CSYE 7245
Big-Data Systems and Intelligence Analytics
Fall 2017 Course Syllabus

Course Information

Professor: Nik Bear Brown
Email: nikbearbrown@gmail.com
Office:
Office hours:
Monday 10:30AM-12PM
Wednesday 10:30AM-11:30AM
Thursday 10:30AM-11:30AM
or by appointment.

Course website: Blackboard (for raw scores, uploading assignments, getting materials, & forums)
Piazza: https://piazza.com/northeastern/fall2017/csye7245

Course Prerequisites

INFO 7250 or INFO 7390 (either may be taken concurrently); engineering students only.

A basic knowledge of the python programming language is required. Assignments may be done in python or R.

Course Description

Offers students an opportunity to learn a hands-on approach to understanding how large-scale data sets are processed and how data science algorithms are adopted in the industry through case studies and labs. This project-based course builds on INFO 7390 and focuses on enabling students with tools and frameworks primarily to build end-to-end applications. The course is divided into three parts: building the data pipeline for data science, implementing data science algorithms, and scaling and deploying data science algorithms

The ability to use python is part of the grade. Students must demonstrate ability to setup data for learning, train, test, and evaluation using either python or R, but all assignment examples and solutions will be presented in python. The assignments include paper exercises designed to reinforce conceptual understanding. A term project is required.

Communication

Communication between instructor and students is through
- Email via the Blackboard distribution list

- Announcements posted on Blackboard
- Notes posted on the Blackboard discussion board
- Private email exchanges

Course Structure

- Regularly test students on paper/algorithmic exercises
- Evaluate students' implementation competency, using assignments that require coding on given datasets
- Evaluate ability to setup data, code, and execute using python language
- Student will be required to do "data digging": run analysis scripts and failure analysis
- Final project is typically asking and answering a "real world" question of interest using machine learning techniques

Course GitHub

The course GitHub (for all lectures, assignments and projects):

https://github.com/nikbearbrown/NEU_COE

nikbearbrown YouTube channel

Over the course of the semester I'll be making and putting additional data science and machine learning related video's on my YouTube channel.

https://www.youtube.com/user/nikbearbrown

The purpose of these videos is to put additional advanced content as well as supplemental content to provide additional coverage of the material in the course. Suggestions for topics for additional videos are always welcome.

Schedule

Week	Topic	Assignments
1) Week 1	Introduction & Essential Concepts in Big Data and Data Science. Intro to python. Data munging with python.	Readings; Assignment 1
2) Week 2	Descriptive Statistics, Probability Theory, Probability Distributions, Bayesian Probability, Inferential Statistics,	Readings

	Test Statistics, Hypothesis Testing, Clustering with python.	
3) Week 3	Support Vector Machine (SVM), Decision Tress, Random Forest, Naïve Bayesian Classifier (NBC), Bayesian Networks, Time Series Analysis with python.	Readings; Assignment 2
4) Week 4	Correlation and Regression, Regularization with python. Research project proposal.	Readings; Project proposal
5) Week 5	Data pipelines in python (Luigi & Airflow)	Readings; Assignment 3
6) Week 6	NoSQL and MongoDB	Readings
7) Week 7	The MapReduce paradigm & Hadoop and HDFS	Readings; Assignment 3
8) Week 8	Apache Spark	Readings; Project progress report.
9) Week 9	Neural Networks Shallow Neural Networks Intro to Deep Learning	Readings; Assignment 4
10) Week 10	Recurrent neural network (RNN) Sequence Modeling with Neural Networks	Readings
11) Week 11	Convolutional Neural Networks	Readings; Assignment 5
12) Week 12	Break	Thanksgiving recess
13) Week 13	Autoencoders Variational Autoencoders Generative Adversarial Networks (GANs) Deep Generative Models Boltzmann Machines	Readings; Assignment 6
14) Week 14	Research Project Presentations	Readings; A draft of the final project for feedback

Teaching assistants

The Teaching assistants for Fall 2017 are:

Programming questions should first go to the TA's. If they can't answer them then the TA's will forward the questions to the Professor.

Learning Assessment

Achievement of learning outcomes will be assessed and graded through:

● Completion of assignments involving scripting in R or python, and analysis of data (85%)
● Completion of a term paper asking and answering a "real world" question of interest using machine learning techniques (15%)

Reaching out for help

A student can always reach out for help to the Professor, Nik Bear Brown nik@ccs.neu.edu. In an online course, it's important that a student reaches out early should he/she run into any issues.

Grading Policies

Students are evaluated based on their performance on assignments, performance on exams, and both the execution and presentation of a final project. If a particular grade is required in this class to satisfy any external criteria—including, but not limited to, employment opportunities, visa maintenance, scholarships, and financial aid—it is the student's responsibility to earn that grade by working consistently throughout the semester. Grades will not be changed based on student need, nor will extra credit opportunities be provided to an individual student without being made available to the entire class.

Grading Rubric:

The following breakdown will be used for determining the final course grade:

Assignment	Percent of Total Grade
Assignments	85%
Research Project*	15%

* Note that the assignments, presentations and drafts related to the research project go to that score rather than the programming assignments. I expect to use the following grading scale at the end of the semester. You should not expect a curve to be applied; but I reserve the right to use one.

Score	Grade
93 – 100	A
90 – 92	A-
88 – 89	B+
83 – 87	B
80 – 82	B-
78 – 79	C+
73 – 77	C
70 – 72	C-
60 – 69	D
<60	F

Scores in-between grades. For example, 82.5 or 92.3 will be decided based on the exams.

Blackboard:

You will submit your assignments via Blackboard. Click the title of assignment (blackboard -> assignment -> <Title of Assignment>), to go to the submission page. You will know your score on an assignment, project or test via BlackBoard. BlackBoard represents only the raw scores. Not normalized or curved grades.

Due dates

Due dates for assignments are usually every other Monday at midnight.

Ten percent (i.e. 10%) is deducted for each day an assignment is late. Solutions will be posted the following Monday. Assignments will receive NO CREDIT if submitted after the solutions are posted.

Course Materials

Required text (All free online)

Some textbooks are all available for free to NEU students via SpringerLink (http://link.Springer.com/). You must access SpringerLink from an NEU IP address to have full access and/or download these books.

If you are off-campus, in order to access resources provided through the Northeastern library outside the network, you should use their bookmarklet to load any page through the proxy:
http://library.northeastern.edu/bookmarklet

Required Texts

The *required* textbooks we will be using in this class are:

An Introduction to Statistical Learning with Applications in python (2013)
Authors: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani
Free online via SpringerLink (http://link.Springer.com/) http://link.Springer.com/book/10.1007/978-1-4614-7138-7

The Definitive Guide to MongoDB: A complete guide to dealing with Big Data using MongoDB (2015)
Authors: David Hows, Peter Membrey, Eelco Plugge, Tim Hawkins
ISBN: 978-1-4842-1183-0 (Print) 978-1-4842-1182-3 (Online)
http://link.springer.com/book/10.1007/978-1-4842-1182-3

Deep Learning - Adaptive Computation and Machine Learning series by Ian Goodfellow, Yoshua Bengio, and Aaron Courville
https://github.com/HFTrader/DeepLearningBook

Beginning Python

From Novice to Professional
Authors: Magnus Lie Hetland 2017
ISBN: 978-1-4842-0029-2 (Print) 978-1-4842-0028-5
https://link.springer.com/book/10.1007/978-1-4842-0028-5

Deep Learning with Python
A Hands-on Introduction
Authors: Nikhil Ketkar 2017
ISBN: 978-1-4842-2765-7 (Print) 978-1-4842-2766-4
https://link.springer.com/book/10.1007/978-1-4842-2766-4

Pro Hadoop Data Analytics
Designing and Building Big Data Systems using the Hadoop Ecosystem
Authors: Kerry Koitzsch 2017
ISBN: 978-1-4842-1909-6 (Print) 978-1-4842-1910-2
https://link.springer.com/book/10.1007/978-1-4842-1910-2

Pro Apache Hadoop
Authors: Sameer Wadkar, Madhu Siddalingaiah 2014
ISBN: 978-1-4302-4863-7 (Print) 978-1-4302-4864-4
https://link.springer.com/book/10.1007/978-1-4302-4864-4

Pro Spark Streaming
The Zen of Real-Time Analytics Using Apache Spark
Authors: Zubair Nabi 2016
ISBN: 978-1-4842-1480-0 (Print) 978-1-4842-1479-4
https://link.springer.com/book/10.1007/978-1-4842-1479-4

Recommended Texts

Pro Python Best Practices
Debugging, Testing and Maintenance
Authors: Kristian Rother 2017
ISBN: 978-1-4842-2240-9 (Print) 978-1-4842-2241-6 (Online)
https://link.springer.com/book/10.1007/978-1-4842-2241-6

Mastering Machine Learning with Python in Six Steps
A Practical Implementation Guide to Predictive Data Analytics Using Python
Authors: Manohar Swamynathan 2017
ISBN: 978-1-4842-2865-4 (Print) 978-1-4842-2866-1
https://link.springer.com/book/10.1007/978-1-4842-2866-1

Introduction to Data Science
A Python Approach to Concepts, Techniques and Applications
Authors: Laura Igual, Santi Seguí 2017
ISBN: 978-3-319-50016-4 (Print) 978-3-319-50017-1

https://link.springer.com/book/10.1007/978-3-319-50017-1

Python Recipes Handbook
A Problem-Solution Approach
Authors: Joey Bernard 2016
ISBN: 978-1-4842-0242-5 (Print) 978-1-4842-0241-8
https://link.springer.com/book/10.1007/978-1-4842-0241-8

Lean Python
Learn Just Enough Python to Build Useful Tools
Authors: Paul Gerrard 2016
ISBN: 978-1-4842-2384-0 (Print) 978-1-4842-2385-7
https://link.springer.com/book/10.1007/978-1-4842-2385-7

Learn to Program with Python
Authors: Irv Kalb 2016
ISBN: 978-1-4842-1868-6 (Print) 978-1-4842-2172-3
https://link.springer.com/book/10.1007/978-1-4842-2172-3

Big Data Made Easy
A Working Guide to the Complete Hadoop Toolset
Authors: Michael Frampton 2015
ISBN: 978-1-4842-0095-7 (Print) 978-1-4842-0094-0
https://link.springer.com/book/10.1007/978-1-4842-0094-0

ggplot2: Elegant Graphics for Data Analysis (2009)
Authors: Hadley Wickham
Free online via (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-0-387-98141-3

R Quick Syntax Reference
Authors: Margot Tollefson
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-1-4302-6641-9

Probability for Statistics and Machine Learning Fundamentals and Advanced Topics (2011)
Authors: Anirban DasGupta Springer Texts in Statistics
Free online via SpringerLink (http://link.Springer.com/ http://link.Springer.com/book/10.1007/978-1-4419-9634-3

The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2011)
Authors: Trevor Hastie, Robert Tibshirani and Jerome Friedman
Free online http://web.stanford.edu/~hastie/local.ftp/Springer/OLD/ESLII_print4.pdf

Data Manipulation with R
Authors: Statistical Computing Facility Phil Spector
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-0-387-74731-6

Beginning Data Science with R (2014)
Authors: Manas A. Pathak
ISBN: 978-3-319-12065-2 (Print) 978-3-319-12066-9 (Online)
http://link.springer.com/book/10.1007/978-3-319-12066-9

The Definitive Guide to SQLite (2010)
Authors: Grant Allen, Mike Owens
ISBN: 978-1-4302-3225-4 (Print) 978-1-4302-3226-1 (Online)
http://link.springer.com/book/10.1007/978-1-4302-3226-1

The Definitive Guide to MongoDB: A complete guide to dealing with Big Data using MongoDB (2015)
Authors: David Hows, Peter Membrey, Eelco Plugge, Tim Hawkins
ISBN: 978-1-4842-1183-0 (Print) 978-1-4842-1182-3 (Online)
http://link.springer.com/book/10.1007/978-1-4842-1182-3

Beginning CouchDB (2009)
Authors: Joe Lennon
ISBN: 978-1-4302-7237-3 (Print) 978-1-4302-7236-6 (Online)
http://link.springer.com/book/10.1007/978-1-4302-7236-6

Beginning Neo4j (2015)
Authors: Chris Kemper
ISBN: 978-1-4842-1228-8 (Print) 978-1-4842-1227-1 (Online)
http://link.springer.com/book/10.1007/978-1-4842-122

A Tiny Handbook of R
Authors: Mike Allerhand
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-3-642-17980-8

R Statistical Application Development by Example Beginner's Guide
Authors: Robert J Knell
PDF: http://www.introductoryr.co.uk/Introductory%20R%20example%20chapters.pdf

Beginning Data Science with R
Authors: Manas A. Pathak
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-3-319-12066-9

Data Mining with Rattle and python The Art of Excavating Data for Knowledge Discovery (2011)
Authors: Graham Williams
Free online via SpringerLink (http://link.Springer.com/) http://link.Springer.com/book/10.1007/978-1-4419-9890-3

Bayesian Essentials with R
Authors: Jean-Michel Marin, Christian P. Robert

Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-1-4614-8687-9

Text Analysis with python for Students of Literature
Authors: Matthew L. Jockers
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-3-319-03164-4

Introductory Time Series with R
Authors: Andrew V. Metcalfe, Paul S.P. Cowpertwait
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-0-387-88698-5

Data Analytics - Models and Algorithms for Intelligent Data Analysis 2012
Authors: Thomas A. Runkler
ISBN: 978-3-8348-2588-9 (Print) 978-3-8348-2589-6 (Online)
http://link.Springer.com/book/10.1007/978-3-8348-2589-6

Computational Social Network Analysis: Trends, Tools and Research Advances 2010
Editors: Ajith Abraham, Aboul-Ella Hassanien, Vaclav Snáˇel
ISBN: 978-1-84882-228-3 (Print) 978-1-84882-229-0 (Online)
http://link.Springer.com/book/10.1007/978-1-84882-229-0

Biostatistics with R
Authors: Babak Shahbaba
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-1-4614-1302-8

Introduction to Probability Simulation and Gibbs Sampling with R
Authors: Eric A. Suess, Bruce E. Trumbo
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-0-387-68765-0

A Modern Approach to Regression with R
Authors: Simon Sheather
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-0-387-09608-7

R by Example
Authors: Jim Albert, Maria Rizzo
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-1-4614-1365-3

Graphical Models with R
Authors: Søren Højsgaard, David Edwards, Steffen Lauritzen
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-1-4614-2299-0

Data Manipulation with R
Authors: Statistical Computing Facility Phil Spector
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-0-387-74731-6

A Tiny Handbook of R
Authors: Mike Allerhand
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-3-642-17980-8

Bayesian Networks in R
Authors: Radhakrishnan Nagarajan, Marco Scutari, Sophie Lèbre
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-1-4614-6446-4

Introducing Monte Carlo Methods with R
Authors: Christian Robert, George Casella
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-1-4419-1576-4

Two-Way Analysis of Variance
Statistical Tests and Graphics Using R
Authors: Christian Robert, George Casella
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-1-4614-2134-4

Applied Spatial Data Analysis with R
Authors: Roger S. Bivand, Edzer Pebesma, Virgilio Gómez-Rubio
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-1-4614-7618-4

Nonlinear Regression with R
Authors: Christian Ritz, Jens Carl Streibig
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-0-387-09616-2

An Introduction to python for Quantitative Economics
Authors: Vikram Dayal
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-81-322-2340-5

The Environment in Economics and Development
Authors: Vikram Dayal
Free online via SpringerLink (http://link.Springer.com/)
http://link.Springer.com/book/10.1007/978-81-322-1671-1

Software

python Anaconda

- https://www.continuum.io/anaconda-overview

R (Statisical programming language)

- R project https://www.r-project.org/

RStudio (IDE)

- RStudio https://www.rstudio.com/products/rstudio/download3/

Python Tutorials

Dive into Python http://diveintopython.org

Python 101 – Beginning Python http://www.rexx.com/~dkuhlman/python_101/python_101.html

The Official Python Tutorial http://www.python.org/doc/current/tut/tut.html

The Python Quick Reference http://rgruet.free.fr/PQR2.3.html

Python Fundamentals Training – Classes http://www.youtube.com/watch?v=rKzZEtxIX14

Python 2.7 Tutorial Derek Banas· http://www.youtube.com/watch?v=UQi-L-_chcc

Python Programming Tutorial - thenewboston http://www.youtube.com/watch?v=4Mf0h3HphEA

Google Python Class http://www.youtube.com/watch?v=tKTZoB2Vjuk

Nice free CS/python book https://www.cs.hmc.edu/csforall/index.html

datacamp.com https://www.datacamp.com/tracks/python-developer

R Tutorials

LearnR
https://youtu.be/p3i7Kz6C_-4?list=PLFAYD0dt5xCwDNFdrqeNoU9t-nhAWkbKe

Try python @codeschool: http://tryr.codeschool.com

Datacamp python Tutorials
https://www.datacamp.com/

rstudio online learning
https://www.rstudio.com/online-learning/

Deep Learning Tutorials

MIT 6.S191: Introduction to Deep Learning http://introtodeeplearning.com/

Stanford Winter Quarter 2016 class: CS231n: Convolutional Neural Networks for Visual Recognition
https://youtu.be/NfnWJUyUJYU

Deep Learning - Adaptive Computation and Machine Learning series by Ian Goodfellow, Yoshua Bengio, and Aaron Courville
https://github.com/HFTrader/DeepLearningBook

Participation Policy

Participation in discussions is an important aspect on the class. Participation on the Piazza discussion forums (e.g. asking and answering questions about assignments, discussing readings, midterm review) also counts towards this grade. It is important that both students and instructional staff help foster an environment in which students feel safe asking questions, posing their opinions, and sharing their work for critique. If at any time you feel this environment is being threatened—by other students, the TA, or the professor—speak up and make your concerns heard. If you feel uncomfortable broaching this topic with the professor, you should feel free to voice your concerns to the Dean's office.

Collaboration Policies

Students are strongly encouraged to collaborate through discussing strategies for completing assignments, talking about the readings before class, and studying for the midterms. However, all work that you turn in to me with your name on it must be in your own words or coded in your own style. Directly copied code or text from any other source is not allowed. In any case, you must write up your solutions, in your own words. Furthermore, if you did collaborate on any problem, you must clearly list all of the collaborators in your submission.

Feel free to discuss general strategies, but any written work or code should be your own, in your own words/style. If you have collaborated on ideas leading up to the final solution, give each other credit on what you turn in, clearly labeling who contributed what ideas. Individuals should be able to explain the function of every aspect of group-produced work. Not understanding what plagiarism is does not constitute an excuse for committing it. You should familiarize yourself with the University's policies on academic dishonesty at the beginning of the semester. If you have any doubts whatsoever about whether you are breaking the rules – ask!

To reiterate: **plagiarism and cheating are strictly forbidden. No excuses, no exceptions.** *All incidents of plagiarism and cheating will be sent to OSCCR for disciplinary review.*

Assignment Late Policy

Assignments are due by 11:59pm on the due date marked on the schedule. Late assignments will receive a 10% deduction per day that they are late, including weekend days. It is your responsibility to determine whether or not it is worth spending the extra time on an assignment vs. turning in incomplete work for partial credit without penalty. Any exceptions to this policy (e.g. long-term illness or family emergencies) must be approved by the professor.

Student Resources

Special Accommodations/ADA: In accordance with the Americans with Disabilities Act (ADA 1990), Northeastern University seeks to provide equal access to its programs, services, and activities. If you will need accommodations in this class, please contact the Disability Resource Center (www.northeastern.edu/drc/) *as soon as possible* to make appropriate arrangements, and please provide the course instructors with any necessary documentation. The University requires that you provide documentation of your disabilities to the DRC so that they may identify what accommodations are required, and arrange with the instructor to provide those on your behalf, as needed.

Academic Integrity: All students must adhere to the university's Academic Integrity Policy, which can be found on the website of the Office of Student Conduct and Conflict Resolution (OSCCR), at http://www.northeastern.edu/osccr/academicintegrity/index.html. Please be particularly aware of the policy regarding plagiarism. As you probably know, plagiarism involves *representing anyone else's words or ideas as your own*. It doesn't matter where you got these ideas—from a book, on the web, from a fellow-student, from your mother. It doesn't matter whether you quote the source directly or paraphrase it; if you are not the originator of the words or ideas, *you must state clearly and specifically where they came from*. Please consult an instructor if you have any confusion or concerns when preparing any of the assignments so that together. You can also consult the guide "Avoiding Plagiarism" on the NU Library Website at http://www.lib.neu.edu/online_research/help/avoiding_plagiarism/. If an academic integrity concern arises, one of the instructors will speak with you about it; if the discussion does not resolve the concern, we will refer the matter to OSCCR.

Writing Center: The Northeastern University Writing Center, housed in the Department of English within the College of Social Sciences and Humanities, is open to any member of the Northeastern community and exists to help any level writer, from any academic discipline, become a better writer. You can book face-to-face, online, or same day appointments in two locations: 412 Holmes Hall and 136 Snell Library (behind Argo Tea). For more information or to book an appointment, please visit http://www.northeastern.edu/writingcenter/.