SYLLABUS E10.20: Advanced Topics in Quantitative Methods: Classification and Clustering
Spring 2009, 2 credits

Course Time and Location:
Tuesday 3:30-6:10; optional lab section Tuesdays 6:20-7:30
NOTE: first class meets March 10, 2009
ROOM TBD
Instructor: Marc Scott
Office Hours: Tuesday 2:30-3:30 and by appointment

Course Description and Prerequisites:
Classification and clustering are important statistical techniques commonly applied in many social and behavioral science research problems. Both seek to understand social phenomena through the identification of naturally occurring homogeneous groupings within a population. Classification techniques are used to sort new observations into pre-existing or known groupings, while clustering techniques sort the population under study into groupings based on their observed characteristics. Both help to reveal hidden structure that may be used in further analyses. This course will compare and contrast these techniques, including many of their variations, with an emphasis on applications.

The prerequisite is the first two semesters of intermediate and advanced quantitative methods (E10.2003, E10.2004) or the equivalent as approved by the instructor.

Course Goals:
By the end of this course, the student will be able to:
1) Determine which technique is appropriate for different problems, based on the goals of the research.
2) Explore multivariate data using visualization techniques (including multi-dimensional scaling) to illuminate the structure in the data.
3) For classification problems, understand the advantages and disadvantages of model-based approaches over their alternatives.
4) For clustering problems, be able to choose a suitable metric and technique, depending on the goals of the research, as well as be aware of the strengths and limitations of the various approaches.

Course Timing:
The course is organized as a half-semester module, and will meet for the first time on March 10. Due to spring break and the April AERA meetings, class will not meet on March 17 and April 14. The compressed nature of this offering requires student attendance at every class meeting. A lab section follows the lecture, emphasizing practical applications of these techniques using statistical software.

Grading:
Grading will be based on class participation (10%), three homeworks (20%), and two projects (35% each) that will involve both data analysis and a thoughtful description of both the analysis and the findings. Depending on the size of the class, some assignments may be done in groups.

**Reading materials**
There is no required textbook for the course. The required readings are either available through e-journals through the library or will be posted on Blackboard. The following texts are recommended, however:


**Outline of course topics and readings:**
The following outline describes the topics that will be covered along with anticipated associated readings. It corresponds roughly to the course weeks though we may end up adjusting time spent on each topic as we go. All readings not available on the web will be put on electronic reserve through Bobst library.

1) **Introduction to classification and clustering:** what is a cluster; visualization techniques, including principal components and multidimensional scaling.


2) **Hierarchical clustering (including linkage methods) divisive vs. agglomerative approaches; distance measures; the dendogram.**


3) **Optimization techniques; use of dissimilarity matrices and the partition about medoids approach; choosing the number of groups; evaluating clusters**


4) **Finite Mixture models (model-based clustering); growth curves (incl. Nagin clusters)**


5) Classical classification techniques: Discriminant Function analysis; adding complexity; interpretation


6) Model-based classification: logistic and other models
