DSCI 100: Introduction to Data Science

Time and Place

Sept-Dec 2019, Tues/Thurs 12:30 - 2:00 pm, ORCH 4074

Description

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

Prerequisite

MATH 12

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:

- Download and scrape data off the world-wide-web.
- Wrangle data from their original format into a fit-for-purpose format.
- Create, and interpret, meaningful tables from wrangled data.
- Create, and interpret, impactful figures from wrangled data.
- Apply, and interpret the output of, a simple classifier.
- Make and evaluate predictions using a simple classifier.

- Apply, and interpret the output of, a simple clustering algorithm.
- Apply, and interpret the output of, a regression model.
- Make and evaluate predictions using a regression model.
- Distinguish between in-sample prediction, out-of-sample prediction, and cross-validation.
- Apply and interpret a bootstrap analysis in a regression context.
- Accomplish all of the above using workflows and communication strategies that are sensible, clear, reproducible, and shareable.

Learning outcomes per lecture are available here.

Teaching Team

Position	Name	email	office hours	office location
Instructor	Tiffany Timbers	tiffany.timbers@stat.ubc.ca		
Instructor	Trevor Campbell	trevor@stat.ubc.ca	Thursday 2pm	ESB3116
TA	Daniel Alimohd			
TA	Alex Chow			
TA	Jordan Bourak			
TA	Grandon Seto			
TA	Petal Vitis			

Assessment

Course breakdown

Deliverable	% grade
Lecture worksheets	5
Tutorial problem sets	15
Group project	20
Three quizzes	60

Group project breakdown

Deliverable	% grade
Proposal	3
Peer review	2
Final report	10
Team work	5

- *It is necessary to pass the final examination to pass the course.*
- *Specific dates for each assessment item are listed here and will be posted on Canvas.*

Schedule

Lectures are held on Thursdays. The tutorials happen on Tuesdays and build on the concepts learned in lecture.

Lecture date	Topic	Description	Lecture pre-reading
	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.	Introduction to Data Science
	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.	Reading in data locally and from the web

Lecture date	Topic	Description	Lecture pre-reading
	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.	

Lecture date	Topic	Description	Lecture pre-reading
	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	

Lecture date	Topic	Description	Lecture pre-reading
	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.	
	Transition week	Quiz 2	
	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.	

Lecture date	Topic	Description	Lecture pre-reading
	Chapter 10: Bootstrap applied to regression	This week will introduce the bootstrap, first by visualizing bootstrap samples and their fitted regression lines for cases where there is a response variable of interest and a single explanatory variable. An intuitive case will be made for what the ensemble of slopes represents, Then we work through examples from multiple regression, emphasizing the scientific interpretation and relevance of the mix of negative/positive slopes. We will emphasize that this is a jumping off point for the study of statistical inference.	
	Chapter 11: Clustering Data Science wrap-up & Work on group project in class	Introduction to clustering using K-means	

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 a quiz or assignment need to provide a doctor's note and make arrangements (e.g., schedule an oral make-up quiz) with the Instructor as soon as possible. Failing to present a doctor's note may result in a grade of zero.

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.

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.

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.

Code Plagiarism

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

- the reproduction (copying and pasting) of code 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-generations software

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

Attribution

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