Machine Learning in Chemistry

Instructor: Dr. Hao Li (Technical University of Denmark)

Email: haoli@dtu.dk

Tel: 45-24595969

Course Description:
The tremendous success of artificial intelligence (AI) in applications ranging from self-driving cars to data mining in social media has brought much attention to the use of machine learning (ML) methods in modern technology. Since the original studies by McCulloch and Pitts in 1943, a diverse range of ML techniques have been developed and applied in both scientific and industrial applications. In recent years, state-of-the-art applications of ML, e.g., AlphaGo, newly developed ML techniques and algorithms, e.g., deep learning and graphical processing units, and the availability of open databases have led to a surge of interest across the scientific community into the practical applications of ML and data analytic approaches. In many fields of chemistry and materials science, the use of ML is now becoming commonplace. This course covers the essential knowledge of programming, data analysis, data processing, ML modeling and validation, and how they can be used in chemical applications.

Key References:
1) *Chemoinformatics: Basic Concepts and Methods,* edited by Engel and Gasteiger, (2018)
2) *Neural Networks for Chemist*, edited by Zupan and Gasteiger, (1993)
3) Müller et al., Assessment and validation of machine learning methods for predicting molecular atomization energies, *Journal of Chemical Theory and Computation,* 2013.
4) On-line tutorials video from IPAM workshops,
visit: http://www.ipam.ucla.edu/programs/workshops/

Exams:
Computational exercise: ~60%
Computational exercises include the practical use of programming languages such as Python and Mathematica for statistical analysis, chemical data collection and processing, and machine learning modeling, validation, and prediction. Examples and notes of the use of programming and data-analysis packages will be provided. Detailed documentations, instructions and code templates will be provided.

Final projects & presentation: ~40%
Based on different primary interests, students will be divided into several subgroups to complete the final projects. Some potential final project topics will be provided for students. This is an important assessment of a student's problem-solving capacity, scientific communication skills, critical thinking capacity, and self-study ability. The solutions for each project are all within the course content, which include data collection and analysis, machine learning modeling and validation, model parameter optimization, and predictions. Each student should submit a write-up with supplementary materials by the end of the due date. On the last day of the class, every student should deliver a 10-15 min presentation which includes:

1. The background of the project
2. The applied methodologies

3. Results and discussion
4. Conclusion

Lecture Outlines

1) Intro to programming I: knowledge of common programming languages for ML applications (e.g., Python and Mathematica)

2) Intro to programming II: applications of programming to fulfill simple computational tasks

3) Basic concepts in ML: background & overview of ML and data mining

4) Intro to Chemical database: an overview of both experimental and computational chemical database

5) Data collection & processing: how to collect, filter, and store data from a chemical database

6) Statistical & feature analysis: feature analysis to select influential independent variables and relevant data normalization

7) Intro to Algorithms & ML backends I: introduction to the most widely used ML methods such as artificial neural networks, decision trees, support vector machine, and other kernel-based methods

8) Intro to Algorithms & ML backends II: introduction to common ML backends such as TensorFlow, Keras, PyTorch, Scikit-Learn, and relevant packages; introduction to advanced ML methods such as deep learning

9) ML Modeling: ML modeling application based on ML backends

10) Model validation: applications of cross-validation and sensitivity tests to avoid under- and over-fitting, with particular emphasis on model hyper-parameter adjustment

11) Model prediction: applications of a well-trained ML model for practical predictions, including interpolation and extrapolation

12) Intro to AI-assisted quantum chemistry: introduction to the ML methodologies to fit the potential energy surfaces from an *ab initio* database

13) Final project discussion I: introduction to the potential final projects

14) Final project presentation II: discussion with students one-by-one about final project assignments

15) Final project presentation.