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 both R and Python throughout this course.

Prerequisites:

- Some experience with programming languages (preferably R or Python) and familiarity with the command line interface (UNIX).
- Laptop with OSX (Mac) or UNIX/Linux operating system

Note to students:
Please come to the first class with R (http://cran.r-project.org/) and the ggplot2 plotting library (http://ggplot2.org/) installed on your machine. We will also use Python later in the course.

Each class will consist of a lecture and a hands-on data analysis session using UNIX, R, and/or Python.

GRADING

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 within 2 - 3 days. 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 MATERIALS/REQUIREMENTS

Students are required to bring a laptop to class everyday.

COURSE TOPICS/LEARNING OBJECTIVES

UNIT 1: THE BASICS

LESSON 1: INTRODUCTION TO DATA EXPLORATION

- Describe the data mining workflow and the key traits of a successful data scientist.
- Extract, format, and preprocess data using UNIX command-line tools.
- Explore & visualize data using R and ggplot2.

LESSON 2: INTRODUCTION TO MACHINE LEARNING

- 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.

UNIT 2: FUNDAMENTAL MODELING TECHNIQUES

LESSON 3: K-NEAREST NEIGHBORS CLASSIFICATION

- Describe the setting and goal of a classification task.
- Minimize prediction error using training & test sets, optimize predictive performance using cross-validation.
- Understand the kNN classification algorithm, its intuition and implementation.
- Implement the "hello world" of machine learning (kNN classification of iris dataset).

LESSON 4: NAIVE BAYES CLASSIFICATION

- Outline the basic principles of probability, including conditional probability and Bayes' theorem.
- Describe inference in the Bayesian setting, including the prior and posterior distributions and the likelihood function.
- Understand the naive Bayes classifier and its assumptions.
- Implement a spam filter using the naive Bayes technique.

LESSON 5: REGRESSION AND REGULARIZATION

- Explain the concepts of regression models, including their assumptions and applications.
- Discuss the motivation for regularization techniques and their use.
- Implement a regularized fit.

LESSON 6: LOGISTIC REGRESSION

- Describe the applications of logistic regression to classification problems and probability estimation.
- Introduce the concepts underlying logistic regression, including its relation to other regression models.
- Predict the probability of a user action on a website using logistic regression.

LESSON 7: K-MEANS CLUSTERING

- Explain the purpose of exploratory data analysis, its applications in continuous and categorical feature spaces, and the interpretation and use of clustering results.
- Discuss the importance of the distance function in cluster formation, as well as the importance of scale normalization.
- Implement a k-means clustering algorithm.

UNIT 3: FURTHERING MODELING TECHNIQUES

LESSON 8: MACHINE LEARNING IN PYTHON

- Introduce Python and its usefulness for data analysis tasks.
- Experiment with scikit-learn, a general-purpose machine learning library for Python.

LESSON 9: ENSEMBLE TECHNIQUES

- Describe general ensemble techniques such as bagging and boosting.
- Build an enhanced classification algorithm using AdaBoost.

LESSON 10: DECISION TREES AND RANDOM FORESTS

- Describe the use and construction of decision trees for classification tasks.
- Create a random forest model for ensemble classification.

LESSON 11: SUPPORT VECTOR MACHINES

- 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.

LESSON 12: DIMENSIONALITY REDUCTION

- Explain the practical and conceptual difficulties in working with very high-dimensional data.
- Understand the application and use of dimensionality reduction techniques.
- Draw inferences from high-dimensional datasets using principal components analysis.

LESSON 13: RECOMMENDATION SYSTEMS

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

UNIT 4: OTHER TOOLS

LESSON 14: DATABASE TECHNOLOGIES

- Introduce concepts and use of relational databases, alternative database technologies such as NoSQL, and popular examples of each.
- Interact with a relational db (MySQL) and a NoSQL db (MongoDB), compare and contrast their use.

LESSON 15: SOCIAL NETWORK ANALYSIS

- Describe the use of graphs and graph theory to analyze problems in network analysis.
- Explore network visualization with Gephi.

LESSON 16: MAP-REDUCE

- Describe the concepts of parallel computing and applications to problems in big data.
- Introduce the map-reduce framework and popular implementations including Hadoop.
- Implement and explore examples of map-reduce tasks.

LESSON 17: FINAL PROJECT WORKING SESSION
LESSON 18: FINAL PROJECT WORKING SESSION
LESSON 19: WHERE TO GO NEXT • Review of concepts and examples from preceding weeks. • Discussion of resources & tools for further study.
LESSON 20: FINAL PROJECT PRESENTATIONS