36-462 Topics in Statistics: Statistical Learning
Spring 2010

Instructor: Rebecca Nugent
Baker Hall 232C
(412) 268-7830
rnugent@stat.cmu.edu
http://www.stat.cmu.edu/∼rnugent
Office Hours: Wed 2-3pm

Teaching Assistant:
Darren Homrighausen
Craig Street 311B
(412) 268-9922
dhomrigh@stat.cmu.edu
Office Hours: Mon 2:30-3:30pm, Porter A18A

Class Meetings: Tuesdays, Thursdays 1:30-2:50pm, Doherty A310

Website: http://www.cmu.edu/blackboard
http://www.stat.cmu.edu/∼rnugent/teaching/CMU462/

Prerequisites: any of 36-226, 36-310, or 36-626

Recommended Textbooks:

- *Finding Groups in Data: An Introduction to Cluster Analysis*. Kaufman & Rousseeuw. Wiley.
- *Multivariate Analysis*. Mardia, Kent, Bibby. Academic Press.
- *The Elements of Statistical Learning*. Hastie, Tibshirani, Friedman. Springer

General Course Plan: In this special topics course, we will explore different supervised and unsupervised learning techniques (i.e. "learning with and without labels") focusing more heavily on clustering and classification methodology. Emphasis will be on using these tools in practice and in particular for finding reduced group structure in multivariate high-dimensional data. We will also focus on diagnostics and validation of our methods (i.e. how do we know if we succeeded? Are there subgroups that need more analysis? Which observations are problematic?). Applications will include, among others, astrostatistics, text mining, and genetics.

Course Objectives:
For each method, we want to be able to:

1. Know when the method is appropriate to use
2. Understand the underlying theory
3. Effectively choose any needed parameters
4. Implement the method and interpret the results

Course Work: Your grade in this course will be determined by homework assignments, a midterm project, and a final group project.

- Weekly homework assignments will be due at the beginning of class (1:30pm) on Thursdays. Assignments should be submitted electronically to the Blackboard Digital Dropbox and in class (B-W is fine for paper version). There have been some issues with the latest Office version and Blackboard. If you use this version, you may want to avoid the .docx extension and perhaps convert to a pdf. Late homeworks are not accepted (exceptions may be made depending on circumstances; instructor permission required in advance).

  Homework Format: name on front page; questions answered in order; ALL answers should be clearly marked and labeled; *just circling answers on R output or attaching graphs with no explanation is not acceptable; answers should be written up in context of problem.* Graphs should be as close to the corresponding problem as possible. Deviating from this format may result in loss of points on homework.

  Please see the TA or instructor during office hours for help with homework problems. Questions posed by email must be sent at least 24 hours before the time an assignment is due in order to guarantee a response.

- The midterm project will be done individually and will involve comparing and contrasting clustering algorithms on a large data set. The final write-up will include a discussion of the data, the algorithms, their performance, and the assumptions made. More details will follow.

- The final project will be group-based and will require constructing your own algorithm on a training data set with the goal of achieving the highest classification rate possible without overfitting the data set. Your project will be written up in report form detailing and justifying the algorithm. The code will also be turned in and run on a test data set live in class. More details will follow.

Grading policy: You are encouraged to discuss homework problems with your fellow students, however the work you submit must be your own. Acknowledge any help received on your assignments. Copied work will receive no credit. Late assignments will not be accepted. Your lowest homework grade will be dropped. **Please come talk to me if there are difficulties; problems/conflicts must be discussed IN ADVANCE.** Cheating/copying will result in a zero for the homework or project and a letter to your dean. Do your own work. Final grades will be computed with the following weights:

Homeworks	.40
Midterm Project	.25
Final Project	.35

You have one week from the day an assignment, exam, etc is handed back in class to bring any grading issues, comments, complaints, etc to the attention of the instructor. Please note that if you are absent the day something is handed back, you will not receive an extension unless arrangements have been made in advance with the instructor.

Final letter grades will be determined as usual: [90,100] = A, [80,89] = B, [70,79] = C, [60,69] = D, [$<$ 60] = R. Grades may be curved at the instructor's discretion (effort, improvement, etc).

Computing: The statistical computing package we will use in this course is R. R is available on many campus computers, and you may download a free version from `www.r-project.org`. You may also use the nearly-identical (but not free) program called S+, available on all campus computers. You can obtain a free temporary version from `myandrew`. This version is good for 1 year; you can keep renewing the license as long as you are a CMU student.

R References: manuals available on R website;
`http://www.stat.cmu.edu/~rnugent/teaching/introR`
Introductory Statistics with R, Peter Dalgaard; Springer-Verlag
Modern Applied Statistics with S-Plus Venables, Ripley; Springer

Laptop Policy: Students are expected to be participating in class; any laptop use during class should pertain directly to the class. Instructor reserves the right to not allow laptop use during class. When the class has a guest speaker, laptops must be turned off and put away.

Cellphones/Pagers, etc: All cellphones, pagers, beepers, and anything else that makes noise should either be turned off or silenced during class.

Communication: Assignments and class information will be posted on Blackboard and class website. Help with using blackboard is available at www.cmu.edu/blackboard/help/.

Academic Integrity: All students are expected to comply with the CMU policy on academic integrity. This policy is online at www.studentaffairs.cmu.edu/acad_integ/acad_int.html

Disability Services: If you have a disability and need special accomodations in this class, please contact the instructor. You may also want to contact the Disability Resources office at 8-2013.

TENTATIVE SCHEDULE: *subject to change*

Date	**Topic**	**Due**
Tue 1/12	Introduction; Learning Examples	
Thu 1/14	Measuring Distance/Dissimilarity	
Tue 1/19	K-means/medoids	
Thu 1/21	K-means/medoids	HW 1
Tue 1/26	Linkage Clustering	
Thu 1/28	Minimal Spanning Trees	HW 2
Tue 2/2	Density-based Dissimilarity	
Thu 2/4	Model-Based Clustering	HW 3
Tue 2/9	Nonparametric Clustering; High Density Clusters	
Thu 2/11	Cluster Trees; Thresholds	HW 4
Tue 2/16	Spectral Clustering; Image Segmentation	
Thu 2/18	Spectral Clustering; Image Segmentation	HW 5
Tue 2/23	Diagnostics; Measuring Performance	
Thu 2/25	Diagnostics; Measuring Performance	
Tue 3/2	Longitudinal Clustering	
Thu 3/4	Clustering Trajectories	**Midterm Project**
	Mon 3/8 - Fri 3/12: Spring Break	
Tue 3/16	Variable/Feature Selection	
Thu 3/18	Projections; Principal Components; MDS	no HW due
Tue 3/23	Linear Discriminant Analysis; Fisher's Discriminant	
Thu 3/25	Kernels; Support Vector Machines	HW 6
Tue 3/30	Diffusion Mapping; Astrostatistics	
Thu 4/1	Diffusion Mapping; Astrostatistics	HW 7
Tue 4/6	Streaming Classification/Clustering	
Thu 4/8	Streaming Classification/Clustering	HW 8
Tue 4/13	Catch-up; Final Project	
Thu 4/15	*No class; Carnival*	
Tue 4/20	Special Topics	HW 9?
Thu 4/22	Special Topics	
Tue 4/27	Special Topics	**Final Project**
Thu 4/29	**Thunderdome**	