Introduction to Data Science I

From Introduction to Data Science

Contents

- 1 Course outline for COMPSCI 4414A/9637A/9114A
 - 1.1 Prerequisites
 - 1.2 Instructor Information
 - 1.3 Course Description and Objectives
 - 1.4 Important Dates
 - 1.5 Course Materials
 - 1.6 Topics (anticipated)
 - 1.7 Evaluation
 - 1.7.1 Midterm 4414/9114: 35% 9637: 30%
 - 1.7.2 Brainstorming Session 10%
 - 1.7.3 Project Proposal 4414/9114: 15% 9637: 10%
 - 1.7.4 Report Draft 5%
 - 1.7.5 Project Report 35%
 - 1.7.6 Peer Review **9637 only:** 10%
 - 1.7.7 Participation and Effort
 - 1.8 Accommodation and Accessibility
 - 1.9 Academic Policies
 - 1.10 Support Services
 - 1.11 Timeline (Tentative)

Course outline for COMPSCI 4414A/9637A/9114A

The University of Western Ontario London, Ontario, Canada Department of Computer Science Course Outline - Fall (September - December) 2018

Note that this course is in high demand. Now that those who submitted a proposal successfully have been registered, the course is open to all computer science students (who can register themselves online) subject to space availability. If there is space remaining after 21 September, students from other Departments and Faculties may be admitted. Those interested should attend lectures anyway.

Note that Master of Data Analytics students are exempt from this and will be registered in 9114A.

Prerequisites

0.5 course from Biology 2244A/B, Statistical Sciences 2035, Statistical Sciences 2141A/B, Statistical Sciences 2143A/B, Statistical Sciences 2244A/B or Statistical Sciences 2858A/B; 1.0 course from Computer Science 1025A/B, Computer Science 1026A/B, Computer Science 1027A/B, Computer Science 1037A/B, Computer Science 2120A/B, Computer Science 2121A/B, Digital Humanities 2220A/B, Digital Humanities 2221A/B, Engineering Science 1036A/B; and 0.5 course from Mathematics 1229A/B, Mathematics 1600A/B, Applied Mathematics 1411A/B; and written permission of the Department obtained by applying as above.

Instructor Information

- Instructor: Dan Lizotte dlizotte at uwo dot ca Office MC363
- Teaching Assistant: Nathan Phelps nphelps3 at uwo dot ca
- Time: Tuesday from 2:30PM 4:30PM, and on Thursday from 2:30PM 3:30PM
- Place: Talbot College TC-205 (http://www.music.uwo.ca/pdf/resources/TC-03.pdf)
- Communication: We will be using OWL (https://owl.uwo.ca) for electronic communication.

Course Description and Objectives

The objective of this course is to introduce students to data science (DS) techniques, with a focus on application to substantive (i.e. "applied") problems. Students will gain experience in identifying which problems can be tackled by DS methods, and learn to identify which specific DS methods are applicable to a problem at hand. During the course, students will gain an in-depth understanding of a particular (substantive problem, DS solution) pair, and present their findings to their peers in the class. Although this course does not assume prior machine learning or visualization knowledge, it does require students to show substantial initiative in investigating methods that are applicable for their project. The lectures give an overview of important methods, but the lecture content alone is not sufficient to produce a high quality course project.

This course is designed for students who:

- Like to read have a desire to understand substantive problems
- Like to think make connections between methods and problems
- Like to wrangle be willing to wrangle (https://en.wikipedia.org/wiki/Data_wrangling) data into usability
- Like to **speak** teach us about what you found

Important Dates

- Pick Brainstorming Slot by Friday, 5 Oct at 5pm
- Project Proposal Due Friday, 26 Oct at 5pm
- Project Draft Due Friday, 16 Nov at 5pm
- Project Report Due Friday, 7 Dec at 5pm
- Paper Reviews Due Friday, 14 Dec at 5pm

Register for a wiki account. You will need to use the wiki to let us all know about data sources you find, indicate which dataset you are using, and slot yourself in for brainstorming. Also, everyone should free to make improvements to any part of the wiki. (E.g. if you find some useful software or other resources.)

Slot yourself in for a brainstorming session in the Timeline portion at the bottom of this page before end of **Friday**, **6 Oct at 5pm** or Dan will pick a slot for you.

Course Materials

- Required Texts
 - JWHT: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York: Springer. [Free through Western (https://www.lib.uwo.ca/c gi-bin/ezpauthn.cgi?url=http://link.springer.com/978-1-4614-7138-7)]
 - HTF: *The Elements of Statistical Learning* by Hastie, Tibshirani and Friedman. Expanded version of required text. [Free online (http://web.stanford.edu/~hastie/ElemStatLearn/)]
 - LW: Leland Wilkinson's *The Grammar of Graphics* (2005). [Free from Springer (https://www.lib.uw o.ca/cgi-bin/ezpauthn.cgi?url=http://link.springer.com/book/10.1007/0-387-28695-0)]

Introduction to Data Science

- ggplot2 book by creator Hadley Wickham (2016). [Free through Western (https://alpha.lib.uwo.ca/rec ord=b6962637~S20)]
- **Review** if you need to catch up:
 - Calculus Review (https://onlinecourses.science.psu.edu/statprogram/calculus_review) from Penn State University. Includes basic mathematical notation.
 - linear algebra review (http://www.cs.mcgill.ca/~dprecup/courses/ML/Materials/linalg-review.pdf) up to and including Section 3.7 - The Inverse
 - Larry Wasserman's (http://www.stat.cmu.edu/~larry/all-of-statistics/) *All of Statistics*. (Available through UWO Library)
 - Devore, J. L., & Berk, K. N. (2007). Modern mathematical statistics with applications. 2nd ed. Springer. [Free through Western (https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi?url=http://link.springer. com/978-1-4614-0391-3)]

Other Resources

- The Data and Software Page
- Cheat Sheets
 - ggplot2 (https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf) cheat sheet
 - Data Wrangling (https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatshe et.pdf) cheat sheet
- Texts
 - Phil Spector. (2008). Data Manipulation with R New York: Springer. [Free through Western (ht tps://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi?url=http://www.springer.com/us/book/97803877473 09)]
 - probability review (http://www.cs.mcgill.ca/~dprecup/courses/ML/Materials/prob-review.pdf) from Stanford University by way of Doina Precup.
 - List of resources (http://www.cs.mcgill.ca/~dprecup/courses/ML/resources.html) from COMP-652 at McGill (courtesy Doina Precup)
 - C. M. Bishop, Pattern Recognition and Machine Learning (2006)
 - R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction (1998)
 - Ethem Alpaydin, "Introduction to Machine Learning", MIT Press, 2004.
 - David J. C. MacKay, "Information Theory, Inference and Learning Algorithms", Cambridge University Press, 2003.
 - Richard O. Duda, Peter E. Hart & David G. Stork, "Pattern Classification. Second Edition", Wiley & Sons, 2001.
- Other Links
 - Data Visualization for Human Perception (https://www.interaction-design.org/literature/book/the -encyclopedia-of-human-computer-interaction-2nd-ed/data-visualization-for-human-perception)
 - Data Journalism (http://datadrivenjournalism.net/news_and_analysis/is_data_journalism_for_ev eryone)
- Software
 - The dplyr package documentation (https://cran.r-project.org/web/packages/dplyr/). The "vignettes" are particularly good.
 - The Tensorflow Library (Python, C++) [1] (https://www.tensorflow.org/)
- Deep Learning Resources (courtesy Ethan Jackson)
 - Tutorials on Word2Vec in Python. Learns semantic relationships between words in very large corpora by mapping each word to a high-dimensional word embedding. Semantic relationships are estimated using contextual frequency, i.e. how often a word appears given a context of other words.
 - https://radimrehurek.com/gensim/models/word2vec.html
 - https://rare-technologies.com/word2vec-tutorial/

- Some ideas about using t-SNE for visualization
 - https://www.jeffreythompson.org/blog/2017/02/13/using-word2vec-and-tsne/
- Digit classification on MNIST dataset using TensorFlow
 - https://www.tensorflow.org/get_started/mnist/beginners
- Autoencoders for MNIST in Keras (a very high level interface for deep learning libraries including TensorFlow)
 - https://blog.keras.io/building-autoencoders-in-keras.html
- Convolutional neural networks for image recognition on CIFAR-10 dataset in TensorFlow. Great starting point for image classification using deep learning.
 - https://www.tensorflow.org/tutorials/deep_cnn

Topics (anticipated)

- Introduction to Data Science

- Definitions
- Components
- Relationships to Other Fields

Data Munging

- Working with structured data: selecting, filtering, joining, aggregating
- Web scraping
- Simple visualizations
- Sanity checking

• (Re)-introduction to Statistics

- Data Summaries
- Randomness, Sample Spaces and Events, Probability
- Random Variables, CDF, PMF, PDF
- Expectation
- Estimation
- Sampling Distributions: Law of Large Numbers, Central Limit Theorem, The Bootstrap
- Inference: Hypothesis testing, P-values, Confidence Intervals
- Multivariate Statistics: conditional probability, correlation, independence

- Supervised Machine Learning, Predictive Models

- Supervised Learning
 - Regression
 - Classification
- Reinforcement Learning and Sequential Decision Making
- Evaluation
 - Variance: Test set, cross-validation, bootstrap
 - Bias: Confounding, causal inference

Unsupervised Machine Learning, Representations, and Feature Construction

- Clustering
- Dimensionality reduction
- Domain-specific Feature Development
 - Images
 - Sounds
 - Text
- Visualization
 - Topics to be determined

9/10/2018

Evaluation

There will be a midterm test but no final exam. Each student will co-lead a brainstorming session, and co-produce a proposal, draft, and report for a course project. **Graduate students (9637)** will additionally submit peer reviews of other class projects. For detailed requirements, see Project Guidelines.

Scholastic offences are taken seriously and students are directed to read the appropriate policy, specifically, the definition of what constitutes a Scholastic Offence, at this website: [2] (http://www.uwo.ca/univsec/pdf/academic_policies/appeals/scholastic discipline undergrad.pdf).

Midterm - 4414/9114: 35% 9637: 30%

Assessing competencies from the fundamentals taught in the first half of the class.

Brainstorming Session – 10%

Each student will prepare a presentation explaining an applied problem, as well as some potential data science methods that could be applied to the problem. The presentation should be **no more than 10 minutes**. We will then discuss the problem as a class, along with possible approaches for solving the problem using data science methods. **The student is expected to be prepared to answer deep questions about the nature of their problem to ensure that they receive high quality feedback** from the brainstorming session.

Project Proposal - 4414/9114: 15% 9637: 10%

Document detailing the plan for the project. See Project Guidelines for detailed requirements.

Report Draft – 5%

A draft of the final report will be due approximately midway through the term. The purpose of the draft is to allow the instructor to provide feedback on the quality of the writing and the direction of the project.

Project Report – 35%

Each student will prepare a research paper detailing a substantive problem, the data available, the applicable data science methods, and empirical results obtained on the problem.

Peer Review – 9637 only: 10%

Each graduate student enrolled in CS9637 will prepare two reviews of their classmates' work.

Participation and Effort

Success of the course as a useful learning experience hinges on active participation and effort of the students. **Students are expected to attend all classes** and are expected to **actively participate in the brainstorming sessions**.

Accommodation and Accessibility

If you are unable to meet a course requirement due to illness or other serious circumstances, you must provide valid medical or supporting documentation to the Academic Counselling Office of your home faculty as soon as possible. If you are a Science student, the Academic Counselling Office of the Faculty of Science is located in NCB 280, and can be contacted at scibmsac@uwo.ca.

For further information, please consult the university's medical illness policy at http://www.uwo.ca/univsec/pdf/academic_policies/appeals/accommodation_medical.pdf.

Academic Policies

The website for Registrarial Services is http://www.registrar.uwo.ca.

In accordance with policy, http://www.uwo.ca/its/identity/activatenonstudent.html, the centrally administered email account provided to students will be considered the individual's official university e-mail address. It is the responsibility of the account holder to ensure that e-mail received from the University at his/her official university address is attended to in a timely manner.

Electronic devices are not permitted for the midterm.

Scholastic offences are taken seriously and students are directed to read the appropriate policy, specifically, the definition of what constitutes a Scholastic Offence, at the following Web site: http://www.uwo.ca/univsec/pdf/academic_policies/appeals/scholastic_discipline_undergrad.pdf.

All required papers may be subject to submission for textual similarity review to the commercial plagiarism detection software under license to the University for the detection of plagiarism. All papers submitted for such checking will be included as source documents in the reference database for the purpose of detecting plagiarism of papers subsequently submitted to the system. Use of the service is subject to the licensing agreement, currently between The University of Western Ontario and Turnitin.com (http://www.turnitin.com).

Computer-marked multiple-choice tests and exams may be subject to submission for similarity review by software that will check for unusual coincidences in answer patterns that may indicate cheating.

Support Services

Please contact the course instructor if you require lecture or printed material in an alternate format or if any other arrangements can make this course more accessible to you. You may also wish to contact Student Accessibility Services (SAS) at 661-2147 if you have any questions regarding accommodations.

The policy on Accommodation for Students with Disabilities can be found here: http://www.uwo.ca/univsec/pdf/academic_policies/appeals/accommodation_disabilities.pdf

The policy on Accommodation for Religious Holidays can be found here: http://www.uwo.ca/univsec/pdf/academic_policies/appeals/accommodation_religious.pdf

Learning-skills counsellors at the Student Development Centre (http://www.sdc.uwo.ca) are ready to help you improve your learning skills. They offer presentations on strategies for improving time management, multiple-choice exam preparation/writing, textbook reading, and more. Individual support is offered throughout the Fall/Winter terms in the drop-in Learning Help Centre, and year-round through individual counselling.

Students who are in emotional/mental distress should refer to Mental Health@Western (http://www.health.uwo.ca/mental_health) for a complete list of options about how to obtain help.

Additional student-run support services are offered by the USC, http://westernusc.ca/services. https://www.csd.uwo.ca/~dlizotte/teaching/IDS/index.php/Introduction to Data Science I

Timeline (Tentative)

- 6 Sep Lectures:
 - 11 Sep Lectures:
 - 13 Sep Lectures:
 - 18 Sep Lectures:
- 20 Sep Lectures:
 - 25 Sep Lectures:
- 27 Sep Lectures:
 - 2 Oct Lectures:
- 4 Oct Pick Brainstorming Slot by 5 Oct 5pm Lectures:
 - 9 Oct Fall Reading Week
- 11 Oct Fall Reading Week
 - 16 Oct Lectures:
- 18 Oct Project Proposal Due 19 Oct at 5pm Lectures:
 - 23 Oct Lectures:
- 25 Oct Lectures:
 - 30 Oct Lectures:
- 1 Nov Lectures:

- 6 Nov Lectures:
- 8 Nov Midterm Review and Q&A
 - 13 Nov Midterm
- 15 Nov GUEST LECTURE
 - 20 Nov Brainstorming: 1,2,3,4,5,6
- 22 Nov Project Draft Due 24 Nov at 5pm Brainstorming: 1,2,3
 27 Nov Brainstorming: 1,2,3,4,5,6
 - 29 Nov Brainstorming: 1,2,3
 - 4 Dec Brainstorming: 1,2,3,4,5,6
- 6 Dec Brainstorming: 1,2,3
- Project Document Due Friday 7 December 5pm
- Reviews (graduate students only) Due Friday 14 December 5pm

Retrieved from "https://www.csd.uwo.ca/~dlizotte/teaching/IDS/index.php? title=Introduction_to_Data_Science_I&oldid=154"

• This page was last edited on 9 September 2018, at 17:32.