Statistical Learning with Big Data
2014 AARMS summer school course
Course outline
(July 23,, 2014 version)

Instructors:
Hugh Chipman, Acadia University, hugh.chipman@gmail.com
Xu (Sunny) Wang, Saint Francis Xavier University, xwang@stfx.ca
Office hours 1:00 - 3:00 MWF during lab time (see below)

Website:
http://www.mathstat.dal.ca/~aarms2014/StatLearn/

Text:
Lectures will follow the text "An Introduction to Statistical Learning with Applications in R" by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. Students are responsible for obtaining their own copies of the text. A free pdf version of the text is available for download from Gareth James' website, http://www-bcf.usc.edu/~gareth/ISL/

(Advanced students may find the text "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Second Edition, by Trevor Hastie, Robert Tibshirani and Jerome Friedman interesting. It is, however a PhD level book. Free online at http://statweb.stanford.edu/~tibs/ElemStatLearn/)

Software and computing:
Students are encouraged to bring a laptop to the course if they are able. Computers are also available in lab Dunn 301A for everyone to use.

The course will make extensive use of the (free) R software environment for statistical computing and graphics. Introductory tutorials/labs will be provided for students unfamiliar with R. Students are encouraged to install R on their laptop before the summer school begins. Copies of R may be downloaded from http://www.r-project.org/ . R runs on a wide variety of UNIX platforms, Windows and MacOS. We encourage students to use the RStudio add-on interface to R (www.rstudio.com) as it provides a convenient set of tools and a uniform interface across different computers.

Outline:
Week 1: Introduction (Ch 1, 2), regression (Ch 3), classification (Ch 4)
Week 2: Resampling methods (Ch 5), other supervised learning methods, including Support Vector Machines (Ch 9)
Week 3: Model selection and regularization (Ch 6), moving beyond linear methods (Ch 7)
Week 4: Tree-based methods and ensembles (Ch 8), unsupervised learning (Ch 10) and statistical graphics.

Lectures:

- Lectures are daily Monday - Friday, from July 21 - August 15.
- Monday August 4 is a holiday and there is no lecture that day.
- Lectures are 9:00 - 10:30 in Dunn 305.
- Hugh Chipman will teach weeks 1 & 2, Sunny Wang will teach weeks 3 & 4.

Labs are from 1-4:30pm daily in Dunn 301A. There will be an introductory lab on R in week 1. For other lab time, attendance is optional. Instructors will try to be in the lab part of the time each day.

Evaluation

Assignment 1	July 21	July 25	20%
Assignment 2	July 25	August 1	20%
Assignment 3	August 1	August 8	20%
Test	August 8	August 8	20%
Assignment 4	August 5	August 15	20%

Assignments are due in class on the day indicated.
The test will be in the lab timeslot (1-2:30pm) on Friday August 8.