INFO 6105: DATA SCIENCE ENGINEERING METHODS AND TOOLS

Course Information

Course Title: Data Science Engineering methods and tools
Course Number: INFO 6105, CRN 16779
Term and Year: Fall 2020
Credit Hour: 4 credits
Location: Online

About this course

Are you looking to explore a high-paying career as a data scientist in the future? Maybe, you are wondering if you have the right skills to pursue data science? Or you are thinking about how best you can hone your existing skills towards becoming a professional data scientist? If you are then, check out INFO 6105. Since we have a very different teaching philosophy, individuals from a variety of backgrounds and strengths will find this course useful. This means that if you currently have just high-school level math knowledge, you can still think about taking this course and, aspire to become a good data scientist. *This course will be taught via more than 22 hrs of lively and, star-rated video lectures and other materials!*

The goal of this online course is to introduce the fundamentals of machine learning (ML) and data science. This course is aimed at individuals who are looking to learn and apply state-of-the-art ML algorithms to real-world datasets.

There are two unique and foundational pillars of this course— (1) We use real-world situational analogies to intuitively understand the algorithmic aspects of data science and ML, and (2) We use real-world datasets and use powerful machine learning libraries in Python, to put these algorithms to practice.

Students enrolled in this course can expect to learn both theory and practice of modern data science in a very interactive and intuitive manner. After completing this course, participants will be able to build effective ML pipelines within minutes.

What will you learn in this course?

1. Instead of doing a survey of all machine learning algorithms, we focus deeply on the three or four algorithms that are commonly used by practicing data scientists in the industry.

2. Understand the central principles of machine learning from a conceptual, mathematical, and programming point of view.
3. Understand the why and how of data pre-processing for machine learning.
4. How to build, refine, and measure the performance of a machine learning model using real-life datasets.
5. Explore the commonalities and differences between some of the state-of-the-art machine learning algorithms including deep learning.
6. How to do all of the above in a cloud computing framework (Google Cloud Platform)
7. Machine learning best practices for the industry as it's currently being practiced.
8. Tips and tricks to troubleshoot problems in real-life data sets. E.g. How to tackle datasets containing predominantly one class, i.e unbalanced datasets
9. Where to go from here to further your data science career goals.

Course Syllabus

In addition to covering the material below, you are expected to work on a project. There are two options to choose a project — (1) From an instructor provided selection of real-world datasets, OR (2) come up with your own project idea and discuss it with the instructor. All discussions and communications during this online course, will be via email, or zoom.

Course introduction — structure, operations, components, and content overview. Why and what is machine learning with some examples. The burning hot data science job market, and how this course will help you land a job in data science. Brief introduction to Python, and an overview of Python libraries commonly used by data scientists.

Supervised and unsupervised learning. Classification Vs Regression. Machine learning best practices — train-valid-test split. The K- nearest neighbor algorithm. Data pre-processing. Bias-variance trade off, and how to improve model performance. Measures of classifier performance. Decision trees.

Extending decision trees to Random Forests. Introduction to ensemble models and bagging. Using excel and Python to create a random forest classifier. More about Random Forests. Hyper-parameter tuning, and exploratory data analysis with random forests.

Introduction to artificial neural networks and deep learning — why are they popular, examples, their relation to linear algebra. Different neural network architectures. Basics of linear algebra. Neural networks as successive transformations of the input vector.

Components of neural networks — forward of activations, error calculation, back-propagation of the error gradients, weights updating. Activation functions - logistic, tanh, ReLU, and softmax. Loss functions — binary and categorical cross-entropy. Introduction to Keras, tensor flow and pytorch deep learning libraries. Stochastic gradient descent, learning rate, and the loss function landscape for deep neural networks. Saddle Points.

Convolutional neural network — theory and practice. Using excel to understand convolutions. Transfer learning with convolutional neural networks. Comparison and contrast with standard neural network architecture.

Unsupervised learning — Clustering and discovering structure in the data. Principal Component Analysis or PCA to visualize high dimensional datasets.

Meet the instructor

Ramkumar Hariharan is currently head of applied AI with Macro-Eyes, Seattle, where he drives diverse projects in the Artificial Intelligence & Healthcare space. Previously, he has led multiple high-impact data-driven projects at some of the leading institutes in Seattle. These include Fred Hutch, University of Washington (UW), and the Institute for Systems Biology. His areas of focus include data analyses, data visualization, and predictive analytics of both structured and unstructured data.

Ram has a 15-year history of developing and delivering more than 20 computational, biomedical, and data science courses at a variety of levels. His courses, lectures, online teaching, and motivational talks have been overwhelmingly well-received in Seattle, Japan and in India. He has also “edutained” on local and national Television and Radio in India. Ram serves as affiliate faculty at Northeastern University, affiliate of UW e-sciences institute, bootcamp leader at General Assembly, and mentor with Springboard. He has also led education and training programs for Fred Hutch. He specializes in using powerful, yet simple analogies to explain seemingly complex computational and data science concepts and math.

Ram’s teaching philosophy is grounded in one strong belief: there is no one size fits all approach to teach, or to learn a new concept.

His brand statement — telling humorous stories with at least one takeaway!

How is the course going to be delivered?

This course is organized as a series of modules, one per week. Each module will contain

(1) One to two hours of lively, high-quality, five-star rated video lectures
(2) Slides from the videos
(3) Jupyter notebooks with python code used in the videos
(4) Data for running the examples, assignments and final project
(5) Links to great resources
(6) Text books suggestions

How are you going to be graded?

There will be three assignments in total, one going out every third week. Each assignment will have questions that either require you to write a response, select a response, or more likely, write Python code to solve a problem. Your performance on the assignments will contribute 70% towards your final grade.

You are required to complete one final project beginning at week 4. You will be provided with links to download the data and the goal of the project will be predefined. You will be asked to submit code that goes all the way from data pre-processing to final results. Machine learning model performance plots are very important. Scores on your project will contribute 30% towards your final grade for this course.

Grade Scale

95-100%	**A**	**87-89.9%**	**B+**	**77-79.9%**	**C+**	69.9% or below F
		84-86.9%	**B**	**74-76.9%**	**C**	
90-94.9%	**A-**	**80-83.9%**	**B-**	**70-73.9%**	**C-**	

Pre-requisites

Some familiarity with programming in any computer programming language!

Attendance Policy
This is an online course and you can take it at your schedule.

Late Work Policy

Students must submit assignments by the deadline in the time zone noted in the syllabus.
Students must communicate with the faculty prior to the deadline if they anticipate work will be submitted late.
Work submitted late without prior communication with faculty will not be graded.

Course reviews by previous INFO 6105 students

“Ram is the friendliest professor I have had…” **—** Spring 2020 student
“Ram can teach machine learning to my grandma and she will completely understand it” **—** Spring 2019 student

“I thoroughly enjoyed the course. would 100% recommend your course to anyone interested in starting out with Data Science” **—** Summer 2019 student

“I found the course to be very interesting as its design is very simple and understandable” **—** Summer 2019 student

“Used techniques from your course for my data science internship. Thank you” — Spring 2019 student

How to ask for help and other benefits

Ram and TA’s will be available by email throughout the duration of this course and will gladly help out students with INFO 6105.

Perks: for active data science job seekers, Ram will be happy to leverage his professional network to pass along CVs of students! This has resulted in some of his previous students landing jobs, or sometimes getting interviews from companies!

Text Books

These are a few suggestions. Please remember that we made this course from scratch and we will not follow any single textbook!

Practical treatment

- Introduction to Machine Learning with Python: A Guide for Data Scientists, Andreas C. Müller and Sarah Guido

- Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow, 2nd Edition, Sebastian Raschka and Vahid Mirjalili

Theoretical treatment

- An Introduction to Statistical Learning: With Applications in R, Daniela Witten, Gareth James, Robert Tibshirani, and Trevor Hastie (Legally free e-book here : https://www-bcf.usc.edu/~gareth/ISL/)

- The Elements of Statistical Learning, Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie (legally free e-book here : https://web.stanford.edu/~hastie/ElemStatLearn/)

- Pattern Recognition and Machine Learning, Christopher Bishop

Academic Integrity

A commitment to the principles of academic integrity is essential to the mission of Northeastern University. The promotion of independent and original scholarship ensures that students derive the most from their educational experience and their pursuit of knowledge. Academic dishonesty violates the most fundamental values of an intellectual community and undermines the achievements of the entire University.

As members of the academic community, students must become familiar with their rights and responsibilities. In each course, they are responsible for knowing the requirements and restrictions regarding research and writing, examinations of whatever kind, collaborative work, the use of study aids, the appropriateness of assistance, and other issues. Students are responsible for learning the conventions of documentation and acknowledgment of sources in their fields. Northeastern University expects students to complete all examinations, tests, papers, creative projects, and assignments of any kind according to the highest ethical standards, as set forth either explicitly or implicitly in this Code or by the direction of instructors.

Go to http://www.northeastern.edu/osccr/academic-integrity-policy/ to access the full academic integrity policy.

Student Accommodations
Northeastern University and the Disability Resource Center (DRC) are committed to providing disability services that enable students who qualify under Section 504 of the Rehabilitation Act and the Americans with Disabilities Act Amendments Act (ADAAA) to participate fully in the activities of the university. To receive accommodations through the DRC, students must provide appropriate documentation that demonstrates a current substantially limiting disability.

For more information, visit http://www.northeastern.edu/drc/getting-started-with-the-drc/.

Library Services
The Northeastern University Library is at the hub of campus intellectual life. Resources include over 900,000 print volumes, 206,500 e-books, and 70,225 electronic journals.

For more information and for Education specific resources, visit http://subjectguides.lib.neu.edu/edresearch.

Diversity and Inclusion

Northeastern University is committed to equal opportunity, affirmative action, diversity and social justice while building a climate of inclusion on and beyond campus. In the classroom, member of the University community work to cultivate an inclusive environment that denounces discrimination through innovation, collaboration and an awareness of global perspectives on social justice.
Please visit http://www.northeastern.edu/oidi/ for complete information on Diversity and Inclusion

TITLE IX

Title IX of the Education Amendments of 1972 protects individuals from sex or gender-based discrimination, including discrimination based on gender-identity, in educational programs and activities that receive federal financial assistance.

Northeastern's Title IX Policy prohibits Prohibited Offenses, which are defined as sexual harassment, sexual assault, relationship or domestic violence, and stalking. The Title IX Policy applies to the entire community, including male, female, transgender students, faculty and staff.

In case of an emergency, please call 911.

Please visit www.northeastern.edu/titleix for a complete list of reporting options and resources both on- and off-campus.