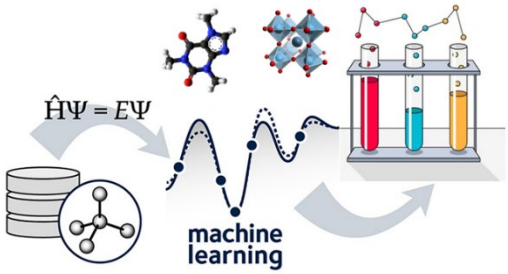
Machine Learning for Molecular Sciences

Spring 2021

Department of Chemistry

Carnegie Mellon University



Course Info

Instructor: Prof. Olexandr Isayev

Office: Mellon Institute 511A

Email: olexandr@cmu.edu

Email Policy: Assume a reply will require about 24 h and 48 h for weekends. For grading questions, please contact me through your Andrew account, not yahoo, gmail, etc. The University forbids faculty to reply to non-Andrew accounts with information concerning grades.

Class schedule:

MW, 4:00- 5:20 PM (Remote)

Piazza: https://piazza.com/cmu/spring2020/09860

Canvas: https://canvas.cmu.edu/courses/13278

Office Hours: By appointment.

Course description

The emergence of contemporary artificial intelligence (AI) and machine learning (ML) methods has the potential to substantially alter and enhance the role of computers in science. At the heart of ML applications, lie statistical algorithms whose performance, much like that of a scholar, improves with training. There is a growing infrastructure of machine learning tools for generating, testing and refining scientific models. Such techniques are suitable for addressing complex problems that involve vast combinatorial spaces or complex processes, which conventional procedures either cannot solve or can tackle only at great computational cost.

The purpose of this course is to provide a **practical** introduction to the core concepts and tools of machine learning in a manner easily understood and intuitive to STEM students. The course begins by covering fundamental concepts in ML, data science, and modern statistics such as the bias-variance tradeoff, overfitting, regularization, and generalization, before moving on to more advanced topics in both supervised and unsupervised learning. Topics covered in the course also

include ensemble models, neural networks, modern deep learning, embedding, clustering and data visualization. Throughout the course, we emphasize application of ML methods to chemical, physical and biological data. A notable aspect of the course is the hands-on use of Python Jupyter notebooks to introduce modern ML/statistical packages.

Learning objectives

Upon successful completion of this course, students will have achieved the following learning objectives:

- Know how to explore and visualize chemical data
- Compare and contrast different types of chemical data and molecular representations.
- Understand core components of data analytics pipeline: visualization, exploratory data analysis, classification, regressions, prediction etc.
- Be able to analyze chemical data using a variety of machine learning approaches.
- Implement and analyze well-known existing ML algorithms.
- Integrate multiple components of practical machine learning in a single system: data preprocessing, learning, regularization, model selection and be familiar with programming tools to accomplish it.
- Hands on experience with real-world cases on how ML could address challenges in molecular sciences.

Course requirements

09-860 is designed for PhD and master students but is open to senior–year undergraduate students too. Prerequisites: (09-231 or 09-344) and (15-110 or 15-112)

This is practical, application oriented course, requiring skill in algorithmic problem solving. We will use Python based data science tools and libraries. Prior programming experience with Python is needed. Students entering the class are expected to have a pre-existing working knowledge of probability, linear algebra, statistical thermodynamics and quantum mechanics.

Reading

The content of this course does not follow any textbook. Readings will be provided on Canvas portal and lectures. The purpose of the readings is to provide a broader and deeper foundation than just the lectures and assessments. The readings for this course are required. We recommend you read them after the lecture. Sometimes the readings include topics that are not mentioned in lecture; but we still encourage you to read those portions. Many authors and publishers allow free use of copyrighted materials for educational purpose. Using material from the existing literature rather than a textbook is also intended to prepare you for your future professional careers. You should be developing skills in reading a variety of materials and assessing usefulness.

Course Outline

The following (approximately) outlines the course material that we will cover through the course. This schedule is subject to change.

- Chemical data and chemical descriptors, molecular similarity
- Databases of materials and chemical data
- RDKit python library for chemical data
- Exploratory data analysis and visualization
- Unsupervised learning, clustering, dimensionality reduction
- Supervised learning, model training and evaluation
- Linear and nonlinear models
- Classification, SVM, kernel methods
- Decision trees and RF
- Probabilistic methods
- Deep learning and neural networks.
- Convolutional neural networks (CNN)
- Sequence modeling, recurrent neural networks (RNN)
- Learnable representations, embeddings
- Generative models
- Advanced applications of ML in chemistry and molecular sciences

Grading

The requirements of this course consist of attendance, participating in class, projects, homework assignments, and readings. The grading breakdown is the following:

5% for attendance

5% for class participation

50% for homework assignments

40% for final open-ended project:

On Piazza, the *Top Student "Endorsed Answer" Answerers* will earn bonus points

Grading Policies

Late-work policy/Flex days: You will have 5 flex days for the entire semesters that you can use for homework submission. You can chose to divide the days up the way you want. After that, submissions will not be accepted.

Re-grade policy: Requests for re-grades must be submitted within 1 week of receiving the grades assignment, paper, or test.

Attendance policy: Attending lectures is mandatory and class attendance will be kept. Each student is allowed two absences (no questions asked) for the entire semester.

Class Participation policy: Class participation will be measured by the level of the student's engagement with Canvas/Piazza discussion forums and during the in-class quizzes.

Group work policy: You are encouraged to work in groups on the graded homework assignments. This means you can work together on the problem, but you must write your own solution. On your homework, you must list the members of your work group as the comment in the code.

Peer Evaluation and Feedback: For both projects we will be using a writing tool Peer Evaluation and Feedback located within Canvas. It provides the opportunity for you to both provide and receive peer feedback with your projects.

Homework

Homework assignments will be due almost week. Assignments will include *programming* components, where you must complete designated tasks, and *kaggle* component where you compete with your peers (or a short write-up if there are no scores). The programming assignments will emphasize understanding of ML methods, implementing ML pipelines, and experimental design. Results of the execution of your ML pipeline (usually as CSV file) will be submitted to the Kaggle (kaggle.com) for automatic evaluation. Data sets will be provided. The write-up part will focus explaining the observed results .

Projects

Final Project. You may work in teams of 2 people or individually. A final project in which you solve a problem of relevance to chemistry (or broadly molecular sciences) and gain experience with the full ML pipeline. This is an open-ended project, where you could your data and select a method of analysis, apply it, and draw conclusions about the data. Data sets and problems could be provided, but students are encourage to use their own projects. You will be graded by the course instructor and other students taking the course.

During the last week of classes, you will present your project to the class.

Laptops and Mobile Phones.

This is a technology-oriented course but there is a time and place to use technology. As the need arises, we will have hands on class sessions where you will need to use your laptop. Other devices (iPads, smart phones etc.) tend to hinder classroom participation and discussions. Hence, unless explicitly stated otherwise, please close or turn off all such devices when in class.

Student recording of class

Not allowed.

Online Resources

Some of the homework question you receive might have solutions online. Looking up the answers to these homework questions is not allowed. Similarly, looking up code for a problem is not allowed. Sometimes, you might need help with a small portion of the code (for example sorting or specific function), such basic things that are not in relation to understanding the material are allowed.

Piazza

We will use Piazza for all course discussion. Questions about homeworks, course content, logistics, etc. should all be directed to Piazza. If you have a question, chances are several others had the same question. By posting your question publicly on Piazza, the course staff can answer once and everyone benefits. If you have a private question, you should also use Piazza as it will likely receive a faster response.

Canvas

Lecture slides, assignments, and syllabus updates will be posted on the Canvas. Lecture slides may be posted before or after the lecture, depending on the nature of the content. Canvas might be also used for quiz-style problems, code and write up submissions.

Kaggle

Kaggle for education classroom (kaggle.com) is used for autoscoring homework submissions and peer competition. Students are expected to perform equally of better to a specified baseline to receive perfect grade. Top submissions might be eligible for bonus points.

Poll Everywhere

Lectures will provide opportunities to engage you in applying course concepts on various levels, problem solving, and assessing your understanding of course material. We will use polleverywhere.com polls for short unannounced quizzes to measure class participation.

Academic Integrity Policy

Academic integrity refers to the implicit commitment that every member makes to all others in the community to practice those principles that underlie the mission of the university and define academic integrity. These are: honesty and good faith; clarity in the communication of core values; professional conduct of work; mutual trust and respect; and fairness and exemplary behavior. In this course, cheating will not be tolerated and could lead to expulsion from the university. Please remind yourself of the policy here: http://www.cmu.edu/academic-integrity/

There is a zero tolerance policy on cheating. If you are found to be cheating on an exam, you will automatically receive a failing grade for the exam (which cannot be dropped) and you will be reported to

the Dean of Students for further disciplinary action. This usually means an academic review board meeting and possible suspension/expulsion from the university.

Policy on Course Accommodations

If you wish to request an accommodation due to a documented disability, please inform the instructors and contact: Disability Resources, 102 Whitfield Hall, 412-268-2013, (access@andrew.cmu.edu) as soon as possible. For ongoing documented classroom accommodations, one week's notice is required prior to each exam.

Statement on student wellness.

As a student, you may experience a range of challenges that can interfere with learning, such as strained relationships, increased anxiety, substance use, feeling down, difficulty concentrating and/or lack of motivation. These mental health concerns or stressful events may diminish your academic performance and/or reduce your ability to participate in daily activities. CMU services are available, and treatment does work. You can learn more about confidential mental health services available on campus at: http://www.cmu.edu/counseling/. Support is always available (24/7) from Counseling and Psychological Services: 412-268-2922.

Let's have a fun and productive course!