Statistics UN3106-001
Applied Data Mining
Spring 2017

Instructor
Gabriel Young
Email: gjy2107@columbia.edu
Office: Room 610, 6th floor of Watson Hall, 612 West 115th Street

Teaching Assistant
Phyllis Wan
Email: pw2348@columbia.edu

Office Hours: TBA

Course Description
Massive data collection and storage capacities have led to new statistical questions:

- Amazon collects purchase histories and item ratings from millions of its users. How can it use these to predict which items users are likely to purchase and like?
- Yahoo news acts as a clearinghouse for news stories and collects user click-through data on those stories. How should it organize the stories based on the click-through data and the text of each story?
- Advances in molecular biology have allowed scientists to gather massive amounts of genomic data. How can this be used to predict gene interactions?
- Large medical labs can receive thousands of tissue and cell samples per day. How can they automatically screen cancerous specimens from non-cancerous ones, preferably with a higher accuracy than doctors?
- Facebook gets millions of photographs annotated by its users. How can it use this data to automatically detect who is in newly uploaded photos?

Many new problems in science, industry, arts and entertainment require traditional and non-traditional forms of data analysis. In this course, you will learn how to use a set of methods for modern data mining: how to use each method, the assumptions, computational costs, how to implement it, and when *not* to

use it. Most importantly, you will learn how to think about and model data analysis problems.

Contact Hours
STAT UN3106 is a 3 credit hour course. The class meets TR from 2:40pm-3:55pm, 233 Seeley W. Mudd Building.

Prerequisites for UN3106
Calculus (integration and differentiation), linear algebra (matrix computations, eigenvalues), and basic probability theory (expectations, conditioning, distributions). Note: There will be some multivariate calculus in this class.

Text/Supplies

- James, G., Witten, D. Hastie, T. and Tibshirani, R. *An Introduction to Statistical Learning* Springer, 2014.
- Torgo, L. *Data Mining with R.* CRC Press, 2011.
- Hastie, T., Tibshirani, R. and Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference and Prediction, 2nd Edition.* Springer, 2009.
- Maybe another textbook.
- Statistical Software `R`

Course Structure
The class follows a traditional lecture environment. New topics are presented in each class. Lecture notes, in class examples and other relevant course materials will be posted on Canvas. Students should print class examples/notes out before each lecture. The solutions are covered during the scheduled contact hours.

Attendance
Students should ideally attend each class meeting. Even though attendance is not required, I will frequently give examples or hints during lecture that may show up on assessments.

Grading

Homework	40%
Midterms	25%
Final	35%

Homework

Homework will contain both written and `R` data analysis elements. Homework assignments will be posted on Canvas at least one week in advance of the due date.

You are encouraged to discuss homework problems with your classmates, but all work submitted must be your own. If multiple students turn in identical solutions, all of them will receive a zero.

Homework Guidelines

1. (a) Homework can be submitted in class before 2:40pm on the stated due date; homework submissions are not accepted during or after lecture.
   (b) Homework can be submitted to the appropriate mailbox in Room 904 SSW within 24 hours of the nominal deadline.
   (c) Zero credit will be awarded for papers submitted after the final deadline.
2. Use only $8\frac{1}{2}$ by 11 paper.
3. Any R portions must be compiled as a pdf using R Markdown.
4. Multiple pages should be stapled together in the top left corner.
5. Your name, the course and section number (STAT UN3106 Section 001), and the assignment number should appear at the top of the first page. Your first and last name must be clearly legible.
6. Homework should be typed or handwritten neatly, and well-organized. Problems should be clearly labeled, and presented in the order assigned (i.e., problem 4 should not precede problem 3 in your write-up).

7. Leave sufficient margins and whitespace that the grader may mark point deductions and other comments.
8. Homework submitted on notebook paper with frayed edges will not be accepted.
9. Do not include plastic binder covers.
10. Any plots created form must statistical software must have titles labeled axis.

A portion of the grade on each assignment will be based on presentation; any paper that fails to comply with the above requirements will not earn the presentation points.

Midterms There will be one midterm exam. The midterm is scheduled for

- 03/02/2017

Final
The final exam will be weighted 35% of the final course grade. The final exam is tentatively scheduled for

- **TBA**

You must take the final at the scheduled time.

Exams
You will have 75 minutes to complete the midterm exam, and 150 minutes to complete the final. Allowable materials will be discussed two weeks prior to the exam. There should be absolutely no communication between students during a test. The sharing of materials (such as a calculator) is strictly prohibited.

Exam absences
Make-up exams will not be given routinely. If you have a legitimate conflict with an exam date, it is incumbent upon you to make arrangements with the instructor to take the test early. An exam missed due to a documented illness or other unforeseeable (and documented) extraordinary circumstances must be made up before the test papers are returned to the class.

Grading Scale

93 or more	A
90 to 92	A-
87 to 89	B+
83 to 86	B
80 to 82	B-
77 to 79	C+
70 to 76	C
60 to 69	D
Below 59	F

The letter grade $A+$ will be given to the top two students in the class (assuming they earned a final grade of at least 93%).

Academic Honesty
The university expressly prohibits academic dishonesty such as cheating, plagiarism, etc. It provides for a number of rather unpleasant consequences for students who are caught in violation of its academic honesty policies. Any suspected cheating on examinations will be referred to the Dean?s Discipline process, possibly resulting in course failure or College dismissal.

Tentative course outline follows on the next page

Date	Topic	Recommended Readings
01/17/2017	Introduction	TBA
01/19/2017	Introduction to R	TBA
01/24/2017	Probability Models	TBA
01/26/2017	Dimension Reduction I	TBA
01/31/2017	Dimension Reduction II	TBA
02/02/2017	Supervised Learning I	TBA
02/07/2017	Supervised Learning II	TBA
02/09/2017	Linear Models I	TBA
02/14/2017	Linear Models II	TBA
02/16/2017	Logistic Regression	TBA
02/21/2017	Linear Discriminant Analysis	TBA
02/23/2017	Naive Bayes	TBA
02/28/2017	Cross Validation	TBA
03/02/2017	Exam 1	TBA
03/07/2017	The Bootstrap	TBA
03/09/2017	Subset Selection	TBA
03/14/2017	Spring Recess	TBA
03/16/2017	Spring Recess	TBA
03/21/2017	Shrinkage	TBA
03/23/2017	Trees I	TBA
03/28/2017	Trees II	TBA
03/30/2017	Boosting	TBA
04/04/2017	Bagging and Random Forests	TBA
04/06/2017	Scalable Machine Learning	TBA
04/11/2017	SVMs I	TBA
04/13/2017	SVMs II	TBA
04/18/2017	SVMs III	TBA
04/20/2017	Clustering I	TBA
04/25/2017	Clustering II	TBA
04/27/2017	A Priori	TBA
05/02/2017	Study week	
05/04/2017	Study week	
05/09/2017	Finals week	
05/11/2017	Finals week	