

INFO 7390: Advances in Data Science and Architecture

Course Information

Course Title: Advances in data science and architecture

Course Number: INFO 7390, CRN 17729

Term and Year: Fall 2022

Credit Hour: 4 credits

Location: Online

About this course

Are you looking to acquire practical data science skills in areas like machine and deep learning? Maybe, you are looking to learn some fundamentals, at the same time applying them on new problems involving big data? 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 7390. Since we have a very different teaching philosophy, individuals from a variety of backgrounds and strengths will find this course useful. *This course will be taught via online videos and live streaming weekly office hours.*

The goal of this course is to foray into advanced machine learning and data science. will cover basic and advanced machine learning with focus on deep learning.

Students enrolled in this course can expect to learn both theory and practice of advanced modern data science in the context of big data, in a very interactive and intuitive manner. After completing this course, participants will be able to build effective ML pipelines for solving industry focused problems.

What will you learn in this course?

1. We start off by doing a pre-defined machine learning project selected by the instructor or by you.
2. Students are expected (with guidance from instructor) to gather data, preprocess it, build deep learning model and evaluate performance.
3. We focus on aspects of big data including choice of algorithms, compute environment, etc....
4. Understand the central principles of machine learning, recommender systems, and transfer learning from a conceptual, mathematical, and programming point of view.
5. Understand the why and how of data pre-processing for machine learning.

6. How to do all of the above in a cloud computing framework (Google Colab or the like)
7. Machine learning best practices for the industry as it's currently being practiced.
8. Where to go from here to further your data science career goals.

Course Syllabus

All these topics will be discussed from a project-centric point of view.

Course introduction — project expectations, presentation, writing up an academic paper based on this project, etc.

Unsupervised learning — importance, clustering, K-means clustering, feature importance in the context of clustering.

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 and differential calculus. Neural networks as successive transformations of the input vector.

Convolutional neural networks — theory and practice. Using excel to understand convolutions. Transfer learning with convolutional neural networks. How to use variable learning rates for transfer learning. Image augmentation techniques and an understanding of some common pretrained deep neural network models (Vgg 19, Resnets, etc.).

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. How to use variable learning rates for transfer learning.

Recommender systems: theory and practice. Collaborative filtering, content-based filtering and hybrid approaches. Live coding with the surprise package.

Basics of Natural Language Processing (NLP). Data preprocessing (removing stop words), tokenization, numericalization, language models and classification models. Concept of embeddings and their importance.

Recurrent neural networks and their applications.

Meet the instructor



Ramkumar Hariharan is currently Associate Teaching Professor and Associate Director of Program Management at Northeastern University, Seattle. He is also an AI consultant for start-ups in the Seattle area. 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 of University of Washington e-sciences institute, bootcamp leader at General Assembly, and mentor with Springboard. He is a recipient of the Springboard mentor inspire award for motivating students. 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.

How is the course going to be delivered?

This course is organized as a series of online videos and livestream discussions.

How are you going to be graded?

There will be one assignment. Your performance on the assignments will contribute 40% towards your final grade.

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 60% towards your final grade for this course. Projects can be as team work, with 2 to 4 members.

Grade Scale

96-100%	A	87-90.9%	B+	77-79.9%	C+	69.9% or below F
		84-86.9%	B	74-76.9%	C	
91-95.9%	A-	80-83.9%	B-	70-73.9%	C-	

Attendance Policy

This is an online course and you can take it at your schedule. But please show up for the live project discussion every week.

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 carry a penalty of 15 marks. To ensure fairness, no exceptions will be made.

Pre-requisites

Familiarity with programming in Python and INFO 6105.

Course reviews by previous INFO 6105, INFO 7390 and CSYE 7245 students

“Ram’s INFO 7390 is one of my favorite courses at NEU” — Spring 2020 student

“Ram is one of the friendliest professors I have come across...” — 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.

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://subject-guides.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.