Syllabus for STAT 3654: Introduction to Data Analytics and Visualization

Spring 2014

Times: TR 9:30-10:45 GYM 219
Instructors: Eric P. Smith, Leanna House, Scotland Leman
Office: 406A Hutcheson Hall
Phone: 231-5657
Email: epsmith@vt.edu, lhouse@vt.edu, leman@vt.edu

My Office hours: 9:30-10:30 Monday –Wednesday or by appointment

Textbook: **Data Mining with Rattle and R:** The Art of Excavating Data for Knowledge Discovery, Graham **Williams, 2011**

Other material will be added to the scholar site as needed.

This course will provide background in the techniques in data analytics, in 'learning from data' and some of the tools that are unique to analysis of large data sets. Some of the topics will include the collection of, storing, accessing, and manipulating standard-size and large datasets; data visualization; predictive (supervised learning) analytics and clustering (unsupervised learning). We plan on using some actual data sets for case studies and to motivate the methods. A good background in mathematics, some statistics and computational skills will be required to take the class.

Prerequisites: This course requires CMDA 2006. A recommended set of courses that would be equivalent:

STAT 3005-6: Statistical Methods
STAT 3104: Probability and Distributions
MATH 1114: Elementary Linear Algebra
MATH 1224: Vector Geometry
MATH 2214: Introductory Differential Equations
MATH 2224: Multivariable Calculus

Computing: We will use the R package. This is available at ://cran.us.r-project.org/. There are packages for linux, mac and windows machines. Also download the packages Rstudio and rattle and make sure they load properly. See the handouts for some other details. Rattle may be tricky.

Grading:	Homework	25%
	Quizzes/presentations	30%
	Participation	5%
	Final Project	40%

Some policies: Homework will be assigned on a regular basis (1 maybe 2 assignments per week). We expect that the work will be neat, tidy, preferably typed. If any code is developed/required please hand in code as an appendix to the homework. We expect code to have comments so that one can understand what you are doing. Late homework is not accepted without prior permission. We

especially do not like output that is simply attached as a solution to a problem. It needs to be summarized or annotated.

Final Exam: The written project will be due the day of the final, by 5PM.

Topics may include (not necessarily in the order covered)

1. Basics of Analytics
    a. Examples
    b. Data collection and cleaning
    c. Terminology, types of analyses
2. R package
3. Data, Database and sql
4. Visualization tools
    a. Basic plots
    b. Visualizing distributions
    c. Visualizing patterns
    d. Dimension reduction
5. Analytics tools
    a. Predictive analytics
        i. Bayes methods
        ii. Linear discriminant analysis
        iii. Logistic regression
        iv. Classification trees
        v. Model evaluation and validation
    b. Segmentation/clustering
        i. K-means
        ii. Hierarchical
6. Other topics as time permits