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, 2018
Meeting time: Thursdays, 6:10 – 8:40 PM
Location: TBD

Instructor

Name: Benjamin S. Harvey, Ph.D.
Campus Address: Science & Engineering Hall (SEH) 1800
Phone: (904) 662-6611
E-mail: bsharve@gmail.com (cc: bsharve@gwu.edu)
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)

- Field Cady. *The Data Science Handbook*. Wiley, 2017. Available for free as a PDF download here (https://universalflowuniversity.com/Books/Computer%20Programming/Data%20Mining%20and%20Data%20Science/The%20Data%20Science%20Handbook.pdf)
- 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 successful 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) and Lab	Readings (Cady)	Speaker	Assignment(s) Due
1	8/30	**Course Introduction: Becoming a Unicorn** **Lab:** Installing and Editing Portfolios	Ch1 (Cady) Ch 1 & 2 (Erl)		
2	9/6	**Software Engineering Best Practices** **Data Science Roadmap and Life Cycle**	Ch 2, 15 (Cady)	GWU Meetup	

		My Personal Toolkit and Portfolio **Lab:** Installing and Editing Portfolios Python and GitHub Installation and Basics, Data Science Toolkit, Visualization, Interpreting, and Communicating Results	Ch 3 & 4 (Erl)		
3	9/13	**Review: Linear Algebra and Computer Science** **Lab:** Exploratory Data Analysis for Classification using Pandas and Matplotlib Programming Languages, Technical Communication and Documentation, Data Structures, Encodings and Formats, and Computer Memory	Ch 3, 9, 12, 20, 21, 22 (Cady) Ch 5 (Erl)	Guest Speaker: Vlad Barash, Graphika	Assignment #1: Intro to R/Python, GitHub, and developing a Portfolio Research Proposal Instructions will be handed out.
4	9/20	**Data Engineering and Data Munging** **Lab:** Web Scraping	Ch 4 (Cady) Ch 6 (Erl)	Guest Speaker: Ankit Saxena, GWU Graduate Student	
5	9/27	**Data Visualizations** **Lab:** Exploratory Data Analysis for Classification using Pandas and Matplotlib	Ch 5 (Cady) Ch. 7 (Erl) Ch 14 (Schutt)	Guest Speaker: Ruth Agada, Bowie State University	
6	10/4	**Big Data and Database Storage Technology** **Lab:** MapReduce	Ch 13 and 14 (Cady) Ch. 8 (Erl)	Guest Speaker: Ms. Laura Wrubel, GW Libraries	Assignment #2: Applying Analysis Techniques and Statistical Inference to Data Assignment #1: Due Students Research Proposals Due: (10/4)
7	10/11	**Machine Learning Classification and Feature Extraction** **Lab:** Scikit-Learn, Regression, PCA	Ch. 8 (Cady) Ch. 2 & 7	Guest Speaker: Donald Braman, Mayors Office, The Lab @ DC	

			(Schutt).		
8	10/18	**Unsupervised Learning and Regression** **Lab:** Bayes, Linear Regression, and Metropolis Sampling	Ch. 10 and 11 (Cady) Ch. 3 & 5 (Schutt).	Guest Speaker: Kato Mivule, Department of Defense	
9	10/25	**Data Analysis I: Artificial Intelligence (AI) and Natural Language Processing** **Lab:** Sampling and Text Processing	Ch. 16 (Cady)	Guest Speaker: Jane Liu, GWU Benjamin Ulloa, Virginia Tech Applied Research Corporation	
10	11/1	**Data Analysis II: Time Series Analysis and Real-Time data analysis** **Lab SVM and Neural Networks**	Ch. 17 (Cady)	Guest Speaker: Dr. Douglas Rose, Modern Technology Solutions Inc. (MTSI)	Assignment #3: Applying Machine Learning Techniques to Large Datasets Assignment #2: Due
11	11/8	**Data Analysis III: Probability, Statistics and Maximum Likelihood Estimation and Optimization** **Lab:** Bias, Variance, Cross-Validation	Ch. 18, 19, 23 (Cady)	Guest Speaker: Dr. Rodney Wallace, IBM	
12	11/15	**Stochastic Modeling and Advanced Classifiers** **Lab:** Bias, Variance, Cross-Validation	Ch. 24 and 25 (Cady)	Guest Speaker: Dr. Kola Ogunlana, Booz Allen Hamilton	
13	11/22	**Thanksgiving Holiday (No class)**			
14	11/29	**Machine Learning: III** **Graph Analysis and Recommender Systems**	Materials will be provided.	Guest Speaker: Dr. Davit Trott, Department of Defense	Assignment #4: Data Science Process, "Big Data", Visualization, Interpreting, and Communicating Results in your Portfolio

		Lab: Networks			Assignment #3: Due
15	12/6	**Machine Learning: IV** **Special Topics in Data Analytics**	Materials will be provided.	Speaker: Student Research Presentation	Assignment #4: Due 12/7 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 and Lab - Part I 5%
- Portfolio Project and Lab - Part II 10%
- Portfolio Project and Lab - Part III 15%
- Portfolio Project and Lab - Part IV 20%
- Final Research Project – Part I 25%
- Final Research Project – Part II 25%

Labs

This course will consist of a lecture portion and a lab portion at the end of each class. The topics for the labs are as follows:

- Installing and Editing Portfolios
- Web Scraping
- Sampling and Text Processing
- Exploratory Data Analysis for Classification using Pandas and Matplotlib
- Scikit-Learn, Regression, PCA
- Bias, Variance, Cross-Validation
- Bayes, Linear Regression, and Metropolis Sampling
- Support Vector Machines
- Networks
- MapReduce

Assignments

This course consists of four portfolio assignments, and a final research project. There will be a total of 500 points: Portfolio projects/assignments (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	10/4
Portfolio Project - Part II	Assignment 2 – Data Analysis, Statistical Inference, and Visualizations.	50	11/1
Portfolio Project - Part III	Assignment 3 - Machine Learning	75	11/29
Portfolio Project - Part IV	Assignment 4 - Data Science Process, "Big Data", Visualization, Interpreting, and Communicating Results in your Portfolio	100	12/7
Final Project – Part I	Research Proposal and Final Paper (1-5 pages)	125	10/4 and 12/7
Final Project – Part II	Portfolio and Research Presentation	125	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