UNIVERSITY OF ILLINOIS AT SPRINGFIELD

CSC 570 Data Science Essentials

Credit Hours: 4.0
Fall 2017

Instructor: Mike Bernico
Office:
Office Hours:

Phone:
Email: mike.bernico@gmail.com

Course Description

The Data Science Essentials course provides a foundation to data science and the practice of analytics. In addition to an introduction to Big Data, it also introduces students to the Data Analytics Lifecycle, which addresses business challenges that leverage big data. It includes a laboratory component to provide a hands-on foundation in basic analytic methods and big data technology and tools, including Apache Spark, MapReduce, and the Hadoop ecosystem. This course assumes no prior knowledge of Big Data Analytics.

Course Objectives/Learning Outcomes

Upon completion of this course, you should be able to:

- Participate and contribute as a data science team member on big data and other analytics projects
- Deploy a structured life-cycle approach to data science and big data analytics projects
- Reframe a business challenge as an analytics challenge
- Apply analytic techniques and tools to analyze big data, create statistical models, create machine learning models, and identify insights that can lead to actionable results
- Select optimal visualization techniques to clearly communicate analytic insights to business sponsors and others
- Use tools such as Python, R, Spark, Hive, Hadoop.
- Explain how advanced analytics can be leveraged to create competitive advantage and how the data scientist role and skills differ from those of a traditional business intelligence analyst

Course Content
There are 5 modules in this course. Each module will consist of:

Lectures and Reading:
In each module I'll talk about the material that's important that week. I'm going to try to do most of this with youtube videos. Additionally, we can expect a few guest speakers. I'll also post several links to videos to watch and other material to read.

Class Participation:
I expect everyone to participate in the online forum and/or slack channel every week. I'll be asking questions, and posting starter topics. I expect you to research the ideas I post on and comment on/discuss them with your classmates.

Labs:
These are hands-on exercises to be completed in the cloud or on your home computer.

Summary Lab / Project presentations
The last assignment of the semester is an application of the Data Analytics Lifecycle to a sample challenge.

Programming Environment
In an effort to make this experience more 'real world' all work will be submitted via GitHub (www.github.com). Your github code repo will be a public portfolio of Data Science work that you can showcase in the future. You will need to supply me with your github username, and repo name.

I will not be providing technical support or help with github. Students seeking an advanced degree in computer science need to be familiar with version control systems and it is your responsibility to build this familiarity. Assignments will only be accepted via github.

You may work on your own machine, or in the school's virtual environment. You may choose to write your code in either R or Python for the first 5 modules, my examples will be done using Python. You will be responsible for learning either R or Python on your own, however I will post some 'getting started' links.

All work needs to be submitted on Github. NO LATE WORK WILL BE ACCEPTED.

Grading
The final grade will be based on
10% Homework
30% Regression/Unsupervised Learning Tests
30% Midterm Project
30% Final Analytics Project

Working Together on Assignments: Students may work together on the homework problems, provided that the methods used promote a learning environment for all involved. Specific rules follow:

1. You may discuss assignments with your classmates.

2. When it comes time to do your assignment (actually write the answer), **you must write your own problem solution and turn in your own homework** (unless specified otherwise in the assignment).

3. In particular, you may **NOT:**

copy in part or in totality from another student's homework problems.
get someone else to do your work.
submit the work of a group as your own work.
submit solutions copied (in whole or in part) from any other source as your own work.

You may use external sources for your work so long as you properly CITE the source within your submission.

Academic Honesty

Acts of cheating and plagiarism will not be tolerated in this class. This included, but is not limited to, using someone else's work as yours, using someone else's original thoughts as your own, paying someone else to do your work, or cheating in other ways not consistent with the spirit of the assignment.

This is a data analytics course, and the instructor will use data science techniques to verify academic honesty. This course has a 0 tolerance policy for academic honesty violations. **Acts of plagiarism or cheating will result in a 0 result in a failing grade for the course.**

Course Schedule (Subject to Change):

Semester Week	Lecture / Notes	Lab	Due Date

Week 1:	**Module 0: Overview** **Module 0: Intro to Data Science** **Final Project Overview**		9/5 by 11:59pm
Week 2:	**Module 0: Data Science Lifecycle Overview** Lesson 1: What kinds of problems can be solved with Data Science? Lesson 2: What's are the steps we go through in the Data Science Lifecycle Lesson 3: How do we know that our Data Product is a good one?	Lab 1: Introduction to our Data Environment Lab 2: Setting Up iPython Notebook and downloading the Titanic dataset Homework Getting Started with Git	9/5 by 11:59pm
Week 3:	**Module 1: Converting Data to Features** Lesson 1: Getting Around Python Lesson 2: Analyzing/Exploring Data with Python Lesson 3: Handling missing values	Homework: Titanic EDA and Data Prep	9/11by 11:59pm
Week 4:	**Module 1: Converting Unstructured Data to Features** Lesson 1: Vectorizing Text using TF/IDF	Homework: TF/IDF Lab	9/18 by 11:59pm
Week 5:	**Module 2: Making Predictions - Regression** Lesson 1: Inputs and Outputs Lesson 2: Handling Categorical Variables Lesson 3: Measuring Regression Performance Lesson 4: How can it fail? Bias and Variance Lesson 4: Regularization and Normalization	Homework: Boston Housing Regression Lab	9/25 by 11:59pm

Week 6:	**Module 2: Making Predictions – Logistic Regression** Lesson 4: Going from Linear to Logistic – Differences Lesson 5: Binary Classification Metrics ROC/AUC, Precision, Recall Lesson 6: Tuning Logistic Regression	Homework: Logistic models applied to the Titanic Dataset	10/2 by 11:59pm
Week 7:	**Module 2: Making Predictions – Random Forest** Lesson 1: Decision Trees Lesson 2: Bootstrapping and Forests of Trees Lesson 3: Tuning Random Forest Lesson 4: Ensembling models	Homework: (optional) Titanic Random Forest and ensemble classifiers	10/9 by 11:59pm
Week 8:	**Module 2: Advanced Validation**	**TEST 1: Supervised Learning Test** Homework: Implementing K-Fold Cross Validation	10/16 by 11:59pm
Week 9:	**MIDTERM PROJECT WEEK**	Midterm Kaggle	10/19 by 11:59
Week 10:	**Module 2: Making Predictions – Other Models: Naive Bayes, SVM, and Deep Learning** Lesson 1: Naïve Bayes Lesson 2: SVM Lesson 3: Deep Learning	Homework: Text Classification with Naïve Bayes	10/30 by 11:59pm
Week 11:	Module 2: Deep Neural Nets and Computer Vision	Homework: Computer Vision	11/6 by 11:59
Week 12:	**Module 2:Unsupervised Machine Learning** Lesson 1: Clustering Lesson 2: Finding K	Homework: K-Means Clustering	11/13 by 11:59pm

	Final Project Proposal Submitted		
Week 13:	**Module 2: Unsupervised Machine Learning** Lesson 1: Recommenders Lesson 2: Evaluating Recommender Performance	Homework: Building a Recommender System	11/20 by 11:59pm
Week 14:	**Module 3: Graph Theory for Data Science** Graph Theory / Advanced Validation Techniques	Homework: Graph Theory	11/27 by 11:59pm
Week 15:	Module 4: Taking Data Products to Production Streaming and Lambda Architectures Encapsulating Models in web services	**TEST 2: UNSUPERVISED LEARNING TEST**	12/1 by 11:59 pm
Week 16:	FINAL PROJECT		Due 12/13 by 11:59pm

Midterm Format
My midterm is a class kaggle competition. You will be solving a data science problem and submitting your results to kaggle.

Students are encouraged to:
1. Start on this assignment as soon as they complete the regression lab.

2. Make a submission to the kaggle leaderboard to make sure you can generate a kaggle submission. You may submit as many entries as you like.

3. Scoring is based:
    50% on data cleaning and prep code
    25% modeling clode
    25% your kaggle results, relative to the class (not submitting to kaggle results in a best possible grade of 75%)

My midterm is hard. It will take longer than you think. Leave yourself ample time for this assignment.

Final Project Format
Details on the final project are posted in Week 0.

Most importantly:

1. You pick the topic. I'm leaving the project open ended because I want to help you develop the creativity necessary to be a good data scientist.
2. Once you pick a topic, you probably need some hypothesis. What specifically would you like to prove, disprove, or model?
3. On or before Midterm you'll need to propose your topic and plan to me. Unapproved final projects will not be graded/accepted.
4. The final presentation format will be a youtube.com video. You will need to verbally present your problem and your solution. This should be done in conjunction with appropriate slides and/or code walk throughs. Presentation matters.