Statistical Learning for Analytics (ANLY-512)

Preliminary Syllabus, Spring 2016

This course is concerned with algorithms that use statistical techniques to find structure or patterns in given data (unsupervised learning) or use given instances of data to predict outcomes in new cases (supervised learning). A well-known method for supervised learning is linear regression, and this will be covered early in the course. Statistical methods for making discrete predictions (classification) such as logistic regression will also be covered. Special emphasis will be placed on techniques for handling high-dimensional data (i. e. instances with many attributes), including variable selection and dimension reduction. The course will also cover ensemble methods such as bagging and boosting that are often used to improve the results of given classification methods. Unsupervised methods covered in this course include model-based and hierarchical clustering. The course will use statistical software (**R**) throughout.

Instructor: Hans Engler, 307 St. Mary's Hall, Phone (202)687-6751 and (301)938-9726 (mobile)
Email: engler@georgetown.edu
Skype ID: gumathprof

Office hours: *To be announced*

Room & Time: Car Barn 301, Tuesdays 6:30PM – 9:00PM.

Textbook: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, In Introduction to Statistical Learning with Applications in R. Springer 2013, corrected 6th printing 2015. Softcover ISBN 978-1-4614-7137-0. Available as free legal download at http://www-bcf.usc.edu/~gareth/ISL/

Course Videos: Available at
http://www.r-bloggers.com/in-depth-introduction-to-machine-learning-in-15-hours-of-expert-videos/
Students are expected to watch videos and read the corresponding book sections before class.

Software and Computers: R will be used throughout the course. It is available as a free download at https://cran.r-project.org/. I recommend that you also obtain RStudio which is available for free at https://www.rstudio.com/. This software is also installed on all Georgetown University Information Services (UIS) computer labs. You are expected to be proficient in R and to be able to use it on weekly assignments. Any laptop with 2GB or more of RAM running Win 7 or higher, Mac OS X or a recent dialect of Linux is sufficient.

Internet: The course will use Blackboard, https://campus.georgetown.edu/. Announcements, homework assignments and solutions, course material such as documentation, links, data sets and R code will all be posted there. You can look up your grades, and online surveys will also be conducted here. Students in this course are required to visit this page once a day. Announcements will usually only appear on the web page. You will also have to access data sets on the Internet and need to be comfortable with this.

Prerequisites: ANLY-511 (Probabilistic Modeling and Statistical Computing) or equivalent, Linear Algebra, Multivariable Calculus. Proficiency in R.

Topics to be covered:

- Basic concepts: Model accuracy, prediction accuracy, interpretability, supervised and unsupervised learning.
- Linear regression.
- Classification, logistic regression, linear discriminant analysis.
- Resampling methods, cross validation.
- Model selection, dimension reduction, and other high-dimensional considerations.
- Support vector machines.
- Unsupervised methods such as PCA and Clustering.
- If time permits: Splines, general additive models, tree-based methods.

Grading: About 12 homework sets = 25%, 1 hour in-class mid-term exam = 20%, comprehensive 2 hour final exam with in-class and take-home components = 40 %, class participation = 15%.
This is subject to modification.

Homework for grading will be assigned weekly on the course web page. For **class participation**, expect to be called up randomly to demonstrate practice problems, explain or suggest code, explain concepts that were discussed previously, formulate questions concerning new material etc.

Grading scheme:

A	95% or more	A-	90% or more	B+	85% or more
B	80% or more	B-	75% or more	C	65% or more
D	55% or more				

The break points for these grades may be lowered, but will not be raised.

Honor Code: Please be aware of the academic integrity rules. They may be found in ch. VI of the the Graduate Bulletin. Academic misconduct includes plagiarism, unacknowledged paraphrase, cheating, fabrication of data, fabrication, alteration, or misrepresentation of academic records, facilitating academic dishonesty (i.e. helping or allowing others to violate these rules), unauthorized collaboration. misuse of otherwise valid academic work , misuse of academic resources, and depriving others of equal access to academic resources. Please look at the Graduate Bulletin for a detailed explanation, and stick to these rules.

In this class, you are encouraged to collaborate with other students when you study and when you do your homework. Some in-class work will also be in small groups. When working on a homework assignment or a practice exercise, start by yourself, then talk to other students, ask questions, and share your ideas, then complete the work on your own. Do not copy homework from others and do not permit others to copy your work, as this will be considered plagiarism or facilitating plagiarism. Do not use help from outside (e.g. online). You are not allowed to collaborate with other students or seek any human help on any exams. Closed notes, closed book, no electronic devices such as computers, smart phones or calculators on any in-class exams. Computers, notes, and books are allowed for take-home portions of exams, but no human help is allowed.

Important dates:

01/19/16:	First class meeting.
01/22/16:	Last day for add/drop
03/08/16:	No class (Spring break)
04/4-16/16:	Preregistration for Fall 2016
04/26/16:	Last class meeting
04/27/16:	Last day for graduate students to withdraw from a course

Dates for the midterm and final exam will be announced later.