The George Washington University

EMSE 6992 – Data Analytics Introduction and Practicum

Course Syllabus

Course and Contact Information

Course: EMSE, Data Analytics Introduction and Practicum, 6992, 13
Semester: Fall, 2017
Meeting time: Thursdays, 6:10 – 8:40 PM
Location: TOMP, 402

Instructor

Name: Benjamin S. Harvey, Ph.D.
Campus Address: Science & Engineering Hall (SEH) 1800
Phone: (904) 662-6611
E-mail: bsharve@nsa.gov (cc: bsharve@gmail.com)
Office hours: before class (5:00 PM – 6:00PM) and by appointment

EMSE 6992. Data Analytics Introduction and Practicum. 3 Credits.

Selected topics in engineering management and systems engineering, as arranged. May be repeated for credit. Basic techniques of data science; algorithms for data mining; basics of statistical modeling and their "Big Data" applications. Concepts, abstractions, and practical techniques.

Prerequisites

None.

Required Text(s)

- Erl, Thomas, Wajid Khattak, and Paul Buhler. *Big data fundamentals: concepts, drivers & techniques*. Prentice Hall Press, 2016. Available for free as a PDF download at here (https://www.slideshare.net/AshishSharma118/big-data-fundamentals-thomas-erl?qid=9d96f0ea-9180-421f-99f3-1c8f7fd8a8f6&v=&b=&from_search=2)
- Schutt, Rachel, and Cathy O'Neil. *Doing data science: Straight talk from the frontline*. " O'Reilly Media, Inc.", 2013.

Optional Text(s)

- Conway, Drew, and John White. *Machine learning for hackers*. " O'Reilly Media, Inc.", 2012.
- McKinney, Wes. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.
- Provost, Foster, and Tom Fawcett. *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc.", 2013.
- Stanton, Jeffrey M. "*Introduction to data science*." (2013). Available for free in the iTunes bookstore or as a PDF download at http://surface.syr.edu/istpub/165/

Learning Outcomes

Upon successful completion of this course, students should have developed some or all of the following areas of skills and knowledge:

- Describe what Data Science is and the tools / skill sets needed to be a data scientist.
- Explain in basic terms what Statistical Inference means. Identify probability distributions commonly used as foundations for statistical modeling. Fit a model to “Big Data”.
- Use R/Python to carry out basic statistical modeling and analysis.
- Explain the significance of exploratory data analysis (EDA) in “Big Data” exploration.
- Apply basic tools (plots, graphs, summary statistics) to carry out EDA.
- Describe the Data Science Process and how its components interact.
- Use APIs and other tools to scrap the Web and collect data.
- Apply EDA and the Data Science process to assignments / case studies. Establish a data science toolkit and create a portfolio for their work.
- An understanding of the nature of the data collection, the data itself, and the analysis processes that relate to the kinds of inferences that can be drawn.
- Understand the limitations of data sets based on their size, contents, and provenance.
- Knowledge of data organization, management, preservation, and reuse.
- Knowledge of what statistical analysis techniques to choose, given particular demands of inference and available data.
- Knowledge of general linear algebra, linear models and classification / clustering analysis methods for statistical analysis.
- Skills and knowledge in preparing data for analysis, including cleaning data, manipulating data, and dealing with missing data
- Skills in analyzing open source “Big Data” sets using open source data analysis tools
- Skills in scripting for data manipulation, analysis, and visualization using R, Python, and a variety of add on packages.

Class Schedule

Week	Date	Topic(s)	Readings (Erl and Schutt)	Speaker	Assignment(s) Due
1	8/31	**Course Introduction**			
2	9/7	**Understanding Big Data** Fundamental Terminology and Concepts, Big Data Motivation and Drivers	Ch 1 & 2 (Erl)	Guest Speaker: Kato Mivule – Data Scientist Department of Defense	
3	9/14	**Big Data Adoption, Planning and Business Intelligence** Big Data Adoption and Planning Considerations, Enterprise Technologies	Ch 3 & 4 (Erl)	Guest Speaker: Dong Hyun Jeong – UDC Associate Professor of Computer	Research Proposal Instructions will be handed out.

		and Big Data Business Intelligence		Science	
4	9/21	**Big Data Storage Concepts** Characteristics of Data in Big Data Environments, Dataset Types in Big Data Environments	Ch 5 (Erl)	Guest Speaker: Michael Monson -- VP integrated Data Services	Assignment #1: Intro to R/Python, GitHub, and developing a Portfolio
5	9/28	Guest Lecturer: Dr. David Broniatowski		Guest Speaker: Laura Wrubel, GW Libraries	Students Research Proposals Due: (9/28)
6	10/5	**Big Data Processing** Large-Scale, Parallel, and Distributed Data Processing, Hadoop	Ch 6 (Erl)	Guest Speaker: Rodney Wallace -- IBM	
7	10/12	**Big Data Storage Technology** Big Data Storage Technologies and Data Engineering **Portfolio: Part 1** Python and GitHub Installation and Basics, Data Science Toolkit, Visualization, Interpreting, and Communicating Results	Ch. 7 (Erl) Ch 14 (Schutt)	Guest Speaker: Howie Huang -- GW Computer and Electrical Engineering	Assignment #2: Applying Statistical Inference to Large Datasets Assignment #1: Due 10/12
8	10/19	**Data Analysis I: Big Data Fundamental Analysis and Analytics** Intro to Machine Learning, Statistical Analysis, and Visual Analysis	Ch. 8 (Erl)	Guest Speaker: Hanan Aldarmaki (Mona Diab's student) -- GW NLP lab	
9	10/26	**Data Analysis II: Inferential Statistics and Exploratory Data Analysis (EDA)** Essential Statistics, Statistical Analysis, and Statistical Measures	Ch. 2 & 7 (Schutt).	Guest Speaker: Jan Neumann -- COMCAST	
10	11/2	**Machine Learning and Statistics: Algorithms and Regression** Differentiating algorithmic and model based frameworks, Regression	Ch. 3 & 5 (Schutt).	Guest Speaker: Shaun Gittens, Ph.D. Principal Data Scientist Applied Technology Group, LLC.	Assignment #3: Applying Machine Learning Techniques to Large Datasets Assignment #2: Due 11/2
11	11/9	**Linear Algebra** Introduction to Vectors and Matrices, Solving Linear Equations, Vector Spaces and Subspaces, Orthogonality, Determinants, Eigenvalues and Eigenvectors	Materials will be provided.	Guest Speaker: Abe Usher -- CTO DigitalGlobe	
12	11/16	**Machine Learning: I**	Ch. 4 & 6 (Schutt)	Guest Speaker: Guest Speaker:	Assignment #4: Data Science Process, "Big

		Supervised Learning Techniques, Naïve Bayes, k-NN		Tim Wood -- GW Cloud Computing Lab	Data", Visualization, Interpreting, and Communicating Results in your Portfolio Assignment #3: Due 11/16
13	11/23	**Thanksgiving Holiday (No class)**			
14	11/30	**Visualization, Interpreting, and Communicating Results: Portfolio Part 2** Visualization, Data Summaries, Model Checking & Comparison, Visual Data Analytics, Visualizing / Interpreting / Communicating Results in a Portfolio	Ch. 9 (Schutt)	Speaker: Student Portfolio Presentation	
15	12/7	**Machine Learning: II** Unsupervised Learning Techniques, Challenges for "Big Data" analytics	Ch. 8, 10-13 (Schutt)	Speaker: Student Research Presentation	Assignment #4: Due 11/30 Student Final Research Papers Due: 12/7

Assignments and Grades

Grading

This course consists of an individual portfolio project and a final exam. The portfolio project consists of building a portfolio and Data Science toolkit in GitHub that you can continuously use throughout you Data Science careers.

- Portfolio Project - Part I 5%
- Portfolio Project - Part II 10%
- Portfolio Project - Part III 15%
- Portfolio Project - Part IV 20%
- Final Research Project – Part I 25%
- Final Research Project – Part II 25%

Assignments

This course consists of four portfolio assignments, and a final research project. There will be a total of 500 points: Portfolio project (250) and Final Research Project (250). Due dates for assignments can also be seen below:

Assignment	Description	Total Points	Due Date
Portfolio Project - Part I	Assignment 1 – Creating a Portfolio: Intro to GitHub, Python, and EDA	25	10/5
Portfolio Project - Part II	Assignment 2 - Statistical Inference	50	10/26
Portfolio Project - Part III	Assignment 3 - Machine Learning	75	11/16
Portfolio Project - Part IV	Assignment 4 - Data Science Process, “Big Data”, Visualization, Interpreting, and Communicating Results in your Portfolio	100	11/30
Final Project – Part I	Research Proposal and Final Paper (1-5 pages)	125	12/7
Final Project – Part II	Research Presentation	125	11/30 or 12/7
	Total Possible Points	**500**	

University Policies

University Policy on Religious Holidays [should be included verbatim]

1. Students should notify faculty during the first week of the semester of their intention to be absent from class on their day(s) of religious observance.
2. Faculty should extend to these students the courtesy of absence without penalty on such occasions, including permission to make up examinations.
3. Faculty who intend to observe a religious holiday should arrange at the beginning of the semester to reschedule missed classes or to make other provisions for their course-related activities

Support for Students Outside the Classroom [should be included verbatim]

Disability Support Services (DSS)
Any student who may need an accommodation based on the potential impact of a disability should contact the Disability Support Services office at 202-994-8250 in the Rome Hall, Suite 102, to establish eligibility and to coordinate reasonable accommodations. For additional information please refer to: gwired.gwu.edu/dss/

Mental Health Services 202-994-5300
The University's Mental Health Services offers 24/7 assistance and referral to address students' personal, social, career, and study skills problems. Services for students include: crisis and emergency mental health consultations confidential assessment, counseling services (individual and small group), and referrals. counselingcenter.gwu.edu/

Academic Integrity Code [NOTE: reference to the code should be made and the url provided]

Academic dishonesty is defined as cheating of any kind, including misrepresenting one's own work, taking credit for the work of others without crediting them and without appropriate authorization, and the fabrication of information. For the remainder of the code, see: studentconduct.gwu.edu/code-academic-integrity