Chemical Engineering 4H03

Big Data Methods in Chemical Engineering

Course Outline - Winter 2020

Course Details

Instructor:		**Dr. Jake Nease**	**che4h3instructor@gmail.com** **+1 (905) 599-3165**	**BSB/B105**
Teaching Assistants:		Hassan Abdulhussain Satyam Agnihotri	abdulh22@mcmaster.ca agnihs1@mcmaster.ca	JHEA407 JHE/256
Website:		Avenue2Learn	avenue.mcmaster.ca	
Lectures:		Tu/Th/Fr	08:30 – 09:20	JHE/A102
Office Hours:	**JAKE**: **TAs**: 	Mo/We We Mo We	08:30 – 09:20 14:30 – 17:00 9:30 – 10:20 15:30 – 16:20	 JHE/345 JHE/345
Prerequisites:	Registration in level IV or above in Chemical or Materials Engineering			
Software:	MATLAB – For modeling, regression, and general algorithm design EXCEL – For processing and visualizing large data sets OTHER – Various big-data software packages such as ProMV, *etc.* *Additional software packages for applied modeling will be used as necessary*			
Course Materials:	Lecture modules, assignments, readings, and solutions will be posted on A2L Grades will also be posted on A2L but are not official			
Recommended Textbooks:	There are no required texts for this class Resources and papers will be posted to A2L throughout the term			

Formal Course Description

Chemical Engineering 4H03 focuses on the application of methods to **parse**, **filter**, **interpret**, **learn** and **extract value** from large industrial and consumer data sets. Topics include dimensionality reduction (PCA, PLS), soft sensors for process monitoring, data clustering, artificial intelligence (neural networks and binary decision trees), and modeling/integrating these and other more fundamental methods to learn from and exploit data.

Informal Course Description

In today's industry, you may encounter A LOT of data. This course will show you ways to visualize, learn from, and eventually exploit that data for process improvement.

Learning Objectives

After completing this course, the student should be able to:

- Demonstrate the concept of identifying the **best model** to explain a data set
- Fit and compute model parameters for principal component analysis (PCA) and partial-least squares (PLS)
- Compare the trade-offs between model **accuracy** and **computational effort**
- Demonstrate the ability to identify if a model is over- or under-fit using statistical significance metrics
- Visualize large data sets to identify trends and key observations
- Provide context to a data set so that a lay-person could understand the key conclusions
- Demonstrate a core understanding of the fundamental background theory behind various big-data driven models and artificial intelligence methods
- Identify **misleading results,** apply appropriate analyses to **judge their accuracy** or applicability, and suggest more appropriate alternatives

Grading Policies

Please be aware of the following grading policies for ChE 4H03:

- Late submissions of any take-home portions of exams will not be accepted without an appropriate MSAF
- Valid MSAF submissions will result in either a make-up examination or rolling of that component's weight into the final exam, depending on the situation
- Any tests and exams will be open-book (any book) and open-notes (any hard copies)
- Any calculator may be used for examinations
- All grades are unofficial until final grades are posted on McMaster's student and faculty software: MOSAIC
- The instructor retains the right to modify course weights or components, typically only enforced for the student's benefit
- Final grades will be converted to the standard McMaster 12-point scale
- All submissions for assignments, projects, and take-home examinations must be done **electronically**
- Any copying of code, formulations, or interpretations from other students, prior versions of this course, or resources online will be considered a violation of McMaster's academic integrity policy

Grading Breakdown

Please note that the grading breakdown for this course will be decided **on the first day of class**.

Weight	Component	Comments
25%	Assignments	Up to five assignments worth up to 5% each
25%	Midterm Test	Up to two midterms (on separate units) worth 12.5% each
20%	Course Project	MUST be included. Groups of three
25%	Final Exam	MUST be included. 2.5 hour written examination
5%	Paper Review	ONE paper review presentation in a random group of 3 (10 mins)

Assignments

There may be **5** assignments for this course depending on student interest, each counting for 5% of the student's final grade, up to a maximum of 25%. If more than 5 assignments are released, only the best 5 results will count at 5% each. Please note the following considerations regarding the assignments:

- Assignments may be completed in groups of **up to two (2) students**
- Assignments must be submitted **electronically** to the appropriate A2L repository prior to the due date
- All relevant code, with appropriate comments and guidance for graders, must also be submitted with each assignment
- Assignments will typically focus on main topics of this course (broken down below), but the instructor reserves the right to modify the coverage of assignments
- All assignments will focus on the **application** of course concepts to real data sets, whereas tests and exams will focus more on the theory and simplified analysis
- All assignments will have a **demerit system** for presentation and professionalism. Any submissions that are poorly formatted, have no discussion, or show any other lack of professionalism will be penalized

Course Project

A significant portion of the student's grade will come in the form of a course project. The project will follow a guided self-directed learning (SDL) format in which the students will develop their own topic by applying course concepts to a problem that they consider to be interesting or noteworthy. Some additional comments about the course project:

- The project **MUST** be tackled in groups of **three (3)**. *Special circumstances* will permit groups of more than three, at the discretion of the instructor
- A **proposal** (10% of project grade) will be due after approximately one month. A specific due date will be communicated by the instructor closer to the date
- After the proposal has been reviewed, each group will meet briefly with the instructor (15-minute meetings) to hammer out **specific project details and objectives**
- The last week of lectures will be reserved for **project presentations** (50% of project grade). Presentation and questioning lengths will depend on the number of students in the class
- A **final report** (40% of project grade) detailing the problem to be solved, solution methodology and results/discussion (not to exceed 10 pages) will be graded by the instructor

Midterm and Exams

The course midterm(s) (???%) and exam (???%) will primarily be written tests performed individually. As mentioned in the grade breakdown, take-home components **may** be assigned **depending on the interest of the class** (sometimes take-home portions for this kind of material are requested). Some other short comments:

- The final exam will be **cumulative** and may test all the components of the course
- The midterm(s) will take place at to-be-determined times and locations throughout the term.
- After any midterm(s), the proceeding lecture will be for reserved for a **collaborative re-write** of the same midterm in randomized groups of five (5). Your midterm score will be comprised of 85% of your individual score and 15% of your collaborative score if it is better than your individual score
- The final exam will be scheduled by the McMaster Registrar's office and will take place in early-mid April

Academic Integrity

You are expected to exhibit honesty and use ethical behaviour in all aspects of the learning process. Academic credentials you earn are rooted in principles of honesty and academic integrity.

Academic dishonesty is to knowingly act or fail to act in a way that results or could result in unearned academic credit or advantage. This behaviour can result in serious consequences, e.g. the grade of zero on an assignment, loss of credit with a notation on the transcript (notation reads: "Grade of F assigned for academic dishonesty"), and/or suspension or expulsion from the university. It is your responsibility to understand what constitutes academic dishonesty. For information on the various types of academic dishonesty please refer to the Academic Integrity Policy, located at http://www.mcmaster.ca/academicintegrity.

The following illustrates only **three forms** of academic dishonesty:

- Plagiarism, e.g. the submission of work that is not one's own or for which other credit has been obtained
- Improper collaboration in group work: this point is **particularly important** and will be strongly penalized in this course
- Copying or using unauthorized aids in tests and examinations (solutions, for example)

Accessibility

The instructor aims to make this class accessible to all students. Please forward and optionally discuss any accommodation granted by Student Accessibility Services (SAS) with the instructor *before the third week of the course*. Please raise any other accessibility issues with the instructor as soon as possible, e.g. accessibility of the course website and course materials.

Course Feedback

Please do not hesitate to let me know your thoughts on the course or what you might want to change at any time. You can reach me at neasej@mcmaster.ca or che4h3instructor@gmail.com. If you would prefer to leave feedback **anonymously**, do not hesitate to use our anonymous 4H3 course feedback form.

Dates to Remember

- First lecture: Tuesday January 07
- Reading week: Week of February 17
- Last day for cancelling 4H03: March 13
- Midterm test(s): **TBD**
- Australian Open (Will Roger Federer win one more major in 2020?): January 20 – February 07
- Last Lecture: April 7

Class Recordings

As has become standard in my courses, classes will be recorded each week and posted on YouTube. Please remember that these recordings should NOT be used as an excuse to skip class, and since they are one-take I cannot make any guarantees of their availability in the event of technical difficulties.

Anticipated Course Schedule and Topics – SUBJECT TO CHANGE

Note that all topics will have engineering and other applications weaved in throughout.

Week Number	Anticipated Topics	Anticipated Module Content (subject to change)
Week 01	Course Overview Review of Modeling Core Concepts	• Course overview • Basic stats review • Sampling methods • Language and lingo • Visualizing data
Week 02	Multivariate Regression Basis Function Regression	• Review of SSE optimization • Goodness of fit • Training/testing sets • Weakness: dependent sets of data
Week 03	Dimension Reduction (Latent Variable Methods)	• Principal Component Analysis (PCA) • Purposes of PCA • Defining loadings, scores, and visualization
Week 04	Dimension Reduction (Latent Variable Methods)	• PCA derivation and the NIPALS algorithm • Interpreting scores • Application: Sports data mining • Application: Soft sensors and equipment health
Week 05	Dimension Reduction (Latent Variable Methods)	• Projection of Latent Structures (PLS) • Modeling with PLS • Application: Spectral data
Week 06	Data Clustering	• K-Means Clustering (algorithm and interpretation) • Integrated clustering and visualization in the reduced-dimension space
RW	**Reading Week**	• Suggested Reading: "Infinite Jest" by David Foster Wallace. It might take you the entire week :|
Week 07	Data Clustering	• Expectation Maximization Clustering • Support vector machines
Week 08	Machine Learning Tools	• Background and core concepts • Supervised learning (AKA regression!)
Week 09	Machine Learning Tools	• Artificial Neural Networks • Single-layer networks • Basis functions
Week 10	Machine Learning Tools	• Multi-layered networks • Examples and analysis
Week 11	Machine Learning Tools	• Support Vector Machines for categorization • Predicting sample classifications
Week 12	Boolean Decisions	• Decision Trees • Random Forests
Week 13	Presentations	• N/A