • Lab: Dimensionality reduction and PCA	**HW & Project Milestones** HW3 Assigned: Final Project Elevator Pitch (Due in class Sat 6/28 and submitted via Schoology)

HW 2: KNN CLASSIFICATION - DUE SATURDAY 6/21 BY 11:59PM VIA SCHOOLOGY

LESSON 4.1: LOGISTIC REGRESSION (6/21) • 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 Sat by 11:59PM 7/12 and submitted via Schoology) *NOTE: HW4 will be graded on a 0-100 scale & is required to receive a Letter of Completion in the course.*
LESSON 4.2: DATABASE TECHNOLOGIES, STRUCTURED DATA, & INTRO TO STRUCTURED QUERY LANGUAGE (SQL) (6/21) • 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** HW5: Final Project Proposal Assigned (Due 7/12 in class and submitted via Schoology)

HW 3: FINAL PROJECT ELEVATOR PITCH – DUE & PRESENTED TO GROUP IN CLASS 6/28

LESSON 5.1: UNSUPERVISED CLUSTERING WITH K-MEANS (6/28)

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

LESSON 5.2: DECISION TREES AND RANDOM FORESTS (6/28)

- 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

*Note, unit 3 and unit 4 schedule subject to alteration based on guest speaker availability.

UNIT 3: MORE ADVANCED ML TECNIQUES*

HW 4/MIDTERM: LOGISTIC REGRESSION – DUE SATURDAY 7/12 BY 11:59PM
HW 5: FINAL PROJECT "FORMAL" PROPOSAL – DUE SATURDAY 7/12 IN CLASS

LESSON 6.1: TEXT MINING & NATURAL LANGUAGE PROCESSING (7/12)

- Vectorizing text for data mining. Normalizing vectors. Term frequency – inverse document frequency (TF-IDF).
- Lab: A deeper look at natural language processing (NLP) with nltk (if time allows)

LESSON 6.2: NON-LINEAR CLASSIFICATION TECHNIQUES & SUPPORT VECTOR MACHINES (7/12)

- Describe the motivation for non-linear classification techniques, as well as the conceptual basis for their use
- Implement a non-linear classifier & compare results with linear classification

HW & Project Milestones

HW6: Project Milestone Assigned (Due 7/19 in class and submitted via Schoology)

HW6.1 install/setup plot.ly for gust lecture on 7/19: http://goo.gl/ryVuJl

HW6: FINAL PROJECT MILESTONE – DUE SATURDAY 7/19 IN CLASS

- Github repo live, include README, pointer to data set to be used, and at least one visualization

LESSON 7.1: RECOMMENDER SYSTEMS (7/19)

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

HW & Project Milestones

HW7: Provide Peer Feedback on the Project Milestone Assigned (Due 7/26 in class).

LESSON 7.2: GUEST LECTURES (7/19)

- Mat Sundquist – Co-Founder and COO at Plot.ly

UNIT 4: GUEST SPEAKERS & FINAL PROJECTS*

HW7: FINAL PROJECT SMALL-GROUP PEER FEEDBACK - DUE WED 7/26
We will split into teams of 2-3 people, review each other's final projects and progress to date, and provide peer feedback.

LESSON 8: GUEST LECTURES/Project Feedback (7/26)

- Nick Elprin – Co-founder Domino labs
- We will split into teams of 2-3 people, review each other's final projects and progress to date, and provide peer feedback.

LESSON 9.1: GUEST LECTURE (8/2)

- Michael Rinehart - Principal Scientist at Elastica

LESSON 9.2: MAP-REDUCE (8/2)

- 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

LESSON 10: FINAL PROJECTS WORKIG SESSION / WHERE TO GO NEXT 8/9

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

LESSON 11: FINAL PROJECT PRESENTATIONS (8/16)

Homework 8 – submission of final project paper/iPython Notebook due via Schoology