Data Science in Chemical Engineering

Christopher Paolucci

Spring, 2023

Instructor: Chris Paolucci WDF 306, cp9wx@virginia.edu
TA: Emily Baum byx3au@.virginia.edu
Office Hours: TA:TBD Prof. TBD, WDF 306

Course Description

This course has been adapted from a popular chemical engineering elective taught by Prof. A.J. Medford at Georgia Tech for three years. The course is structured to provide a practical introduction to data analysis and machine-learning for chemical engineers. Topics covered will include data storage and retrieval, dimensional reduction, classification algorithms, regression algorithms, resampling and regularization, and case studies in chemical engineering. The course will emphasize practical programming skills using Python implementations and Jupyter notebooks.

Prerequisites (or equivalent courses from other departments)

1. Intro to Programming CS1110
2. Multivariable Calculus APMA2120
3. Differential Equations APMA2130
4. Applied Probability and Statistics APMA3110
5. Modeling & Simulation in Chemical Engineering CHE2216

Supplementary Materials

- "The Elements of Statistical Learning" Hastie, Tibshirani, Friedman (2008).
  Available: https://web.stanford.edu/~hastie/Papers/ESLII.pdf
- "Pattern Recognition" Duda, Hart, Stork (1973)
  Available: http://cns-classes.bu.edu/cn550/Readings/duda-etal-00.pdf
- "Numerical Python: A Practical Techniques Approach for Industry (2015)" Johansson
  Available: https://github.com/jrjohansson/scientific-python-lectures/

- "Python Data Science Handbook" J. VanderPlas (2017)
  Available: https://jakevdp.github.io/PythonDataScienceHandbook/
- "Numerical Linear Algebra" Lloyd, Trefethen, Bao (1997)

Course Objectives

1. Utilize existing Python libraries to interact with data
2. Store and retrieve data from databases using APIs
3. Visualize high-dimensional data in low-dimensional space
4. Select appropriate algorithms for clustering and classifying data
5. Quantitatively identify regression and classification models with optimum complexity

Course Structure

Class Structure

Lectures will focus first on foundational programming skills, followed by an overview of various techniques and algorithms. Substantial time will be devoted to model selection and validation, and various case studies in the chemical engineering will be explored.

Assessments

Assessments will include 5 homework assignments consisting of hands-on programming problems, with one or more problems that are Graduate only. There will be a midterm exam that focuses on conceptual understanding and assess comprehension of materials covered in lecture. The final will be a class project which will be based on self-defined (graduate students, optional for undergraduates) or data sets provided by me (undergraduate students). I will drop each student's lowest HW grade for the course.

Late Assignments

Late assignments will not be accepted; homework solutions will be posted online shortly after assignment due dates. I reserve the right to extend the deadline for an assignment (for the entire class) if it conflicts with an exam date for another required chemical engineering graduate course or other important events arise. I will not extend due dates on an individual basis. I am dropping the lowest score homework for each student as mentioned above, and if you turn in a late homework I will still grade it even though you will not receive points for it.

Class Project

The project will seek to solve real engineering problems using advanced data analysis methods. Graduate students should define their own project, while undergraduates can choose from predefined projects (but are encouraged to define their own). The project definition must follow the

provided requirements, and must be approved. The project will have defined milestones that culminate in a final grade, and final report.

Grading Policy

A numerical grade will be computed based on the following formula:

- **30%** Project
- **40%** Homework assignments
- **30%** Midterm exam

Laptops and Software

Students are strongly encouraged to bring a laptop to class in order to follow along with in-class programming exercises as appropriate. The course will utilize the Python programming language along with numerous related modules and packages. The software tools are compatible with all operating systems, but the instructors are most familiar with Linux and OSX. Instructors will do their best to solve IT issues such as software installation on all systems, but cannot guarantee support for all computer models and operating systems. Students are ultimately responsible for software installation and maintenance issues.

Honor Code

I trust every student in this course to fully comply with all of the provisions of the University's Honor Code. By enrolling in this course, you have agreed to abide by and uphold the Honor System of the University of Virginia

I encourage you to discuss assignments with your peers after first working through the assignment yourself. Your solutions must be your own. Homeworks which have a significant coding element will be run through software that analyzes those codes for similarity.

Schedule

1. Introduction to Python, NumPy, and Matplotlib
2. Numerical Methods Review
   (a) Linear Algebra
   (b) Linear Regression
   (c) Numerical Optimization and Non-linear Regression
3. Machine Learning & Regression
   (a) Non-parametric Models
   (b) Model Validation
   (c) Complexity Optimization
   (d) High Dimensional Data
4. Classification
   (a) Overview of Classification Methods

   (b) Generalized Linear Models
   (c) Support Vector Machines
   (d) Alternate Classification Methods
5. Data Management
   (a) Data Cleaning and Organization
   (b) Online Data Access and APIs
6. Unsupervised Methods
   (a) High Dimensional Data Part II
   (b) Dimensionality Reduction
   (c) Clustering
   (d) Generative Models
7. Feature Engineering
   (a) Feature Transformations and Auto-Feature Generation
   (b) Time Series Data Analysis

Changes to Syllabus

The schedule and syllabus are subject to change. Given that this is a new course, some changes are to be expected.