Northeastern University

Course ID: CS6220 Data Mining Techniques
Spring, 2019

Instructor: Reid Johnson
Teaching Assistant: TBD
Office Hours: TBD
Email: re.johnson@northeastern.edu

Course Description
This course will cover a variety of practical aspects of data mining, starting with data understanding and manipulation, which in turn will allow us to dive into supervised and unsupervised learning. The latter half of the course is concerned with evaluation techniques, and advanced topics in supervised learning such as ensemble methods. 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.

Course Format & Methodology
This course runs for a total of 15 weeks starting on January 10. Each week (or module), we will cover material pertaining to a small number of lessons, which will be organized around specific course topics, and contain readings and multimedia presentations, practice activities that you complete individually or in small groups, and assessments. **Please note that all 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.

Textbook & Materials
There is one suggested textbook for this course, from which relevant readings will be suggested as a supplement to class lectures. The suggested textbook is the following:

Introduction to Data Mining, 1st Edition (Tan, Steinbach, & Kumar, 2006)
"Introduction to Data Mining presents fundamental concepts and algorithms for those learning data mining for the first time. Each concept is explored thoroughly and supported with numerous examples. The text requires only a modest background in mathematics. Each major topic is organized into two chapters, beginning with basic concepts that provide necessary background for understanding each data mining technique, followed by more advanced concepts and algorithms."

Supplementary material for this textbook is available here.

Recommended Textbooks & Resources

Python for Data Analysis (McKinney, 2017)

"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."

Introduction to Data Analysis with Python on Lynda.com

"This online resource provides an opportunity to learn about the Pandas and NumPy libraries for python. Michele Vallisneri shows how to set up your analysis environment and provides a refresher on the basics of working with data containers in Python. Then he jumps into the big stuff: the power of arrays, indexing, and DataFrames in NumPy and Pandas. He also walks through two sample big-data projects: one using NumPy to analyze weather patterns and the other using Pandas to analyze the popularity of baby names over the last century. Challenges issued along the way help you practice what you've learned."

Python Data Science Handbook (VanderPlas, 2016)

"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."

Course Outcomes

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

- Recognize or describe key principles, emergent methods, and applications of data mining.
- 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 classification or regression techniques in a given set of circumstances.
- Apply appropriate performance evaluation criteria for comparing algorithms.
- Describe social and ethical implications of data mining.

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.

Communication / Submission of Work
Guidelines for completing and submitting each assignment are posted along with the assignment in the course site. 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:

- **Weekly Reading/Discussions:** Weekly readings and multimedia presentations provide the background knowledge, terminology, and practical examples you need in order to understand and correctly apply fundamental course concepts. You are responsible for completing the assigned textbook, workbook and other readings, for viewing the presentations and demonstrations, and for completing the concept checks included in the lessons. All provided material should be completed in the order in which they are presented, and by the due dates specified, within the weekly module.
- **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. Each student will be expected to setup a free GitHub account, where you will then create a repository that will host your assignment code files, which you will share with your instructor for grading. 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.

- **Midterm and Final Exam:** The midterm exam will cover concepts from roughly the first half of the course. The final exam will be cumulative, with content weighted so that roughly two-thirds of the exam covers concepts from the second half of the course.

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-class / online participation	Throughout	5%
2.	Quizzes	In-class quizzes	Throughout	5%
3.	Assignments	Assessed on your individual submission	Throughout	20%
4.	Group Project Proposal	From Groups and propose potential project ideas	Module 7	5%
5.	Midterm Exam	Covering concepts from roughly the first half of the course	Module 7	15%
6.	Group Project Presentation	Group presentation	Module 13	10%
7.	Group Project Report	Incorporate feedback from instructor and peers to create a 10 page report of your findings.	Module 14	25%
8.	Final Exam	Cumulative, but with roughly 2/3 covering concepts from the second half of the course	Module 15 (finals week)	15%

Class Schedule / Topical Outline

Please note: This outline may go through slight adjustments throughout the semester.

Module	Topics	Readings	Assignments
1	Introduction to Data Mining	Introduction to Data Mining: pp. 1–11	Assignment 0: Setting up your Python environment
2	Data Exploration	Introduction to Data Mining: pp. 23–36, 97–105, 105–131	Assignment 1: Creating Statistics and Visualizations
3	Data Quality and Standardization	Introduction to Data Mining: pp. 36–52, 701–705	Assignment 2: Data Preprocessing
4	Association Rules	Introduction to Data Mining: pp. 327–338, 350–352, 363–386	Assignment 3: Transaction Data
5	Clustering	Introduction to Data Mining: pp. 490–513 & 515–526	Assignment 4: Clustering Data Group Project: Forming groups and brainstorm ideas
6	Regression	Introduction to Data Mining: pp. 729-737	Assignment 5: Regression Group Project: Form groups and preliminary project ideas
7	Supervised Learning	Introduction to Data Mining: pp. 145–149, 223–238	Midterm Exam
8	Rule Learners and Decision Trees	Introduction to Data Mining: pp. 150–172, 207–223	Assignment 6: Classification
9	Evaluating Classifiers	Introduction to Data Mining: pp. 172–186. 294–304	Assignment 7: Model evaluation
10	More Evaluation Techniques and Concepts	Introduction to Data Mining: pp. 186–192	Assignment 8: Class project draft documents and data description

11	Ensemble Methods	Introduction to Data Mining: pp. 281–285	Continue work on project
12	Artificial Neural Networks	Introduction to Data Mining: pp. 246–256	Continue work on project
13	TBD		Continue work on project
14	Ethical Data Mining		Group Project Report, Final Exam
15			Group Project Presentation

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