Introduction to Data Science I

From Introduction to Data Science

Contents

- 1 Course outline for COMPSCI 4414A/9637A/9114A
  - 1.1 Objective
  - 1.2 Prerequisites
  - 1.3 Logistics
  - 1.4 Important Dates
  - 1.5 Materials
  - 1.6 Topics (anticipated)
  - 1.7 Evaluation
    - 1.7.1 Daily Quizzes – 5%
    - 1.7.2 Midterm - 35%
    - 1.7.3 Brainstorming Session – 5%
    - 1.7.4 Project Proposal – **4414:** 15% **9637:** 10%
    - 1.7.5 Report Draft – 5%
    - 1.7.6 Project Report – 35%
    - 1.7.7 Peer Review – **9637 only:** 5%
    - 1.7.8 Participation and Effort
  - 1.8 Accessibility and Support Available at Western
  - 1.9 Missed Course Components
- 2 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) 2017

From Dan: This is a very high-demand course that interests students in various programs across campus. I think this is great because the diversity of backgrounds assembled in the class makes for a better learning experience for all. (Myself included!) However, space is limited. Because of the volume of requests I receive, I am not able to manage a wait list. Students will have to monitor the registration website for available spots. However, all are welcome to sit in the room if there is space.

Objective

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 **hack** - be willing to munge (http://en.wikipedia.org/wiki/Data_munging) data into usability
- Like to **speak** - teach us about what you found

Prerequisites

At least one undergraduate programming course (e.g. CS2035) and at least one statistics course (e.g. STAT1024.) This course entails a significant amount of self-directed learning and is directed toward fourth-year undergraduate and graduate students.

Logistics

- **Instructor**: Dan Lizotte – dlizotte at uwo dot ca – Office MC363
- **Teaching Assistant**: Brent Davis - bdavis56 at uwo dot ca - Runs Q/C Hour (see below)
- **Time**: Tuesday from 2:30PM – 4:30PM, and on Thursday from 2:30PM – 3:30PM
- **Place**: Middlesex College **MC-105B** (http://accessibility.uwo.ca/doc/floorplan/bf-mc.pdf)
- **Question and Collaboration Hour:** Tuesday from 4:30pm - 5:30pm **Location MC 320**
- **Communication**: We will be using OWL (https://owl.uwo.ca) for electronic communication.

Important Dates

- Pick Brainstorming Slot by Friday, 6 Oct at 5pm
- Project Proposal Due Friday, 27 Oct at 5pm
- Project Draft Due Friday, 17 Nov at 5pm
- Project Report Due Friday, 8 Dec at 5pm
- Paper Reviews Due Friday, 15 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.

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/cgi-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://www-stat.stanford.edu/~tibs/ElemStatLearn/)]
    - **LW**: Leland Wilkinson's *The Grammar of Graphics* (2005). [**Free** from Springer (https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi?url=http://link.springer.com/book/10.1007/0-387-28695-0)]
    - ggplot2 book by creator Hadley Wickham (2009). [**Free** through Western (https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi?url=http://www.springer.com/us/book/9780387981406)]

- **Review** if you need to catch up:

    - Larry Wasserman's (http://www.stat.cmu.edu/~larry/all-of-statistics/) *All of Statistics*. [**Free** from Springer (http://link.springer.com/book/10.1007/978-0-387-21736-9)]
    - 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)]
    - linear algebra review (http://www.cs.mcgill.ca/~dprecup/courses/ML/Materials/linalg-review.pdf) - up to and including Section 3.7 - The Inverse

- **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-cheatsheet.pdf) cheat sheet
    - Texts
        - Phil Spector. (2008). *Data Manipulation with R* New York: Springer. [**Free** through Western (https://www.lib.uwo.ca/cgi-bin/ezpauthn.cgi?url=http://www.springer.com/us/book/9780387747309)]
        - 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_everyone)
    - 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/)

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

Evaluation

There will be a midterm test but no final exam. Each student will lead a brainstorming session, 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).

Daily Quizzes – 5%

Starting on the second lecture, there will be a very short quiz at the beginning of class covering the previous day's materials. The final quiz will be on 31 Oct. The lowest quiz mark will be dropped. **Quiz marks will only be excused for medical reasons.**

Midterm - 35%

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

Brainstorming Session – 5%

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: 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: 5%

Each **graduate** student 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**.

Accessibility and Support Available at Western

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 Services for Students with Disabilities (SSD) at 661-2111 ext. 82147 if you have questions regarding accommodation. Support Services 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. The website for Registrarial Services is http://www.registrar.uwo.ca.

Missed Course Components

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 WSC 140, and can be contacted at 519-661-3040 or scibmsac@uwo.ca. Their website is http://www.uwo.ca/sci/undergrad/academic_counselling/index.html. A student requiring academic accommodation due to illness must use the Student Medical Certificate (https://studentservices.uwo.ca/secure/medical_document.pdf) when visiting an off-campus medical facility. For further information, please consult the university's medical illness policy at http://www.uwo.ca/univsec/pdf/academic_policies/appeals/accommodation_medical.pdf.

Timeline (Tentative)

- 7 Sep - Lectures:
    - 12 Sep - Lectures:
- 14 Sep - Lectures:
    - 19 Sep - Lectures:
- 21 Sep - Lectures:
    - 26 Sep - Lectures:
- 28 Sep - Lectures:
    - 3 Oct - Lectures:
- 5 Oct - **Pick Brainstorming Slot by 6 Oct 5pm** - Lectures:
    - *10 Oct - **Fall Reading Week***
- *12 Oct - **Fall Reading Week***
    - 17 Oct - Lectures:
- 19 Oct - Lectures: **Guest Lecture by Amanda Holden** of SAS. Topic TBA.
    - 24 Oct - Lectures:
- 26 Oct - **Project Proposal Due 27 Oct at 5pm** - Lectures: **Guest Lecture by Dr. Kemi Ola** on Visualization
    - 31 Oct - Lectures:
- 2 Nov - Lectures:
    - 7 Nov - **Midterm**
- 9 Nov - Brainstorming: *slot1*, *slot2*, Sachi Elkerton
  9637 Slots 3:30pm-4:30pm: *slot4*, *slot5*, *slot6*, *slot7*
    - 14 Nov - Brainstorming: *slot1*, *slot2*, *slot3*, Duff Jones, *slot5*, *slot6*
- 16 Nov - **Project Draft Due 17 Nov at 5pm** - Brainstorming: Kerlin Lobo, *slot2*, *slot3*
  9637 Slots 3:30pm-4:30pm: *slot4*, *slot5*, *slot6*, *slot7*
    - 21 Nov - Brainstorming: *slot1*, Angela Zhao, *slot3*, *slot4*, *slot5*, *slot6*
- 23 Nov - Brainstorming: *slot1*, *slot2*, *slot3*
  9637 Slots 3:30pm-4:30pm: *slot4*, *slot5*, *slot6*, *slot7*
    - 28 Nov - Brainstorming: *slot1*, *slot2*, Vanessa Zhu, *slot4*, *slot5*, *slot6*
- 30 Nov - Brainstorming: *slot1*, *slot2*, *slot3*
  9637 Slots 3:30pm-4:30pm: *slot4*, *slot5*, *slot6*, *slot7*
    - 5 Dec - Brainstorming: *slot1*, *slot2*, *slot3*, *slot4*, *slot5*, *slot6*
- 7 Dec - Brainstorming: *slot1*, *slot2*, *slot3*
  9637 Slots 3:30pm-4:30pm: *slot4*, *slot5*, *slot6*, *slot7*

- **Project Document Due Friday 8 December 5pm**
- **Reviews (graduate students only) Due Thursday 15 December 5pm**

Retrieved from "https://www.csd.uwo.ca/~dlizotte/teaching/IDS/index.php?title=Introduction_to_Data_Science_I&oldid=36"