

INFO 7390: Advances in Data Science and Architecture

Course Information

Course Title: Advances in data science and architecture

Course Number: INFO 7390,

Year: Spring 2024

Credit Hour: 4 credits

Location: Arlington

About this course

This course aims to delve into the intricate realms of advanced machine learning and data science. We will explore a range of machine learning techniques, placing a special emphasis on deep learning.

Throughout this course, students will gain knowledge in both the theoretical underpinnings and practical applications of contemporary data science, particularly within the framework of big data. This learning experience will be highly interactive and intuitive. Upon completion, course participants will be equipped with the skills to construct efficient machine learning pipelines tailored for real-world industry challenges.

In this course, you will engage in the following learning experiences:

1. The course commences with a machine learning project, which can be determined either by the instructor or yourself.
2. Under instructor guidance, students will be involved in data collection, preprocessing, constructing a deep learning model, and assessing its performance.
3. The course places a significant emphasis on big data elements, including the selection of appropriate algorithms and computing environments.
4. Gain a comprehensive understanding of the fundamental concepts of machine learning, recommender systems, and transfer learning, encompassing theoretical, mathematical, and coding perspectives.
5. Learn the rationale and techniques behind data preprocessing in the context of machine learning.
6. Apply the aforementioned concepts using a cloud-based computing environment, such as Google Colab.
7. Acquire knowledge of current industry-standard best practices in machine learning.
8. Receive guidance on potential future directions to enhance your career in data science.

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.

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.

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.

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.).

Introduction to Pytorch deep learning libraries. Stochastic gradient descent, learning rate, and the loss function landscape for deep neural networks. Saddle Points.

Programming with Large Language Models (LLMs) — Introduction to LLMs like GPT or others. Accessing and utilizing pre-trained models for NLP tasks. Practical applications and ethical considerations.

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.

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

William Claster is a seasoned educator and author. Dr. Claster brings a wealth of knowledge in data science, machine learning, and NLP. He is the author of a comprehensive textbook on machine learning and has contributed to various peer-reviewed journals.

His expertise extends to hands-on industry experience in fintech, banking, and IT, along with consulting experience for prominent financial institutions. Dr. Claster has developed and implemented a robust Data Science curriculum and has guided students through intricate data science. His teaching philosophy emphasizes a practical, project-based approach, ensuring that students gain real-world skills alongside theoretical knowledge.

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-	

Late Work Policy

Students must submit assignments by the deadline 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.

How to ask for help and other benefits

Professor Claster and TA's will be available by email throughout the duration of this course and will gladly help out students.

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 genderbased 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.