STAT5703 Data Mining I Winter 2021

This is an **Asynchronous** course - an online course where the instructor and students share information, ideas, and learning experiences in a **virtual course space**. While there is a scheduled time associated with the course *for registration*, students move through course material on their own schedule. (Please try also to hold the class time for possible Zoom or Teams meetings to discuss topics and ask questions.)

Class time: Tuesdays 5:35 pm-8:25 pm Jan. 12-April 13, 2021

Labs start the week of January 18.

Instructor: Dr. Shirley Mills

5203 Herzberg Bldg.,(Note: U. offices are closed due to COVID-19)
Phone: 613-825-0480 (home)
Email: smills@math.carleton.ca
Webpage: math.carleton.ca/~smills

Office hours: via Zoom (TBA). These will be posted on CuLearn.

Extensive course notes will be posted. **Please note that all of my course materials are protected by copyright. You may make a copy for your own use but you may not sell or distribute these materials without my prior written consent.**

In addition I recommend the following texts:

- Elements of Statistical Learning 2nd Ed. , Trevor Hastie, Robert Tibshirani, Jerome Friedman A free electronic version obtainable at
  http://www-stat.stanford.edu/~tibs/ElemStatLearn/

- Pattern Recognition and Machine Learning, Christopher Bishop

Software: The course will use free software:

R available at http://www.cran.r-project.org/
Ggobi available at www.ggobi.org
Tinn-R available at http://sourceforge.net/projects/tinn-r/
Rstudio available at http://www.rstudio.com/

Ottawa U students: You must fill out a form and submit it to FGPA to be able to access CULearn. The form can be found at https://gradstudents.carleton.ca/wp-content/uploads/Access-to-CULearn.pdf Once filled out, email it to FGPA. The instructions are on the form.

Tentative Course Outline:

Topic 1: What is Data Mining, Intro. to R and Ggobi
Data visualization: Scatterplots, scatterplot matrix, Co-plots, grand tour, brushing, linking, parallel coordinates plots, cluster separation, Stereo

Topic 2: Association rules/ market basket analysis: support, confidence, lift.

Topic 3: Data Reduction- Modelling: Polynomial Interpolation, overfitting, splines, Running means, Regression: linear, quadratic, cubic; piecewise: constant, linear; all subsets, Ridge, training/test sets, Cross Validation, best subset, LOESS, Supersmoother.

Topic 4: Dimension Reduction: Curse of dimensionality, Principal Components (PCA), Factor analysis, Extensions for nonlinearity: Principal Curves and Surfaces, NLCA; low-d representations [MDS and SOM]

Topic 5: Dimension Reduction: Projection Pursuit and Independent Component Analysis.

Topic 6: Case/Data Reduction - Clustering (unsupervised learning): Voronoi tesselations and one-nearest neighbour, Vector Quantization (VQ), K-means, Image compression, LBG-VQ.

Topic 7 : Self-organizing Maps (SOM), Local PCA.

Topic 8: Classification (supervised learning): Neural Nets

Topic 9: Classification: Logistic, k-nearest neighbour (kNN), Discriminant Analysis: Linear, Quadratic, Flexible, Penalized.

Topic 10: Classification: Classification Trees (recursive partitioning), prune, snip, bag, boost.

Topic 11: Data Reduction/ Modeling Revisited: Regression Trees, Neural nets for regression, Project Pursuit Regression (PPR), GAMS, Multivariate Adaptive Regression Splines (MARS).

Marking scheme:
2 assignments @ 10% + 20%; research project @ 20%; final project (Take-home) @ 50%
Collaboration is permitted and encouraged unless otherwise indicated by the course professor. However, not all work to be evaluated may be done collaboratively; course professor will indicate when collaboration is NOT permitted.

Notices:

1. You need a computer that can run some version of R (free software) and a webcam/mic to participate in labs and discussions. Notes, sample code, and videos will be posted, along with a suggested weekly pace but students may work at their own pace. Please do not try to cram – there is a lot of material and it builds upon earlier material in the course so please follow the order of presentation. If you do not understand something, ask the TA or contact me via email or phone.
2. Communication of course material will be via CuLearn as well as Zoom or Teams.
3. Students MUST send email from their Carleton email accounts for all course-related correspondence. Responses generally will be the same day.

4. Late assignments are NOT accepted unless you have received prior authorization from me. In that case, I will advise you how and by when to submit material for grading.

Other info and resources:

1. Classes start the week of January 11 and end the week of April 5, 2021..
Beyond this, I refer you to the official listing of academic dates for the details on holidays, the Fall Break, etc. http://calendar.carleton.ca/academicyear/

2. For Academic Regulations:
Please refer to
https://calendar.carleton.ca/grad/gradregulations/

3. Academic Accommodation:
Please refer to https://calendar.carleton.ca/search/?search=academic+accommodation
You may need special arrangements to meet your academic obligations during the term. For an accommodation request, the processes are as follows:

a. Pregnancy obligation: Write to me with any requests for academic accommodation during the first two weeks of class, or as soon as possible after the need for accommodation is known to exist.
b. For students with disabilities: Please refer to **https://calendar.carleton.ca/undergrad/regulations/academicregulationsoftheuniversity/regulations-for-students-with-disabilities/**
c. For students with Religious obligations: Please refer to https://calendar.carleton.ca/undergrad/regulations/academicregulationsoftheuniversity/regulations-for-students-with-religious-obligations/

4. Students wishing to see their examination papers must make an appointment within one week of the examination results being posted; it is not an opportunity to argue about the marking!

5. Academic Integrity:
Please refer to

http://www2.carleton.ca/senate/ccms/wp-content/ccms-files/academic_integrity_policy-21.pdf

6. Please note that all course materials are protected by copyright. You may make a copy **for your own use** but **you may not sell or distribute these materials** without prior written consent.

7. **TA opportunities** within the School for future terms. Information on how to apply can be found on our School web page. In hiring undergraduate TAs, the priority shall first be given to students who have passed some of the following Honours courses: MATH 1002, 1102, 2000, 2100, STAT 2655, 2559 with grades A- or better.