ChBE 4745/6745: Data Analytics for Chemical Engineers

A. J. Medford

Spring 2020

Instructor: A.J. Medford
Email: andrew.medford@chbe.gatech.edu
Graduate TA: Gabriel Gusmão
Email: gusmaogabriels@gatech.edu
Office Hours: See Canvas
Class Hours: T/H 4:30 - 5:45pm
Office Hours Location: See Canvas
Class Room: ES&T L1125

---

1 Textbooks

The course will roughly follow the topics outlined in "Machine Learning Refined". Chapters 1-4 are most relevant, and are available for free online. However, having a full copy of the textbook is recommended.

- "Machine Learning Refined" Watt, Borhani, & Katsaggelos (2016)
  Partially Available: https://github.com/jermwatt/machine_learning_refined

1.1 Supplementary Textbooks

The following sources are excellent resources for practical implementations of numerical methods and machine learning algorithms in Python. These are highly recommended for anyone who has not worked with Python before, and are both freely available online.

- "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/

1.2 Recommended Graduate-level Textbooks

The following textbooks contain more mathematical details and are recommended for graduate-level students who are interested in the underlying theory.

- "The Elements of Statistical Learning" Hastie, Tibshirani, Friedman (2008).
  Available: https://web.stanford.edu/~hastie/Papers/ESLII.pdf

- "Pattern Recognition and Machine Learning" Bishop (2006).
  Available: http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf
- "Numerical Linear Algebra" Trefethen & Bau (1997).
  Partially Available: http://people.maths.ox.ac.uk/~trefethen/text.html

2 Overview:

This course will cover the basic principles of machine learning and data analysis. The course will provide chemical engineers with a basic literacy in these topics as well as experience applying the techniques to chemical engineering data sets.

2.1 Description

This course will introduce data science and machine learning techniques for analyzing, interpreting, and visualizing datasets commonly encountered in chemical and biomolecular engineering. Students will learn practical tools needed to understand which data analysis approaches are appropriate for different datasets, how to apply them, and correctly interpret their results. Topics covered will include: regression, classification, data storage and retrieval, dimensionality reduction, clustering, time series analysis, and cross-validation. The course will be taught in the Python programming language, and will include a short introduction for students familiar with MATLAB or other languages. In addition to lectures, the course will feature homeworks from "case studies" that will step students through the application of the data analysis techniques introduced in the course to datasets common in chemical engineering including spectroscopy data, materials property predictions, and chemical process data.

2.2 Prerequisites:

For ChBE 4745 the prerequisite is ChBE 2120.
For ChBE 6745, graduate standing is required, but there are no specific prerequisites. However, students should be familiar with basic concepts in numerical methods, optimization, and programming to be successful in the course.

2.3 Learning Objectives

1. Utilize Python libraries to analyze, visualize, and organize data
2. Store and retrieve data from databases using APIs
3. Evaluate the accuracy of a model with quantitative metrics
4. Quantitatively identify a supervised learning model with optimum complexity
5. Apply unsupervised and supervised techniques to improve model performance through feature engineering
6. Visualize high-dimensional data in low-dimensional space
7. Apply data science techniques to realistic data sets

3 Grading

3.1 Numerical Grade Criteria

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

- **20%** Exercises 4745 (**10%** Exercises 6745)
- **20%** Homeworks (additional questions for 6745)
- **15%** Midterm Exam
- **20%** Final Exam
- **20%** Final Project (**30%** Final Project 6745)
- **5%** Peer Grading

3.2 Letter Grade Criteria

Final letter grades will be assigned based on "clusters" of students identified by large gaps in the grading distribution to minimize sensitivity to individual assignments. Letter grades will be assigned by evaluating samples of work from each "cluster" of students based on the following guidelines:

- Students receiving A grades will have demonstrated, by the end of the semester, mastery in most topics and proficiency in every topic covered in the course.
- Students receiving B grades will have demonstrated competency in all but the most difficult topics in the course.
- Students receiving C grades will have demonstrated familiarity with most topics, but have exhibited unresolved technical challenges with numerous topics covered in this course.
- Students receiving D or F grades will have exhibited deficiencies in numerous topics covered in this course, and may not even be familiar with some topics due to lack of participation or missing assignments.

There are no pre-defined quotas for the number of letter grades assigned, but historically 30-60% of the course has received A's.

3.3 Assignments

3.3.1 *Exercises*

Exercises are designed to reinforce concepts and coding skills from the lectures and will be interspersed with the lecture notes. Exercises will occasionally be discussed or worked in class through active learning exercises or hands-on sessions, so attending and participating in class will reduce the amount of effort needed for the exercise. Exercises will be submitted as a single .pdf via Gradescope, and you will **receive a single 0-5 grade for each Exercise assignment based on completeness rather than correctness**. A 0 grade means the assignment was not turned in, and a 5 means that the solution clearly shows intermediate work and a complete (but possibly

incorrect) answer to each exercise. Detailed feedback on each exercise will not be provided by graders, but exercises can be discussed during office hours or class. Exercises will typically be due within 1 week of the end of a module, but it is expected that you work on the exercise throughout the module.

3.3.2 *Homeworks*

Homeworks are designed to test your ability to extend concepts from lectures to new chemical engineering datasets. Homeworks will involve answering specific questions about datasets. Questions may be open-ended and may not always have a single correct answer. Homeworks for ChBE 6745 may also include additional theory-based questions. Homeworks should be submitted as a stand-alone .pdf document that answers each question along with sufficient evidence, including plots and figures. The .pdf document should be uploaded via Gradescope, and the report along with any supporting code or data should be submitted in a .zip file via Canvas. Homeworks will be graded by peers as well as instructors. **You will receive a 1-5 grade for each question in a Homework assignment**, and the grade will ultimately be determined by the instructor after considering comments from peer graders. Homeworks will typically be due within 1 week of the end of a module, but it is expected that you work on the problems throughout the module.

3.3.3 *Midterm/Final Exam*

There will be two written, closed book exams. The first midterm exam is weighted slightly lower since the final exam will be comprehensive. An equation sheet will be provided with all relevant equations, and calculators are recommended but will not be strictly necessary. Exams will revisit concepts from both Exercises and Homeworks, and also include problems that test students' ability to extend concepts to new problems. Exams may include additional or alternate questions for ChBE 6745.

3.3.4 *Final Project*

ChBE 4745: Students in ChBE 4745 will be allowed to choose between pre-defined projects with associated datasets and project objectives. Projects will be selected by the midterm, and the approach and findings will be documented using a Jupyter notebook along with a 2-3 page final report and a presentation outlining findings.

ChBE 6745: Students must propose a project based on a dataset of their choice, preferably from a real research problem. The objectives and goals of the projects will be defined in a project proposal (1-2 pages) due by the midterm. The approach and findings will be documented using a Jupyter notebook, along with a 3-4 page final report and presentation outlining findings. The final report must include a section describing proposed future work.

ChBE 4745/6745: Groups should be composed of 4-5 students, and it is recommended that students only select groups within 4745/6745. Mixed groups will be graded based on 6745 standards. Project grades will be based on a series of intermediate assignments as follows:

- **10%** Group/project selection (and project proposal for 6745)

- **15%** Baseline supervised learning model including validation strategy*
- **15%** Variations to baseline model based on feature engineering approaches*
- **10%** Draft report and peer grading
- **25%** Final report
- **15%** Final presentation
- **10%** Individual contributions**

These assignments will primarily be graded on a per-group basis, but some assignments (marked with *) will be graded as individual assignments. However, group members are allowed to submit identical work. Peer grading will be used along with instructor grading for these assignments. Individual contributions (marked with **) will be assigned by instructors based on (1) deviations between individual contributions and group report for baseline model and variations, (2) individual contributions as described in the final report, and (3) peer evaluations. This means that a **total of 30% of the grade is controlled by individual contributions**.

3.4 *Peer Grading*

The course will utilize "peer grading" administered through Canvas for various assignments. This gives you an opportunity to see how others approach open-ended problems, and practice applying critical thinking skills.

When you grade your peers, you will not provide a numerical grade, but will instead leave a comment that clearly identifies the **greatest strength(s)** and **greatest weakness(es)**. You must identify at least one strength/weakness of each assignment you grade, but may identify more than one. All comments should be **constructive criticism**. These comments will have no formal impact on the grade assigned by TA's, but may assist the TA's in identifying strengths or weaknesses of the assignment. The "Peer Grading" portion of your grade will be based on your response rate. As long as you provide at least one strength/weakness in a constructive manner you will receive full credit for peer grading. All peer grading is due within 7 days of the assignment. Failure to submit an evaluation in time, or failure to follow instructions will result in zero credit.

3.5 *Dropped Grades and Bonus Points*

If the CIOS response rate exceeds 90% the lowest exercise score will be dropped, and if it exceeds 95% the lowest exercise and homework scores will be dropped.

Bonus exercises and programming challenges may be offered. Bonus points will be applied to the "Exercises" category with a maximum of 100%.

4 Course Structure

4.1 Class Format

The course will be organized into modules. Each module will have exercises and an associated homework. The exercises will be graded based on completeness and homeworks will be graded

based on correctness. Some class time will be provided for working on exercises, but it is expected that you study the lectures outside of class to ensure you are prepared. You are allowed to seek help from your group members and classmates, but ultimately **all homework and case studies should be completed and submitted independently**.

4.2 Online Content

Some modules may be provided as online modules. In this case, students are expected to watch all content outside of class and course time will be used as "flipped" classes to work on exercises and homeworks.

4.3 Office Hours and Recitation

The instructor and TA will each hold a weekly office hour which will serve as an opportunity to ask questions, but there will be no prepared material. In addition, there will be a biweekly "recitation" session where the TA will work exercises or relevant problems. Discussions of highly specific or personal matters should be handled through individual meetings scheduled with the instructor.

4.4 Modules

The modules along with tentative sub-topics are given below:

- Numerical Methods
  - Python basics
  - Linear algebra
  - Linear regression
  - Multivariate optimization
  - Dataset: IR Spectrum Analysis
- Regression
  - Cross validation
  - Ridge regression
  - Kernel ridge regression
  - Neural Networks
  - Dataset: Impurity concentrations
- Classification
  - Class imbalance
  - Logistic regression
  - Support Vector Machines
  - Naive Bayes

  - k-Nearest Neighbors
  - Dataset: Fault detection
- Data Management
  - Spreadsheets
  - Application Programming Interfaces
- Exploratory Data Analysis
  - Principal component analysis
  - Hierarchical analysis
  - K-means/Gaussian mixture clustering
  - Mean shift clustering
- Feature Engineering
  - Feature selection
  - Partial Least Squares
  - Linear Discriminant Analysis
- Time Series Data
  - Forecasting
  - ARMAX modeling

4.5 Schedule

Week 01, 01/06 - 01/10: *Numerical Methods*

Week 02, 01/13 - 01/17: *Numerical Methods*

Week 03, 01/20 - 01/24: *Regression*

Week 04, 01/27 - 01/31: *Regression*

Week 05, 02/03 - 02/07: *Classification*

Week 06, 02/10 - 02/14: *Classification*

Week 07, 02/17 - 02/21: *Data Management*

Week 08, 02/24 - 02/28: *Data Management*

Week 09, 03/02 - 03/06: Review, **Exam 1**

Week 10, 03/09 - 03/13: *Exploratory Data Analysis*

Week 11, 03/16 - 03/20: *Exploratory Data Analysis*

Week 12, 03/23 - 03/27: *Feature Engineering*

Week 13, 03/30 - 04/03: *Feature Engineering*

Week 14, 04/06 - 04/10: *Time Series Data*

Week 15, 04/13 - 04/17: *Time Series Data*

Week 16, 04/20 - 04/24: Final Presentations

Final Exam, 4/28 2:40 - 5:30pm

5 Statements and Disclaimers

5.1 Attendance

As per Georgia Tech policy, you are permitted to be absent from class to participate in athletic events, official field trips, and religious observances. Since attendance is not formally taken, you will be responsible for contacting your classmates or instructors/TA's to catch up on anything you missed. If your absence may conflict with an exam, please provide written notice of your upcoming absence at least two weeks before the event, and ideally within the first two weeks of class. Please see http://catalog.gatech.edu/rules/4/ for more information about receiving official notice from the Registrar about the nature and timing of your upcoming Institute-approved absence.

5.2 Academic Integrity

The Georgia Tech Honor Code prohibits unlawful collaboration (see www.honor.gatech.edu). This requires each course to define what collaboration is and is not permitted. In this course students may collaborate on exercises, homeworks, and studying for exams. Students are permitted and encouraged to discuss general aspects of the homeworks, the algorithms involved and the implementation of the algorithm, as well as using online resources such as Stack Overflow. However, students must **write and apply their code independently and turn in their own work** for all assignments except for the final project. Any similarities between homeworks that indicate copying all or part of a question, including effectively copying (i.e. copying code and changing comments and variable names or similar acts) will be considered an honor code violation. A good rule of thumb is that you may look at someone else's code if you are helping them, but you should not look at others' code if you need help. We may use the "Measure of Software Similarity" (MOSS) algorithm to identify potential copying of code. Potential violations will be flagged for a warning or forwarded to the Dean of Students Office as appropriate.

5.3 Diversity and Disability Statement

Georgia Tech values diversity and inclusion; we are committed to a climate of mutual respect and full participation. Our goal is to create learning environments that are usable, equitable, inclusive and welcoming. If there are aspects of the instruction or design of this course that result in barriers to your inclusion or accurate assessment or achievement, please notify the instructor within the first week of class. Students with disabilities should contact the Office of Disability Services to discuss options of removing barriers in this course, including accommodations. ODS can be reached at 404.894.2563, dsinfo@gatech.edu, or disabilityservices.gatech.edu. Students requiring testing accommodations must schedule exams within 1 hour of the regularly-scheduled exam. This may require scheduling well in advance of the exam.

5.4 Health and Well-being

Your health and wellness, including mental health, are more important than your grades. Students are encouraged to pursue wellness through making proactive, healthy choices. The Office of Health Initiatives offers a free online screening so that you can determine quickly and easily whether a professional consultation might be helpful to you. You are encouraged to be proactive regarding mental health to avoid a crisis situation. However, if you or a friend are experiencing a crisis that requires immediate attention you may speak with a counselor at any time 24 hours a day, 7 days a week. During regular business hours (Monday-Friday 8-5) Students are initially assessed through GT CARE, a separate department working closely with the Counseling Center to meet student needs. All students currently registered in a degree-seeking program are eligible for services at the Counseling Center.

5.5 Changes to Syllabus

The schedule and syllabus are subject to change. Given that this is a new course, changes are to be expected; however, we will do our best to notify you of any changes and implement them as fairly as possible. In the event of conflicting information between instructors, the information on the most recent posted syllabus will take priority. If a policy is ambiguous or absent from the syllabus then instructors will collaborate to resolve the ambiguity as fairly as possible.