Introduction to Machine Learning in the Physical Sciences: Student Experiences

Simon J. L. Billinge[1], Sanat Kumar[2], Sarah Alsharif[1], Candice Chiu[4], Joshua Crawford[2], Elizabeth Katzman[2], Alperen Koc[3], Tymon Nieduzak[3], Benjamin Schwartz[2], Zoe Zachko[1], Tieqiong Zhang[1], Grant Zhou[3]
[1]Applied Physics and Applied Mathematics, [2]Chemical Engineering, [3]Civil Engineering, [4]Biomedical Engineering

Overview and Motivation

What: A one-semester accelerated machine learning (ML) course applied to STEM research problems.

Participants: graduate students and physical science undergraduates interested in applying ML to their research, this semesters cohort is a mix of Materials Science, Chemical, Biomedical, and Civil Engineering.

Motivation: Building an understanding of machine learning through interaction with Edexes, then applying these new skills to research projects in each students respective disciplines.

Scope: unsupervised and supervised learning, decision trees, logistic regression, neural nets, etc., applied to physical science research problems

Prerequisites: Physical science knowledge and basic programming. Prior ML knowledge is NOT required.

Learning Materials: EdEx's

What: EdEx stand for educational examples. They are hands-on tutorials designed to walk students through machine-learning applications from published research.

EdExes:

- integrate with the theoretical lecture component of the course
- based on published research conducted primarily here at Columbia Engineering
- Students work together in pairs or teams to solve the given problem. We are working primarily in Jupyter Notebooks. At the end of each Edex, one group is chosen to present their work to the class.

Tools introduced: Jupyter notebooks, conda, git, scikit-learn, keras, tensorFlow.

Status and team

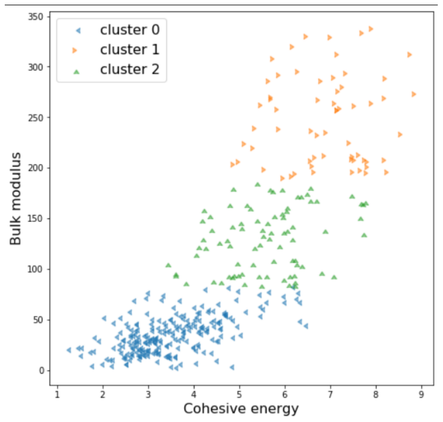
Status: the course is currently being taken by 10 students and is under development as MSAE4990.

Funded by:

COLLABORATORY AT COLUMBIA UNIVERSITY

The Students: The students are a mix of Masters candidates, PHD candidates and undergraduate students across four different disciplines. The three primary research areas are EV safety, batteries, and ligament repair. The class is primarily discussion based. We are enthusiastic about sharing our individual passions and improving our problem solving capabilities so we can all be more effective communicators and engineers.

Edex 1: Predicting melting temps



Goal: Predict the melting temperature of inorganic materials given just the chemistry of the constituents.

Problem: We need quick and reliable low-cost predictions of melting temperature for metal extraction. Chemical variability means that one model does not work over high-variance data.

Student Experience: A variety of different methods were used, including different clustering methods and SVR in an attempt to boost performance.

Model: k-Means clustering, various other clustering and regression models implemented via scikit-learn

Gharakhanyan, V., Urban, A., In preparation (2023).

Edex 2: Space Group Classifier

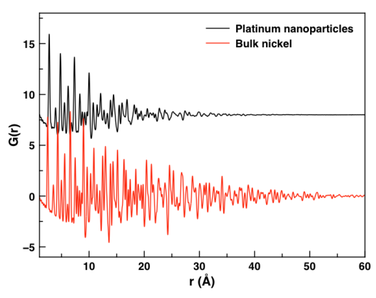


Image source: Billinge Group

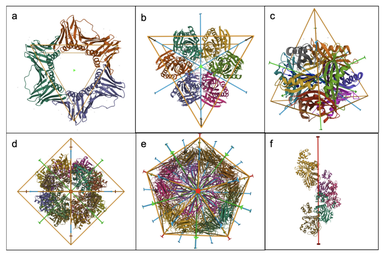


Image source: RCSB Protein Databank

Goal: predict the space group (SG) of a crystal structure given the measured atomic pair distribution function (PDF).

Problem: SG encodes structural symmetries of atomic arrangements in a material. PDF is an x-ray measurement of the material. There is no direct way of getting SG from the PDF.

Motivation: The ML model can quickly predict the most likely space groups and give insights into the structure-property relationships.

Model: convolutional neural networks.

Training: 40,000 PDFs that are calculated from 8 of the most common space groups.

Liu, C. H., et al. Acta Crystallographica Section A: Foundations and Advances, 75(4), 633-643.

Class Outcomes

Adapting Course Materials: Students are currently working on developing new Edexes, as well as providing feedback on the course. The new Edexes are being developed to acknowledge problems that have arisen as we learn about ML. These new Edexes help streamline the class experience and add to its impact. Some examples of new material are unit testing and package development.

ML Solutions to Research: Students present their active research. Currently, we are brainstorming ML solutions to some of the obstacles in the projects. We will then narrow down which problem to investigate further. The goal is to work together as a class to apply our new found knowledge from the Edexes to a relevant problem and publish a paper.

Further Machine Learning Knowledge: Throughout the rest of the term, we plan to expand our ML skill set, and continue to apply it to Materials Science problems. An upcoming Edex is focused on membrane permeability for individual gas separation. This will teach us about permutation importance and Shapley Values.

Community Focus: This class encourages open collaboration between disciplines, through extensive classroom participation and interaction with the larger ML community online. This is a course designed to be accessible and engaging for anyone interested in learning a new problem solving framework.