Purdue University
School of Chemical Engineering

ChE 597: Data Science in Chemical Engineering

Spring 2021
Tues/Thurs, 9:00-10:15am, ME 2061

Instructors: Brett Savoie (FRNY 2043A, bsavoie@purdue.edu)
Administrative Assistant – Jason Thorpe – FRNY 2043
Office Hrs. Thurs. 5-6 pm (https://purdue.webex.com/meet/bsavoie)

Supplemental Texts (optional):

"Machine Learning: A Probabilistic Perspective"
K. P. Murphy; MIT Press, 2012
A comprehensive textbook with many advanced topics. Where applicable all algorithms are derived or discussed in the general context of probability theory.

"The Elements of Statistical Learning"
T. Hastie, R. Tibshirani, J. Friedman; Springer, 2017
A comprehensive text with derivation and statistical treatment and many common algorithms. Covers similar ground to Murphy, but many topics have complementary treatment.

"Python Machine Learning"
S. Racha, V. Mirjalili; Packt Publishing, 2017
A beginner text with lots of python examples.

"Introductory Lectures on Convex Optimization"
Y. Nesterov; Springer, 2004
A classic text by a central figure in numerical optimization.

"Forecasting: Principles and Practice"
R. Hyndman, G. Athanasopoulos; oTexts, 2018 (https://otexts.com/fpp2/)
An online and free introductory textbook on time-series prediction. Examples are in R and with a focus on business examples, but the treatment is comprehensive and great for the price.

Objectives

The intent of this course is to present data analysis and machine learning from a practical perspective focused on applications, use-cases, and the limitations of various approaches to problems in chemical engineering. The focus is on learning by doing, with theoretical material supplemented by concrete coding examples and programming-based homework.

Keeping in mind that most students in this class have limited programming experience, there will be a rapid 2-week review of the python programming language at the start of the course. These first two weeks will be critical for students to get familiar with the language and to be successful in the remainder of the course. Python will be used throughout the course, so these skills will be reinforced throughout the semester through in-class examples and homework.

Course material is divided approximately equally between (i) general topics in data science and machine learning and (ii) specific machine learning methods, their applications, and limitations. The material covered in the first half will thus be revisited and reinforced in the context of specific modeling problems in the second half of the semester. When discussing specific machine-learning models, lectures will typically be broken into a theoretical component followed by a practical implementation component to give students both a general understanding and an opportunity to see concrete applications and coded examples. Given the short time available to cover this large topic, we can only cover a subset of the most popular and/or illustrative machine learning methods in depth. To supplement this limited scope, the course will conclude with a survey of contemporary topics in machine learning that will be provided through student presentations on topics selected in consultation with the instructor.

Grading

Final grades will be based on a group presentation (further described below) and homework. Homework will be assigned approximately weekly in the form of Jupyter Python notebooks and datasets uploaded to Blackboard. Students will need to use a Jupyter client like Anaconda, Google Colab, or Purdue Scholar, to complete these assignments (these options are reviewed on the first day of class). Completed assignments will be uploaded to Blackboard and graded by the instructor. Homework constitutes an important part of the course and should be done conscientiously. NO LATE HOMEWORK WILL BE ACCEPTED.

The final course grade will be weighted as follows:

Group Presentations	25 points
Weekly Homework	75 points
TOTAL	100 points

In the event that the University closes for a period of time during the semester (e.g., due to an outbreak or other unforeseen disaster), we will attempt to continue ChE 597 through assigned reading, problem sets, etc. where I will try to provide lecture material over the web. Communication through email will be critical. If there is a disruption I except that each student will stay connected via your Purdue email account.

Academic Honesty

Group discussions concerning the homework are encouraged, since the sharing of ideas is an excellent way to learn. However, *you need to write your own Jupyter notebooks to submit to Blackboard.*

Academic integrity is one of the highest values that Purdue University holds. Individuals are encouraged to alert university officials to potential breaches of this value by either emailing integrity@purdue.edu or by calling 765-494-8778. While information may be submitted anonymously, the more information that is submitted provides the greatest opportunity for the university to investigate the concern.

The highest standards of Academic Honesty are expected in CHE 597. Any participation in an academically dishonest practice such a copying on work, etc. will result in an F in CHE 597 as well as forwarding your case to the Dean of Students for appropriate disciplinary action.

Mental and Physical Health

Purdue University is committed to advancing the mental health and well-being of its students. If you or someone you know is feeling overwhelmed, depressed, and/or in need of support, services are available. For help, such individuals should contact Counseling and Psychological Services (CAPS) at (765)494-6995 and http://www.purdue.edu/caps/ during and after hours, on weekends and holidays, or by going to the CAPS office of the second floor of the Purdue University Student Health Center (PUSH) during business hours.

Group Presentations

This is a highly accelerated course on data science and machine learning which necessarily leaves a lot of topics out. To provide a supplemental overview of contemporary developments, 20 min group presentations will be prepared and delivered by students in the last three lectures of class. Groups will be assigned mid-semester and the topic for each presentation will be made in consultation with the instructor. The following list of provisional topics for this semester are:

- Reinforcement Learning
- Generative Models
- Active Learning
- Ensemble Methods
- Transfer Learning

The presentations should provide (*i*) a general overview of the topic, including its basic definitions, the classes of problems it is concerned with, and how it relates to material covered in the course and (*ii*) a detailed review of one paper published within the last two years on the chosen topic. These presentations will constitute one quarter of the final grade.

Note: *Depending on how the COVID-19 situation evolves, we may end up converting this assignment into individual term papers rather than group projects.*

Lecture Schedule – Spring 2020

	Date	Topic
Part 1: Python Introduction		
1	Jan. 19	Course overview and introduction to basic python syntax.
2	Jan. 21	Jupyter notebooks, Python syntax, data types, operators, and functions.
3	Jan. 26	Data structures, control statements, and comprehensions.
4	Jan. 28	Python Classes, reading/writing files, and importing.
Part 2: Data Analysis Using Scipy and Pandas		
5	Feb. 2	Introduction to numpy array, matrix and tensor operations.
6	Feb. 4	Introductions to Pandas dataframe.
7	Feb. 9	Visualizing data, Standardizing, and Inputation.
8	Feb. 11	Elements of Data Analysis: Outlier Detection and Dimension Reduction.
Part 3: Optimization Algorithms		
9	Feb. 16	Optimization: Theory of gradient-based methods

10	Feb. 18	Optimization: Implementation of gradient-based methods
11	Feb. 23	Optimization: Theory of Global and Heuristic Methods
12	Feb. 25	Optimization: Implementation of Global and Heuristic Methods

Part 4: Supervised and Unsupervised Learning

13	Mar. 2	Model Training – Data Splits, Validation, Unbalanced Data, Metrics
14	Mar. 4	Model Training – Implementation and Error Analysis
15	Mar. 9	Supervised Learning – Regression, Featurization, and Regularization
16	Mar. 11	Supervised Learning – Theory of common classifiers
17	Mar. 16	Supervised Learning – Implementation of common classifiers
	Mar. 18	No Class – Reading Day
18	Mar. 23	Supervised Learning – Theory of Random Forests and Ensemble Methods
19	Mar. 25	Supervised Learning – Implementation of Random Forests and Ensemble Methods
20	Mar. 30	Supervised Learning – Theory of Neural Networks
21	Apr. 1	Supervised Learning – Implementation of Neural Networks
22	Apr. 6	Supervised Learning – Theory of Time-Series Analysis and Forecasting
23	Apr. 8	Supervised Learning – Implementation of Time-Series Models
	Apr. 13	No Class – Reading Day
24	Apr. 15	Unsupervised Learning – Overview of Algorithms
25	Apr. 20	Unsupervised Learning – Implementation of Algorithms
26	Apr. 22	Student Presentations
27	Apr. 27	Student Presentations
28	Apr. 29	Student Presentations