CS6220: Data Mining

For all general course information such as credit hours, format, meeting times and location, please refer to the registrar system for the latest information.

Instructor Information: Dr. Mirek Riedewald
Office Hours, Email, TA: this information will be posted on Canvas

Following university policy, this course is currently scheduled as a regular in-person lecture on the Boston campus. Students are expected to attend all lectures in person and take the exam(s) in the classroom in person.

Please be aware of the following policies:

- There are **no** deadline extensions or make-up assignments/exams, except if you have a major emergency. You must provide evidence to claim such an emergency and you must inform the instructor *as soon as possible*. The following are examples for situations that do *not* qualify as emergencies: (i) I have job/co-op/internship interviews. (ii) My other course has an exam. (iii) My other course has a major homework or project deadline.
- We understand that some weeks are busier than others, but that's how things will be in your future job, too. By announcing deadlines well in advance, we give you the opportunity to plan and schedule your work accordingly. Make sure you start early so that you have the flexibility for dealing with unexpected issues.
- **Honor Code**: All students must adhere to the Northeastern University honor code available on the Northeastern web site and the graduate student handbook.
    - Please note that you are *not* allowed to share homework solutions with others or copy anybody else's homework entirely or in parts.
    - If you use someone else's code, text, etc, you must clearly indicate the copied material and properly cite the source. This also applies to material that you "slightly" modify. If in doubt, cite it and briefly explain or highlight how you modified it.
    - Please read carefully our course policy about the use of AI tools.

AI Policy for Assignments and Project: AI-based chatbots such as ChatGPT, Copilot, or Claude have become essential tools for software development in industry. We therefore want to encourage all students to use them in an *appropriate and responsible* manner. This means the following:

- Always keep in mind that if the AI can do all your work, then why should anyone hire you? Hence, we designed the course to teach you deeper understanding and problem-solving skills. The exam(s) will test those skills.
- We generally require you to design your own solution in "pseudo-code" without AI help. Then you can use AI to "fill in the syntax", i.e., to convert your solution description into actual code.

    - Note that "pseudo-code" here does not only refer to an algorithm, but also the description of major steps of a data-mining pipeline.
    - We will ask you to include in your assignment reports the "pseudo-code" and the major prompts you used for the AI tools.
- Example: Assume you have a CSV file that you want to split into training and test data to train a decision tree with scikit-learn.
    - Pseudo-code = you describe each step concisely and precisely.
        - Create a scikit-learn pipeline with the following steps.
        - Load the CSV file into a pandas dataframe, using the first row as the column names.
        - Process column X by applying normalization Y.
        - Split the dataframe randomly 80-20 into training and test data.
        - Train a decision tree with hyperparameter Z=z on the training data.
        - Evaluate the model's accuracy and ROC score on the test data and also report the confusion matrix.
    - You report the corresponding prompts, which should be similar to the pseudo-code:
        - "You are given CSV file my_data.csv. Create a scikit-learn pipeline in Python. The first step loads my_data.csv into a pandas dataframe, using the names in the first line of the file as the column names."
        - "Add a step that does XYZ…"

Course Prerequisites and Description: See the official information in the course catalog.

Course Format & Methodology: This course contains online content accessible through the Northeastern University Canvas system. Homework and project solutions will be managed in Gradescope. For source-code management, the instructor will create GitHub Classroom repositories. **Please note that all due dates and times are specified according to the local Boston time (Eastern US time zone)**.

Recommended Textbook & Materials: The course is designed to be self-contained, but at times you may want to consult additional material, e.g., to clarify details of a method. Any standard textbook on data mining and machine learning would be a great resource—pick one whose style you like. There are also great videos and free lectures on the Internet. For those, like for discussion groups and when asking AI chatbots, please use caution and common sense. Not everything on the Internet is correct…

Course Outcomes: This course has the following main objectives and content:

- You can describe the main steps of a typical data mining pipeline from pre-processing to result evaluation and visualization.
- You can implement your solution in a common environment, possibly with the help of AI tools to "fill in the syntax".
- You can critically evaluate the success of your approach.
- You can describe important methods for supervised and unsupervised learning, and you can implement some of their functionality.

    - What problem does it solve (well)?
    - How does it work?
    - How do I effectively tune its hyper-parameters?
    - Will it scale to big data?
- Important principles covered (may change at the discretion of the instructor):
    - Bias-variance tradeoff and overfitting
    - Success measures for data-mining methods
    - Basic statistics and linear algebra
- Techniques covered (may change at the discretion of the instructor):
    - Data summarization
    - Classification and prediction:
        - Decision Trees
        - Nearest Neighbor
        - Naïve Bayes and Bayesian Belief Networks
        - Perceptron, neuron, and Artificial Neural Networks
        - Support Vector Machines (SVM)
        - Ensemble methods
        - Deep Learning and LLMs
    - Itemset and sequence mining:
        - Apriori
        - FP-growth
        - GSP
        - PrefixSpan
    - Clustering
        - K-Means
        - Hierarchical clustering
        - DBSCAN

Please note that AI is a rapidly evolving field, continuously churning out new methods applicable to data mining. This course is designed to give you a solid understanding of representative methods and the principles behind them. This will build a foundation, enabling you to more easily apply new methods and understand their tradeoffs.

Participation and Engagement: Your presence in peer-to-peer activities serves as an indicator of your level of engagement and effort throughout the course. Frequent and varied (e.g., synchronous/asynchronous/face-to-face) opportunities to receive feedback, help, and clarification on course material from the instructor are provided throughout the term. The following activities count towards class participation:

1. Asking or answering questions in class.
2. Submitting solutions for in-class exercises when requested by the instructor.
3. Answering questions or posting relevant information on the discussion boards. Participation points are awarded based on quality and quantity of contributions.

Communication/Submission of Work: Make sure you receive course-related announcements the day they are made. Guidelines for completing and submitting each

assignment are posted along with the assignment. Late and early homework submission policies will be announced with the individual assignments.

Course Activities and Assignments:

- **Exam (tentative information; will be finalized by end of week 5 of the semester)** You will complete an exam, possibly two exams, designed to test your understanding of the course concepts. The exam is **closed-book**, i.e., you cannot bring any material. In fact, you can only bring pencils/pens and (optionally) some blank pages for scratch paper. Students must be present in the lecture room for the exam. Exceptions are possible for students with disabilities who can provide an official letter from the corresponding Northeastern office at the beginning of the semester.
- **Homework/project** You will complete multiple homework assignments that give you the opportunity to practice the concepts you learn. More information about these assignments and the course project is or will be available on Canvas and Gradescope.

Course Grading Criteria:

- Participation: 15%
- Exam: 60%
- Homework/project: 25%

Class Schedule / Topical Outline: (This schedule is subject to updates.)

Individual homework assignments: weeks 2 to 9

Project: weeks 10 to 15

Exam: around week 12

Topics:

- Overview, introduction: 1 week
- Pre-processing: 1 week
- Prediction techniques: 7 weeks
- Itemset and sequence mining: 2 weeks
- Clustering: 2 weeks
- Advanced topics: 1 week

How to Succeed in this Course

Data mining software and AI tools greatly simplify the process of creating data mining pipelines. However, you may find it more challenging and insightful to understand the mathematical and algorithmic principles behind a method, when it is most effective, and how it scales to big data. This deeper understanding is exactly what the lectures focus on. Therefore, it is essential that you attend all lectures and participate in online discussions.

Homework is designed to help you understand the material and prepare for the exam. The following often works well:

1. When going through the material covered in a lecture, make notes about questions you have or about material you find difficult to understand. Then share these questions through the online forum or in class.
2. Consult additional resources (see above) for details and alternative presentations of relevant material. While the instructor relies on teaching methods that work well "in general", everyone learns differently, and you may find that some textbook or video posted by another data mining expert may help you deepen your understanding.
3. Use old-school in-person interactions: Try to explain the material to a friend. This way you can better judge if you understand it. Once you identified things that need clarification, try to find the answer yourself. If you cannot find the answer with reasonable effort, ask others for help (online discussion forum, office hours, and in-class discussions).
4. Start working on homework assignments as soon as they come out. This way you have time to ask questions and get help.

Is This the Right Course for You?

The main point of the course is *not* "syntax", i.e., going through a bunch of method calls in some established data mining or AI package. Nowadays you do not even need a computer science degree to use a modern AI chatbot to build a ready-to-run data mining pipeline.

Our main goal is to teach deeper understanding and problem-solving skills: What is the purpose of the various methods? How do they work and when are they effective and efficient? Which method scales when I have a lot of data records and/or these records have a lot of features? How do I know that I have found good values for the hyperparameters? Answering these questions will often rely on mathematical tools like basic linear algebra and basic statistics, but also concepts from data structures and algorithms.

For practical assignments, we want to give you the flexibility to work in your favorite environment—be it the commandline/terminal, Jupyter notebooks or a cloud service. However, we cannot provide support for all these environments. Our default environment of choice will be Jupyter notebooks with popular Python libraries like scikit-learn. And while you can pick your favorite programming language to implement aspects of a technique to better understand it, in class, we will often use Python-style examples.

Special Accommodations: If you have specific physical, psychiatric or learning disabilities that may require accommodations for this course, please contact Northeastern's Disabilities Resource Center (DRC) at (617) 373-2675. The DRC can provide you with information and assistance to help manage any challenges that could affect your performance in the course. 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.

If the Disability Resource Center has formally approved you for an academic accommodation in this class, please present the instructor with your "Professor Notification Letter" *during the first week of the semester*, so that we can address your specific needs.