STA314: Statistical Methods for Machine Learning I (Fall 2019)

Instructor: Dushanthi Pinnaduwage

Office Hours: Monday 14:00–15:00 (Public Health Sciences at 155 College St in open area (381))
Tuesday 15:00–16:00 (Public Health Sciences at 155 College St in open area (381))
Other times: I will be around after class for 30min to answer questions.

Email: d.pinnaduwage@utoronto.ca

Email Policy: I'll try to get back to you within 2 work days, so if you email me at 7pm Wednesday, I'll get back by 5pm Friday. I don't check email on weekends.

Lectures:
Monday 11:00–13:00, NF 003 (Northrop Frye Hall, 73 Queen's Park Crescent East)
Tuesday 13:00–14:00, BR 200 (Brennan Hall, 81 St. Mary Street)

Tutorials (choose 1):
Friday 13:00–14:00 RS 310, BA 2195
Friday 12:00–13:00 WB 119, WB 219, BA 2165

Teaching Assistants (TAs):
Michal Malyska
David Veitch
Jinda Yang
Lu Yu

Course Outline:
This course introduces machine learning to students focussing on standard statistical learning algorithms used in supervised and unsupervised learning. The course will concentrate more on the applications of the algorithms. At the end, the course will introduce some concepts of linear algebra, probability theory and maximum likelihood estimation in order for student to smoothly transfer to the course 'Statistical Methods for Machine Learning II' offered in the winter term.

What you will learn:

- Statistical modeling vs. statistical learning and machine learning

- Supervised vs unsupervised learning, regression vs classification, overfitting and generalization, and Bias-variance trade-off
- Linear regression and least squares, logistic regression, and linear discriminant analysis
- Resampling methods: cross-validation and bootstrap
- Improving linear model by Feature (predictor) selection*: best subset selection methods* such as stepwise selection and use of AIC, BIC etc., *shrinkage methods* such as ridge regression and Lasso, and *dimensionality reduction methods* such as principal components analysis (PCA) and partial least squares (PLS)
- Clustering methods in unsupervised learning
- Concepts of probability theory and maximum likelihood estimation

Assessment:

Final Exam: 55% (Tentatively on 16 December, location to be announced)

- Two and Half hours
- Covers entire course (1/3 from first half, 2/3 from second half)

Mid Term Test: 25% (15 October, 13:00–14:00)

- One Hour (in lecture time)
- No make ups. If you cannot make the exam, you get 0.
- The test will be written in a room other than the lecture room (location to be announced).
- If the test is missed for a valid reason, you must provide appropriate documentation, such as the University of Toronto Medical Certificate, University of Toronto Health Services Form, or College Registrar's Letter. You must submit this documentation within one week of the test. If documentation is not received in time, your test mark will be zero. If a test is missed for a valid reason, its weight will be shifted to the final exam.

Homework: 20%

- 5 throughout the semester worth 4% each
- Late policy: If you are traveling, you may email your solution to your TA or course instructor in advance of the deadline. 10% of the homework value will be deducted for each day a homework is late. No credit will be given for homework submitted after solutions have been posted. Exceptions will be made for documented emergencies.
- Material covered in these assignments will be assessable on the exams
- Due 27 September, 11 October, 25 October, 15 November, 22 November. All assignments will be due at 12:00pm on the due day.
- There will be an optional 6th homework assignment due 12:00pm on 6 December. If you choose to do this, I will use the best five of the homework assignments when computing your final mark.

Lateness policy:

Homeworks are due sharply at the appointed time and will receive significant penalties if late**.**

Re-grading policy:

Re-grading requests should only be made for genuine grading errors, and should be initiated by writing or typing a complete explanation of your concern (together with your full name, student number, and e-mail address) on a separate piece of paper, and giving this together with your original unaltered homework/test paper to the instructor within one week of when the graded item was first available. Warning: your mark may end up going down rather than up.

Textbook and slides:

- Our text will be *An Introduction to Statistical Learning with Applications in R* by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani.
- The book is freely available online from this address: http://faculty.marshall.usc.edu/gareth-james/ISL/ISLR%20Seventh%20Printing.pdf (PDF-ISLR - University of Southern California)
- Occasionally we will go beyond the textbook, in which case alternate references will be provided.
- Lecture slides will be available on Quercus before each lecture.
- Tutorials will be available on Quercus. Posting days will be announced in the class.

Computing:

- The course will be run using the R computing environment.
- Download R (https://cran.r-project.org) or RStudio (https://www.rstudio.com), which are free.
- All instructions in the course will assume that you have the latest version of both RStudio and R installed. We will not answer any R related questions unless both of these things are true.
- The best resource for R help is always google.