STAT4601/5703 Data Mining I Winter 2020

Meeting: Thursdays January 9- April 2, 2020
Time: 17:35 - 20:25 for lecture. **Labs start the week of January t13th.** You must also register in one of the labs for this course.

Professor: Dr. Shirley Mills,
5203 Herzberg Bldg.
Email: shirley_mills@carleton.ca
Webpage: http://www.math.carleton.ca/~smills
Phone:613-520-2600 ext 2199 (office); 613-825-0480 (home - if urgent only!)

Office Hours: Thursday 2:30-4:00 p.m.; **other hours by appointment only**

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

Pattern Recognition and Machine Learning, Christopher Bishop

Software: The course will use the 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/
Also available is Cloud computing on SAS webservers for using SAS Enterprise Miner.

Course Material will be posted in CuLearn (Ottawa U students: You must 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 and lift.

Topic 3: 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: Projection Pursuit and Independent Component Analysis.

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

Topic 7 (STAT5703 only) : Self-organizing Maps (SOM), Local PCA

Topic 8: Classification (supervised learning) Neural Nets

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

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

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

Grading: STAT5703: 2 assignments @ 10% + 20%; research project @ 20%; final project @ 50%
STAT4601: 3 assignmeents @ 10% +15% +25%; final project @50%

NO LATE ASSIGNMENTS WILL BE ACCEPTED WITHOUT MY PRIOR AUTHORIZATION. 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. **Academic Year** :

https://calendar.carleton.ca/academicyear/

2. **Holy days**:
https://carleton.ca/registrar/academic-integrity/

3.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!

4. **Academic Integrity**
https://carleton.ca/registrar/academic-integrity/

5. **Paul Menton Centre** (for students with disabilities)
http://www2.carleton.ca/pmc/information-for-faculty/accommodation-statement-for-course-outline/

6. More **general accommodation requests** such as pregnancy or religious obligations. See (includes PMC statement):
http://www2.carleton.ca/equity/accommodation/academic/course-outline-wording/

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.

8. 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.