DAMG 7245- Big Data Systems and Intelligence Analytics

Course Information:

Course ID: DAMG 7245

Course Title: Big Data Systems and Intelligence Analytics

Term and Year: Fall 2024

Credit Hour: 4

Class time: Thursday 5:20 pm – 8:40 pm

Instructor Information:

Name: Garret Vo, PhD

Email: g.vo@northeastern.edu

Office hour: By appointment

Instructor Biography:

Dr. Garret Vo is a research scientist at the Naval Research Laboratory. His research intersects machine learning, optimization, and signal processing.

Course Description:

This course is a project-based course, which offers students an opportunity to learn a hand-on approach to perform processing on large-scale dataset and build machine learning/data science algorithms to answer relevant business questions.

In this course, students will be first provided with knowledge to build data pipelines, develop and deploy machine learning algorithms, monitor these algorithms' performance, and visualization of the results. Afterwards, students will use the knowledge to work on projects throughout the course.

The course includes the following topics:

a. Understanding and capturing business needs for data analytics and machine learning.
b. Building data processing pipeline.
c. Feature engineering.
d. Building machine learning algorithms.
e. Deploy and monitor machine learning algorithms (MLOPs).
f. Visualization

Student Learning Outcomes:

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

a. Understand and capture business needs for data analytics and machine learning.
b. Design and build data applications and pipelines.
c. Design and build machine learning algorithms for data applications.
d. Visualize and present the results to business stakeholders.

Prerequisite:

Students must have grade B in DAGM 6105 or INFO 6105

Software:

The software for the class is Python

Textbook:

The course has no required textbooks. These books below are for references:

1. Pattern Recognition and Machine Learning – Christopher Bishop
2. Deep Learning – Ian Goodfellow, Yoshua Bengio and Aaron Courville
3. Deep Learning with Pytorch – Thomas Viehmann, Eli Stevens, and Luca Pietro Giovarnni Antiga
4. Data Pipeline Pocket Reference – James Densmore
5. Machine Learning in Production: Developing and Optimizing Data Science Workflow and Applications – Andrew Kelleher and Adam Kelleher
6. Machine Learning Production System – Robert Crowne, Hannes Hapke, and Emily Caveness – will be published in September 2024.
7. Data Science in Production: Building Scalable Model Pipelines with Python – Ben G Weber

Lectures:

Lectures will include material knowledge, discussion, illustration of methodologies, and in-class exercises. Lectures and in-class exercises will be posted two days after class. We will have guest lectures and in-class exercises.

Grading:

Final project: 100%

Attendance Policy:

You are expected to attend all lectures and participate in class.

If you plan to miss a class, you must email the instructor about your absence. To get participation credit for the class you miss, you must submit a summary of the lecture of the day

you miss. If you miss more than 2 days and do not submit the reports, you will lose participation credit.

If there is a class exercise the day you miss, you must do the in-class exercises to get credit for in-class exercises.

In-class Exercises:

None

Case Study Projects:

None

Final Project:

The final project is selected by students. Students and instructors will work through the project as the course progresses.

Students need to deliver final report and codes associated with the project.

Late Work Policy:

Students must submit the assignments by the deadline in the time zone noted in Canvas.

Students must communicate with the instructor before the deadline if they anticipate the work will be submitted late.

Assignments submitted late without prior communication with the instructor will be subjected to a late penalty (i.e. reduction) of 10% per day.

Tentative Course Schedule:

Week	Topic	Note
1	Business needs for Data Science	
2	Data-driven applications and pipelines	
3	Data Source and Format	
4	Feature Engineering	Project given
5	Machine Learning and Deep Learning	
6	Feature Engineering and Model Building with Deep Learning and Machine Learning	
7	Introduction to the Cloud and HPC environment	
8	Parameters and hyper-parameters tuning	
9	Model Saving, and Deployment	
10	Model Monitoring and Orchestration	
11	Visualization	
12	Frontier topics: Deploying machine learning models on hardware	
13	Frontier topics: Physics-informed machine learning	
14	Final Project Presentation	Final report and code due