UTC-IASE
UNIVERSITY OF CONNECTICUT

UCONN
SCHOOL OF ENGINEERING

CSE/SE 5095: Machine Learning for Physical Science

Course Instructor: **Qian Yang, Ph.D.**

Catalog Description. 3 Credits. This course is designed to provide students with foundational knowledge of applied aspects of machine learning, including methods for handling uncertain, small, and imbalanced data; feature selection and representation learning; and model selection and assessment. Students will also gain exposure to state-of-the-art research on interpretability of machine learning models, stability of machine learning algorithms, and meta-learning. Topics will be discussed in the context of recent advances in machine learning for materials, chemistry, and physics applications, with an emphasis on the unique opportunities and challenges at the intersection of machine learning and these fields. Some basic familiarity with machine learning is assumed as a prerequisite.

Pre-Requisites. Students should be familiar with basic concepts in machine learning, linear algebra, optimization, and statistics (optional supplementary material will be provided for review). A background and interest in applications in the physical sciences is preferable but not required.

Intended Audience. The course is designed for all graduate students in engineering.

Course Delivery Method. The course will be offered online, asynchronously, in small recorded modules according to the course syllabus. Direct and live communication with the instructor will be available once a week for discussion and student presentations (weeks 3 – 14).

Anticipated Student Outcomes. By the end of SE 5095, a student will be able to:

(1) Evaluate the efficacy of using machine learning to solve a particular problem in the physical science domain.
(2) Implement strategies for dealing with small, noisy, and/or imbalanced data.
(3) Use good design principles for feature engineering to construct features.
(4) Describe different deep learning strategies for representation learning.
(5) Choose appropriate methods for model selection and model assessment.
(6) Evaluate the results of machine-learned models reported in the literature.

Course Organization. The course is organized into four learning modules.

(1) Machine Learning Overview
(2) Data

(3) Features
(4) Evaluation

The structuring of these learning modules into 14 lectures of a one semester course, along with the topics and references, is described in the following:

Course Outline

----------Module 1: Machine Learning Overview----------

Lecture 1: Machine Learning Overview

Topics:

- What types of problems is machine learning best suited for? When should we not use machine learning?
    - main themes in current research using machine learning for physical science: predicting structure-property relationships; speeding up existing computational methods; automating analysis of experimental data; etc.
    - advantages/disadvantages of scientific machine learning
- Brief review of different types of machine learning algorithms
    - supervised, semi-supervised, and unsupervised learning
    - generative vs. discriminative models
    - clustering, classification, regression
    - dimensionality reduction
    - reinforcement learning
    - probabilistic graphical models
- Components of machine learning

----------Module 2: Data----------

Module Application: Learning Structure-Property Relationships for Molecules and Materials

Lecture 2: Sample Complexity

Topics:

- generalization error; empirical error
- approximation vs. estimation error
- PAC learning framework
- Rademacher complexity
- VC dimension
- generalization bounds

- non-uniform learnability
- non-I.I.D. data samples

Lecture 3: Small Data

Topics:

- constraints
- data augmentation
- transfer learning
- active learning

Lecture 4: Noisy Data

Topics:

- noisy features vs. noisy labels
- generative models for noise
- prediction of missing features
- ensemble-based filters for eliminating noisy samples
- anomaly detection
- noise-tolerant learning methods
- multi-fidelity methods
- regularization

Lecture 5: Imbalanced Data

Topics:

- re-sampling methods
  - over- and under-sampling methods
  - ensemble methods
- weighted cost functions
- modified evaluation metrics
- evaluating model performance with imbalanced datasets

----------*Module 3: Features*----------

Module Application: Computer Vision and Scientific Imaging

Lecture 6: Feature Engineering & Selection

Topics:

- design principles for feature engineering

- categorical features
- feature selection
    - regularization
    - backward elimination/forward selection
- decision trees

Lecture 7: Dimensionality Reduction

Topics:

- motivations: computational; visualization; feature extraction
- Principle Components Analysis
- nonlinear dimensionality reduction; t-SNE
- random projections; Johnson-Lindenstrauss lemma
- sufficient dimension reduction
- topological data analysis

Lecture 8: Representation Learning

Topics:

- autoencoders
- variational autoencoders
- restricted Boltzmann machines
- word2vec
- convolutional neural networks

Lecture 9: Generative Adversarial Networks

Topics:

- theory of GANs
- implementation challenges
- BiGAN for adversarial feature learning
- constrained GANs
- discrete GANs
- semi-supervised learning with GANs

Lecture 10: Time Series

Topics:

- hidden Markov models
- recurrent neural networks

----------Module 4: Evaluation----------

Module Application: Machine Learning for Molecular Dynamics and Turbulence Modeling

Lecture 11: Model Selection & Assessment

Topics:

- model selection vs. model assessment
- splitting data into training, validation, and test sets
- cross-validation and nested cross-validation
- hyperparameter tuning
- bootstrap
- performance metrics

Lecture 12: Stability

Topics:

- relationship between stability and generalization
- using stability to evaluate algorithms

Lecture 11: Interpretability

Topics:

- definitions of interpretability
- design principles for interpretability
- criticism for interpretability
- interpretable representation learning
- concept activation vectors

Lecture 11: Meta-Learning

Topics:

- one-shot and few-shot learning
- learning across domains
- model-agnostic model learning
- learning to optimize

Useful Reading. Reading lists will be provided for each module on the course website. There is no formal textbook for this class.

Copyright. Copyrighted materials within the course are only for the use of students enrolled in the course for purposes associated with this course and may not be retained or further disseminated.

Grading. Grading of the course will be based on quizzes (24%), a paper presentation (26%), and a final course project (50%). There will be one short multiple-choice quiz based on the video lectures in each of weeks 2-13. Additionally, each student will be required to record (voice or video) one 15-minute paper presentation on a paper applying machine learning to a scientific domain, to be shared with the class in weeks 3-14. The paper presentation grade will include participation in an online class discussion forum. A detailed rubric and recommended list of papers to choose from will be provided on the course website; other papers may be presented with instructor permission. The final course project will focus on applying machine learning methods to address a problem in the physical sciences. The course project will be graded on the basis of a project proposal, midterm report, and final report. A detailed description and rubric will be provided on the course website.

Grade	Letter Grade	GPA
97-100	A+	4.3
93-96	A	4.0
90-92	A-	3.7
87-89	B+	3.3
83-86	B	3.0
80-82	B-	2.7
77-79	C+	2.3
73-76	C	2.0
70-72	C-	1.7
67-69	D+	1.3
63-66	D	1.0
60-62	D-	0.7
<60	F	0.0

Due Dates and Late Policy. All course due dates are identified in the Course Schedule. Deadlines are based on Eastern Standard Time; if you are in a different time zone, please adjust your submittal times accordingly. The instructor reserves the right to change dates accordingly as the semester progresses. All changes will be communicated in an appropriate manner.

Student Conduct: http://www.dosa.uconn.edu/student_code.html. Students are responsible for adherence to the University of Connecticut student code of conduct. Pay attention to the section on Student Academic Misconduct, "Academic misconduct is dishonest or unethical academic behavior that includes, but is not limited, to misrepresenting mastery in an academic area (e.g., cheating), intentionally or knowingly failing to properly credit information, research or ideas to their rightful originators or representing such information, research or ideas as your own (e.g., plagiarism)." Examples of academic misconduct in this class include, but are not limited to: copying solutions from the solutions manual, using solutions from students who have taken this course in previous years, copying your friend's homework, looking at another student's paper during an exam, lying to the professor or TA and incorrectly filling out the student workbook.

Attendance. Students should make every effort to attend the live sessions and to talk with students in the class discussion forum. It is practically impossible to follow the class if classes are missed.

Absences. Students involved in official University activities that conflict with class time must inform the instructor in writing prior to the anticipated absence and take the initiative to make up missed work in a timely fashion. In addition, students who will miss class for a religious observance must "inform their instructor in writing within the first three weeks of the semester, and prior to the anticipated absence, and should take the initiative to work out with the instructor a schedule for making up missed work."

Adding or Dropping a Course. If you should decide to add or drop a course, there are official procedures to follow:

- Matriculated students should add or drop a course through the Student Administration System.
- Non-degree students should refer to Non-Degree Add/Drop Information located on the registrar's website.

You must officially drop a course to avoid receiving an "F" on your permanent transcript.

Simply discontinuing class or informing the instructor you want to drop does not constitute an official drop of the course. For more information, refer to the online Graduate Catalog,

Academic Calendar. The University's Academic Calendar contains important semester dates.

Students with Disabilities. Students needing special accommodations should work with the University's Center for Students with Disabilities (CSD). You may contact CSD by calling (860) 486-2020 or by emailing csd@uconn.edu. If your request for accommodation is approved, CSD will send an accommodation letter directly to your instructor(s) so that special arrangements can be made. (Note: Student requests for accommodation must be filed each semester.)

Course Schedule.*

Date[1]	Topic	Module No	Project Dates
Aug 26	*Lecture 1: Review of Machine Learning Algorithms*	1	
Sept 2	*Lecture 2: Sample Complexity*	2	
Sept 9	*Lecture 3: Small Data*	2	First week of paper presentations
Sept 16	*Lecture 4: Noisy Data*	2	
Sept 23	*Lecture 5: Imbalanced Data*	2	Project Proposal Report
Sept 30	*Lecture 6: Feature Engineering & Selection*	3	
Oct 7	*Lecture 7: Dimensionality Reduction*	3	
Oct 14	*Lecture 8: Representation Learning*	3	
Oct 21	*Lecture 9: Generative Adversarial Networks*	3	
Oct 28	*Lecture 10: Time Series*	3	Project Mid-Term Report
Nov 4	*Lecture 11: Model Selection & Assessment*	4	
Nov 11	*Lecture 12: Stability*	4	
Nov 18	*Lecture 13: Interpretability*	4	
Dec 2	*Lecture 14: Meta-Learning*	4	Project Final Report

* Schedule is tentative and may change.

1 Date indicates release of lecture modules.

Instructors' Contact Information:

- Qian Yang: qyang@uconn.edu
- Office Hours: TBA

Helpful Links:

- Virtual Computer Lab at UConn: http://skybox.uconn.edu/
- Course Material: https://lms.uconn.edu
- Institute for Advanced Systems Engineering: http://www.utc-iase.uconn.edu/