Introduction to Data Science

COM SCI X450.1

4.0 Credits

Spring 2019

Class Meeting Information

- Start date: 4/2/2019
- End date: 6/11/2019
- UCLA Extension Lindbrook Center in Westwood
- 10 meetings, 3.5 hours per meeting

Instructor Information

Name:	Daniel D. Gutierrez
Phone:	(310) 231-7900
Email:	daniel@amuletanalytics.com
Website:	www.amuletanalytics.com
Twitter:	@AMULETAnalytics
GitHub:	https://github.com/AMULETAnalytics/UCLAIntroDataScience

Students should feel free to contact me for the duration of the course in order to get answers to questions about the course materials including – lectures and coursework. E-mail contacts are preferred.

My name is Daniel D. Gutierrez. I am a practicing data science consultant specializing in data science, machine learning, AI, and deep learning using R and Python. In addition, I am a tech journalist serving as Managing Editor for insideBIGDATA.com. I have taught at UCLA Extension for over 15 years as an instructor (courses and seminars). I hold a BS degree in Mathematics/Computer Science from UCLA.

Course Description

With the unprecedented rate at which data is being collected today in almost all fields of human endeavor, there is an emerging economic and scientific need to extract useful information from it. Data science is the process of making predictions and classifications, plus the automatic discovery of patterns, clusters, associations and anomalies in massive data sets. Data science is a highly inter-disciplinary field representing a confluence of disciplines including computer science, mathematical statistics, probability theory, machine learning algorithms, data analysis, data visualization, and database/data warehouse/data lake systems.

The open source R statistical environment is the choice for a growing number of data scientists worldwide for data analysis and machine learning. In order to gain a foothold in the growing field of data science a sound knowledge of R programming is valuable.

Introduction to Data Science is a comprehensive introduction to the field of data science and the R statistical environment and programming language. Students already may be skilled programmers in other languages but complete novices to R programming, or possess some knowledge of R but no programming. The common objective is to write R code for diverse problem domains and applications. No mathematical or statistical knowledge is necessary. The course covers programming-related skills that are not specific to any one problem domain. The course makes use of many extended examples of R programs for different purposes. RStudio, a popular open source Integrated Development Environment (IDE) for R programming, is utilized in the course.

Prerequisites — Classes or Knowledge Required Before Taking This Course

There are no UCLA Extension prerequisites for this class. Some knowledge of computer programming, statistics and computer science, however, will be helpful in mastering the subject matter in this course.

Course Objectives

- Employ and document use of "the data science process."
- Demonstrate how to succinctly state goals for a data science project
- Use the R programming language to perform exploratory data analysis and data visualization
- Demonstrate the data acquisition process
- Employ basic data munging techniques
- Use R programming language to demonstrate an understanding of supervised and unsupervised learning algorithms in the context of various business applications

Course Material

Required Textbook:
Machine Learning and Data Science: An Introduction to Statistical Learning Methods with R, Daniel D. Gutierrez, Technics Publications, 2015, ISBN-9781634620963

Optional Textbook:
The Art of R Programming: A Tour of Statistical Software Design, Norman Matloff, No Starch Press, 20011, ISBN-10: 1-59327-384-3

Course Outline

Week 1	**Topics/Objectives:** **R LANGUAGE PART 1**	**Key Topics:** 1. The Data Science Process 2. Brief history of R 3. Installing, configuring, and using R and RStudio 4. R scripts using basic language constructs and data types 5. Atomic classes in R 6. Assignment statements 7. Useful R objects: vectors, lists, matrices, factors, data frames, arrays 8. Creating sequences 9. Object attributes: names, dimensionality 10. Commenting your R code 11. Coercion 12. Na, NaN, and NULL
	Reading	Read Gutierrez Chapter 1 Read Matloff Chapters 1 - 4
	Assignments	Hand out quiz #1 (not graded)

Week 2	**Topics/Objectives:** **R LANGUAGE PART 2**	**Key Topics:** 1. Learning R language constructs for manipulating data 2. These techniques will be useful for data munging 3. Extracting parts of vectors, matrices, lists (subsetting) 4. Managing NA values found in data sets
	Reading	Read Matloff Chapters 5 and 6
	Assignments	Hand out homework assignment #1 Hand out quiz #2 (not graded)

Week 3	**Topics/Objectives:** **R LANGUAGE PART 3**	**Key Topics:** 1. Vectorized operations 2. If control structure 3. Logical expressions 4. For control structure, nested loops 5. While control structure 6. Repeat control structure 7. Defining and using functions in R, recursion, argument passing 8. `lapply()` loop function 9. `sapply()` loop function

	Reading	Read Matloff Chapter 7 and 10
	Assignments	Hand out quiz #3 (not graded)

Week 4	**Topics/Objectives:** **R LANGUAGE PART 4**	**Key Topics:** 1. `lapply()` loop function 2. `tapply()` loop function 3. `split()` function 4. `mapply()` loop function 5. Generating random numbers – normal, Poisson, binomial 6. `sample()` function for random sampling 7. Dates and times in R
	Reading	
	Assignments	Hand out quiz #4 (not graded)

Week 5	**Topics/Objectives:** **DATA ACCESS**	**Key Topics:** 1. Accessing data sources 2. Downloading files from the web 3. Comma separated value (CSV) 4. Excel 5. JSON 6. Web page scraping 7. SQL databases 8. SQL equivalents in R 9. Writing data
	Reading	Read Gutierrez Chapter 2
	Assignments	Hand out homework assignment #2 Hand out quiz #5 (not graded)

Week 6	**Topics/Objectives:** **DATA MUNGING**	**Key Topics:** 1. Examine the process of data munging, aka data wrangling, aka data transformation 2. Present a variety of commonly used techniques to add to your data science toolbox 3. You'll be able to draw upon these methods for future projects
	Reading	Read Gutierrez Chapter 3 Read Matloff Chapter 12
	Assignments	Hand out quiz #6 (not graded)

Week 7	**Topics/Objectives: EXPLORATOR DATA ANALYSIS (EDA)**	**Key Topics:** 1. Use numeric EDA for knowledge discovery and statistical analysis 2. Perform simple data analysis 3. Use basic R statistical functions 4. Explore levels of factor variables (categorical) 5. Find number of non-missing values 6. Independent study: common statistical tests for continuous random variables, and discrete data (categorical)
	Reading	Read Gutierrez Chapter 4 Read Matloff Chapter 8
	Assignments	Hand out class project Hand out quiz #7 (not graded)

Week 8	**Topics/Objectives: DATA VISUALIZATION**	**Key Topics:** 1. Learn to use `hist()`, `boxplot()`, `barplot()`, density plots, scatterplots with `plot()`, `qqplot()`, heatmaps with `image()` 2. Explore big data visualization techniques: random sample, `smoothscatter()`, count bins with `hexbin()` and `plot()` 3. Techniques for additional variables: color, size of data point, plot symbols 4. Missing value plots 5. Correlation plots with `pairs()` 6. Expository plots with axis labels, legends, titles, multiple panels 7. Create plot PDF and image files
	Reading	Read Gutierrez Chapter 4
	Assignments	Hand out homework assignment #3 Hand out quiz #8 (not graded)

Week 9	**Topics/Objectives: SUPERVISED MACHINE LEARNING**	**Key Topics:** 1. Be able to employ a supervised statistical learning technique – linear regression 2. Using the `lm()` algorithm in R 3. Training the model 4. Making predictions using the trained model 5. Explore the use of both single and multiple linear regression

	Reading	Read Gutierrez Chapter 5
	Assignments	Hand out quiz #9 (not graded)

Week 10	**Topics/Objectives: UNSUPERVISED MACHINE LEARNING**	**Key Topics:** 1. Overview unsupervised learning methods 2. Manually step through process yielding distinct clusters showing groupings and similarities in the data 3. Review the hierarchical clustering algorithm using R's `hclust()` function to compute clusters and use data viz to display 4. Review the K-means clustering algorithm using R's `kmeans()` function to compute clusters and use data viz to display
	Reading	Read Gutierrez Chapter 8
	Assignments	Hand out quiz #10 (not graded)

Evaluation and Grading

Evaluation of Student Performance Weighted as Percentages of the Total Grade

Ungraded practice quizzes (10)	
Homework assignments (3)	60%
Class project	40%
	100%

Grading Scale

A = 90% – 100%
B = 80% – 89%
C = 70% – 79%
D = 60% – 69%
F = 59% or less