Northeastern University

Course ID: CS6220 Data Mining Techniques

Fall 2023

Instructor: Everaldo Aguiar (*e.aguiar@northeastern.edu*)
TAs: Vrinda Bisani (*bisani.v@northeastern.edu*)

Course Description

This course will cover a variety of practical aspects of data mining, starting with data understanding and manipulation and transitioning into supervised and unsupervised learning. The latter half of the course will focus on a number of other topics that pertain to the deployment of machine-learning based python services, as well as evaluation techniques and a select list of advanced topics in supervised learning. The class project will allow you to put all of the concepts that you will learn throughout the course into practice as you work on a real problem that can be addressed by the insightful use of data.

Important Notes

This term we will be using Slack for class discussions, announcements, the sharing of course material, and as the main point of contact for anything happening outside of our normal lecture time. The system is highly catered to getting you help fast and efficiently from classmates and your instructor. Rather than emailing questions to the teaching staff, I encourage you to post your questions to Slack and to become comfortable with using that tool (and setting up notifications for it) in your first week.

You should have received an invite to join our group via email, but you can also do so by using the following link: https://cs6220spring2024.slack.com/

Course Format & Methodology

This course runs for a total of 15 weeks starting on January 11. We will cover material pertaining to a number of concepts during our weekly lectures. Additionally, there will be weekly assignments aimed at reinforcing the concepts covered that week. **Please note that all assignment due dates and times are specified according to the Pacific Standard Time zone (PST);** plan to complete and submit all assignments accordingly. In addition, there will be optional online discussions related to current topics that will reinforce the material being presented in the course. The second half of the semester will focus on your project, and during that time weekly assignments will be replaced by project deliverables that will build up to a final write-up and group presentation.

Textbook & Materials

Below you will find a few ***suggested*** text books. Note that no readings or assignments will be assigned from these and they are listed here merely as reference material should you wish to acquire them as resources.

Python for Data Analysis (McKinney, 2022)

"Python for Data Analysis is concerned with the nuts and bolts of manipulating, processing, cleaning, and crunching data in Python. It is also a practical, modern introduction to scientific computing in Python, tailored for data-intensive applications. This is a book about the parts of the Python language and libraries you'll need to effectively solve a broad set of data analysis problems. This book is not an exposition on analytical methods using Python as the implementation language. Written by Wes McKinney, the main author of the pandas library, this hands-on book is packed with practical cases studies [sic]. It's ideal for analysts new to Python and for Python programmers new to scientific computing."

Python Data Science Handbook (VanderPlas, 2017)

"For many researchers, Python is a first-class tool mainly because of its libraries for storing, manipulating, and gaining insight from data. Several resources exist for individual pieces of this data science stack, but only with the Python Data Science Handbook do you get them all—IPython, NumPy, Pandas, Matplotlib, Scikit-Learn, and other related tools. Working scientists and data crunchers familiar with reading and writing Python code will find this comprehensive desk reference ideal for tackling day-to-day issues: manipulating, transforming, and cleaning data; visualizing different types of data; and using data to build statistical or machine learning models. Quite simply, this is the must-have reference for scientific computing in Python."

Data Science Projects with Python (Stephen Klosterman, 2019)

"Data Science Projects with Python is designed to give you practical guidance on industry-standard data analysis and machine learning tools, by applying them to realistic data problems. You will learn how to use pandas and Matplotlib to critically examine datasets with summary statistics and graphs, and extract the insights you seek to derive. You will build your knowledge as you prepare data using the scikit-learn package and feed it to machine learning algorithms such as regularized logistic regression and random forest. You'll discover how to tune algorithms to provide the most accurate predictions on new and unseen data. As you progress, you'll gain insights into the working and output of these algorithms, building your understanding of both the predictive capabilities of the models and why they make these predictions."

Course Outcomes

By the end of this course you will be able to:

- Recognize key principles, emergent methods, and applications of data science.
- Apply statistical methods and visualization to explore and prepare data.
- Given a dataset, perform data preparation steps prior to analysis.
- Compare and contrast classification and regression techniques.
- Apply appropriate performance evaluation criteria for comparing algorithms.
- Deploy a machine learning model as a service.
- Monitor the inner-workings of that service in real-time.

Participation and Engagement

Your participation in peer-to-peer activities, and your performance on assignments, serve as indicators of your level of engagement and effort. Frequent and varied opportunities to receive feedback, help, and/or clarification on course material from the instructor are provided throughout the term. Those students who struggle with the material but take advantage of self-checks and opportunities provided for instructor help and/or peer-to-peer mentoring, will be successful in this course.

Note that given the collaborative nature of our class, participation and engagement is even more important. To reflect that, 10% of your final grade will be assigned to this metric. In addition to the previously mentioned points, you are highly encouraged to participate during lectures, to join online discussions and external presentations, and to take advantage of that time to interact with your instructors and peers.

Communication / Submission of Work

Guidelines for completing and submitting each assignment are posted along with the assignment description in Slack. Please note that if you are unable to complete an examination within the period it is assigned, a documented compelling excuse (such as hospitalization) is required.

Course Activities and Assignments

This course includes the following required activities and assignments:

- **Assignments**: In the assignments, you will have the opportunity to put together the concepts that you learned throughout the lesson, the code snippets that you were provided, and your creative thinking to solve miniature problems that rely on toy datasets. Submission for the assignments will be made via sharable GitHub links. You will be graded on the application of the modules' topics, the completeness of your answers to the questions in the assignment description, and the clarity of your writing and code.
- **Group Project:** For the final project in this course, you will work in teams to explore a real-world data set. Your team will select a data source, apply the various techniques you

will learn in class, and present on your findings and process to the class. The topic that you choose is completely up to your group, and I encourage you to find subjects that seem particularly interesting to you and methods that you want to explore deeply as part of this project. Previous groups have presented on topics as diverse as basketball strategies or the existence of black holes. You will be working on this project throughout the course, with several opportunities for feedback from your instructor and each other.

Course Grading Scale

The grading scale will break down as follows:

A =	93–100%	C =	73–76%
A- =	90–92%	C- =	70–72%
B+ =	87–89%	D+ =	67–69%
B =	83–86%	D =	63–66%
B- =	80–82%	D- =	60–62%
C+ =	77–79%	F =	Below 60%

Grading/Evaluation Standards:

	Assignment	Description	Due in Modules	Points
1.	Participation	In-person / Slack / Zoom participation	Throughout	10%
2.	Quizzes	In class quizzes	Throughout	5%
3.	Assignments	Weekly individual homework	Throughout	30%
4.	Group Project Proposal	Form Groups and propose potential project ideas	Week 8	5%
5.	Group Project Presentation	Group presentation	Module 13	15%
6.	Group Project Report	Incorporate feedback from instructor and peers to create a 10-page report of your findings.	Module 14	35%

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.

Northeastern University Copyright Statement

This course material is copyrighted and all rights are reserved by Northeastern University. No part of this course material may be reproduced, transmitted, transcribed, stored in a retrieval system, or translated into any language or computer language, in any form or by any means, electronic, mechanical, magnetic, optical, chemical, manual, or otherwise, without the express prior written permission of the University.

Tutoring and Workshops by Global Learner Support (GLS)

Global Learner Support offers **one-to-one tutorials** for NU learners in the areas of academic writing, academic presentations, APA/MLA citation, English language conversation, and professional communication.

- To make a tutoring appointment, please visit the GLS booking page: https://gls.northeastern.edu/gls-tutoring/

Global Learner Support (GLS) also offers **monthly virtual and in-person workshops** on topics related to avoiding plagiarism, paraphrasing, APA/MLA guidelines, grammar and punctuation, academic presentations, writing professional emails, etc.

- Please visit https://gls.northeastern.edu/gls-workshops/ to register for upcoming workshops.

To view additional GLS services, visit their website at https://gls.northeastern.edu/

Wellness Specialist

Physical, mental, and emotional health are huge factors for success during your academic journey and out in the world. Northeastern is invested in your wellness and provides many on-campus, virtual, self-guided, and one-on-one resources so you can get immediate and ongoing support. Visit the Health and Wellness webpage for more information.

Student Affairs

Explore personal and professional development opportunities with Student Affairs. From student interest groups to professional development funding, they're here to assist you. Visit [Student Affairs Website] to learn more.