DSCI 100: Introduction to Data Science

Time and Place

Jul-Aug 2020, Tues/Thurs 1-4pm

Description

Use of data science tools to summarize, visualize, and analyze data. Sensible workflows and clear interpretations are emphasized.

Prerequisite Mathematical Knowledge

- distance between points on a graph
- percentages, average
- powers, roots, basic operations, logarithm, exponential
- equation of a line / plane

As an example, British Columbia's Math 12 or Pre-Calculus 12 courses would satisfy the prerequisite.

Textbook

We are using an open source textbook available free on the web: https://ubc-dsci.github.io/introduction-to-datascience/

Expanded Course Description

In recent years, virtually all areas of inquiry have seen an uptake in the use of data science tools. Skills in the areas of assembling, analyzing, and interpreting data are more critical than ever. This course is designed as a first experience in honing such skills. Students who have completed this course will be able to implement a data science workflow in the R programming language, by "scraping" (downloading) data from the internet, "wrangling" (managing) the data intelligently, and creating tables and/or figures that convey a justifiable story based on the data. They will be adept at using tools for finding patterns in data and making predictions about future data. There will be an emphasis on intelligent and reproducible workflow, and clear communications of findings. No previous programming skills necessary; beginners are welcome!

Course Software Platforms

Students will learn to perform their analysis using the R programming language. Worksheets and tutorial problem sets as well as the final project analysis, development, and reports will be done using Jupyter Notebooks. Students will access the worksheets and tutorials in Jupyter Notebooks through Canvas. Students will require a laptop, chromebook or tablet in both lectures and tutorials. If a student does not their own laptop or chromebook, students may be able to loan a laptop from the UBC library.

Learning Outcomes

By the end of the course, students will be able to:

- Read data using computation from various sources (local and remote plain text files, spreadsheets and databases)
- Wrangle data from their original format into a fit-for-purpose format.
- Identify the most common types of research/statistical questions and map them to the appropriate type of data analysis.
- Create, and interpret, meaningful tables from wrangled data.
- Create, and interpret, impactful figures from wrangled data.
- Apply, and interpret the output of simple classifier and regression models.
- Make and evaluate predictions using a simple classifier and a regression model.
- Apply, and interpret the output of, a simple clustering algorithm.
- Distinguish between in-sample prediction, out-of-sample prediction, and cross-validation.
- Calculate a point estimate in the context of statistical inference and explain how that relates to the population quantity being estimated.
- Accomplish all of the above using workflows and communication strategies that are sensible, clear, reproducible, and shareable.

Teaching Team

Note that your TAs may have class right before their DSCI100 office hours, so they may run a few minutes late. Please be patient!

Position	Name	Email	Office Hours	Office Location
Instructor	Melissa Lee	melissa.lee@stat.ubc.ca	Tuesday 4 PM	Collaborate Ultra
TA	Daniel Alimohd		Friday 5 PM	Collaborate Ultra
TA	Wasay Hayat		Thursday 6 PM	Collaborate Ultra
TA	Pramoda Sachinthana Jayasinghe		Friday 5 PM	Collaborate Ultra
TA	Alice Kang		Wednesday 6 PM	Collaborate Ultra
TA	Kevin Le		Thursday 6 PM	Collaborate Ultra
TA	Yutong Li		Wednesday 6 PM	Collaborate Ultra

TA	Grandon Seto	Friday 5 PM	Collaborate Ultra
TA	Jennifer Vincent	Wednesday 6 PM	Collaborate Ultra

Assessment

In each class (lecture and tutorial) there will be an assignment. **Lecture and tutorial worksheet due dates are posted on Canvas.** To open the assignment, click the link (e.g. `worksheet_01`) from Canvas. To submit your assignment, just make sure your work is saved (File -> Save and Checkpoint to be sure) **on our server** (i.e., using the link from Canvas) before the deadline. Our server will automatically snapshot at the due date/time.

Course breakdown

Deliverable	Percent Grade
Lecture worksheets	5
Tutorial problem sets	15
Group project	20
Two quizzes	40
Final exam	20

Group project breakdown

Deliverable	Percent Grade
Proposal	3
Final report	10
Team work	5
Peer review	1
Group contract	1

Schedule

Lectures are held on Tuesdays. Tutorials are held on Thursdays and build on the concepts learned in lecture.

Topic	Description
Chapter 1: Introduction to Data Science	Learn to use the R programming language and Jupyter notebooks as you walk through a real world data Science application that includes downloading data from the web, wrangling the data into a useable format and creating an effective data visualization.
Chapter 2: Reading in data locally and from the web	Learn to read in various cases of data sets locally and from the web. Once read in, these data sets will be used to walk through a real world data Science application that includes wrangling the

	data into a useable format and creating an effective data visualization.
Chapter 3: Cleaning and wrangling data	This week will be centered around tools for cleaning and wrangling data. Again, this will be in the context of a real world data science application and we will continue to practice working through a whole case study that includes downloading data from the web, wrangling the data into a useable format and creating an effective data visualization.
Chapter 4: Effective data visualization	Expand your data visualization knowledge and tool set beyond what we have seen and practiced so far. We will move beyond scatter plots and learn other effective ways to visualize data, as well as some general rules of thumb to follow when creating visualations. All visualization tasks this week will be applied to real world data sets. Again, this will be in the context of a real world data science application and we will continue to practice working through a whole case study that includes downloading data from the web, wrangling the data into a useable format and creating an effective data visualization.
Transition week	Quiz 1
Chapter 6: Classification	Introduction to classification using K-nearest neighbours (k-nn)
Chapter 7: Classification, continued	Classification continued
Chapter 8: Regression	Introduction to regression using K-nearest neighbours (k-nn). We will focus on prediction in cases where there is a response variable of interest and a single explanatory variable.
Chapter 9: Regression, continued	Continued exploration of k-nn regression in higher dimensions. We will also begin to compare k-nn to linear models in the context of regression.
Transition week	Quiz 2
Chapter 10: Clustering	Introduction to clustering using K-means
Chapter 11: Introduction to statistical inference	Introduce sampling and estimation for sample means and proportions.
Chapter 12: Introduction to statistical inference, continued	Introduce confidence intervals, and calculating them via boostrapping.
Exam period	Final Exam

Policies

Late/Absence

Regular attendance to lecture and tutorials is expected of students. Students who are unavoidably absent because of illness or other reasons should inform the instructor(s) of the course as soon as possible, preferably, prior to the start of the lecture/tutorial. Students who miss quizzes 1 or 2 or an assignment need to provide a self-declaration and make arrangements (e.g., schedule an oral make-up quiz) with the Instructor as soon as possible. Failing to present a declaration may result in a grade of zero.

Late lecture and tutorial worksheets will receive a grade of 0. For other assessments, a late submission is defined as any work submitted after the deadline. For a late submission, the student will receive a 50% deducation of their grade for the first occurrence. Hence a maximum attainable grade for the first piece of work submitted late is 50%. Any additional pieces of work that are submitted late will receive a grade of 0 for subsequent occurrences.

Autograder Policy

Many of the questions in assignments are graded automatically by software. The grading computer has exactly the same hardware setup as the server that students work on. No assignment, when completed, should take longer than 5 minutes to run on the server. The autograder will automatically stop (time out) for each student assignment after a maximum of 5 minutes; **any ungraded questions at that point will receive a score of 0.**

Furthermore, students are responsible for making sure their assignments are *reproducible*, and run from beginning to end on the autograding computer. In particular, **please ensure that any data that needs to be downloaded is done so by the assignment notebook with the correct filename to the correct folder.** A common mistake is to manually download data when working on the assignment, making the autograder unable to find the data and often resulting in an assignment grade of 0.

In short: whatever grade the autograder returns after 5 minutes (assuming the teaching team did not make an error) is the grade that will be assigned.

Re-grading

If you have concerns about the way your work was graded, please contact the TA who graded it within one week of having the grade returned to you. After this one-week window, we may deny your request for re-evaluation. Also, please keep in mind that your grade may go up or down as a result of re-grading.

Device/Browser

Students are responsible for using a device and browser compatible with all functionality of Canvas. Chrome or Firefox browsers are recommended; Safari has had issues with Canvas quizzes in the past.

Missed Final Exam

Students who miss the final exam must report to their faculty advising office within 72 hours of the missed exam, and must supply supporting documentation. Only your faculty advising office can grant deferred standing in a course. You must also notify your instructor prior to (if possible) or immediately after the exam. Your instructor will let you know when you are expected to write your deferred exam. Deferred exams will ONLY be provided to students who have applied for and received deferred standing from their faculty.

Academic Concession Policy

Please see UBC's concession policy for detailed information on dealing with missed coursework, quizzes, and exams under circumstances of an acute and unanticipated nature.

Academic Integrity

The academic enterprise is founded on honesty, civility, and integrity. As members of this enterprise, all students are expected to know, understand, and follow the codes of conduct regarding academic integrity. At the most basic level, this means submitting only original work done by you and acknowledging all sources of information or ideas and attributing them to others as required. This also means you should not cheat, copy, or mislead others about what is your work. Violations of academic integrity (i.e., misconduct) lead to the breakdown of the academic enterprise, and therefore serious consequences arise and harsh sanctions are imposed. For example, incidences of plagiarism or cheating may result in a mark of zero on the assignment or exam and more serious consequences may apply if the matter is referred to the President's Advisory Committee on Student Discipline. Careful records are kept in order to monitor and prevent recurrences.

A more detailed description of academic integrity, including the University's policies and procedures, may be found in the Academic Calendar at http://calendar.ubc.ca/vancouver/index.cfm?tree=3,54,111,0.

Plagiarism

Students must correctly cite any code or text that has been authored by someone else or by the student themselves for other assignments. Cases of plagiarism may include, but are not limited to:

- the reproduction (copying and pasting) of code or text with none or minimal reformatting (e.g., changing the name of the variables)
- the translation of an algorithm or a script from a language to another
- the generation of code by automatic code-generation software

An "adequate acknowledgement" requires a detailed identification of the (parts of the) code or text reused and a full citation of the original source code that has been reused.

The above attribution policy applies only to assignments. **No code or text may be copied (with or without attribution) from any source during a quiz or exam. Answers must**

always be in your own words. At a minimum, copying will result in a grade of 0 for the related question.

Repeated plagiarism of any form could result in larger penalties, including failure of the course.

Attribution

Parts of this syllabus (particularly the policies) have been copied and derived from the UBC MDS Policies.