Psychology 993
Methods for Clustering and Classification, Spring 2006
Syllabus

Contact Information
Jonathan Templin
449 Fraser Hall
jtemplin@ku.edu
785-864-4261

Course Information
Tuesdays and Thursdays, 2:30 pm – 3:45 pm
- 547 Fraser Hall

Office Hours
Thursdays, 11am – 1pm, 4pm – 5pm,
and by appointment.

Online Communications
MSN Messenger: jtemplin@ku.edu
Blackboard Enabled: http://courseware.ku.edu

Course Objectives

This course provides a survey of methods for clustering and classification. Beginning with well-established techniques, we discuss methods such as discriminant analysis, hierarchical clustering methods, and K-means clustering techniques. The second half of class is devoted to model-based approaches stemming from latent class and finite mixture models.

This course will be taught at multiple levels, mixing applications with rigorous mathematical and statistical topics (such as model estimation). The student is expected to work to understand all levels of difficulty and topics covered within the course, regardless of their perceived relevance to the student's specific research interests.

Textbook
None directly assigned, readings will be from photocopies and downloadable documents.

Course Website/Technology
This course will feature extensive use of Blackboard. All necessary information will be posted on Blackboard, which is accessible to students who have registered for the course at http://courseware.ku.edu.

I plan to make use of the message board features of Blackboard to allow for additional comments and questions. Finally, if you use MSN messenger, you may instant message me your questions if you find me online. If you have a webcam, you are welcome to have video conferences with me, too.

Statistical Computing

For this course, we will primarily use two computing packages: SAS and R. Unlike many point-and-click packages (such as SPSS), SAS is a code based language that does many statistical procedures. Of course, if you are already familiar with another package, you may use that package. Be advised that all examples and solutions will feature SAS, and I cannot provide you any help with any other package. Note I **strongly** recommend you complete the course assignments in SAS due to the specific nature of information requested by homework assignments.

SAS is available to you in two ways:

1. You can purchase SAS through campus at http://www.ku.edu/acs/stats/StatisticalSoftware.shtml.
2. The Budig and Fraser computer labs have SAS installed on all computers; for a list of labs and hours visit http://www.computerlabs.ku.edu/, and select "Labs on Campus" from the top left.

Although SAS can be used for many of the classical methods for clustering and classification, not all methods can be estimated within SAS. For the remaining topics, we will use the R statistical computing package. R is a freely available statistical computing package that is extremely malleable to a user's needs. We will spend the first two weeks of class discussing SAS and R, which should allow for easy transitions when discussing topics within this course.

To obtain a copy of R for free, visit http://www.r-project.org.

Course Grading System

The final grade will be determined based on the weighted average of the homework assignments, and the two tests using the following weights:

Homework	40%
Project	40%
Presentation	20%

Course grades will be determined by the weighted average of the homework, presentation, and project grades, and will be given according to the following scale (pluses and minuses will be given for the differing thirds of a grade category). I reserve the right to round grades upward in the event I misjudge the difficulty of the course, but grades will never be rounded downward.

My goal is for everyone to succeed in this course, learn the material, and receive an A. Here is a rough picture of the grading scale I will use:

A	B	C	D	F
85% - 100%	70%-84%	60%-70%	50-60%	Below 50%

Course Structure

Homework

This course will feature several homework assignments to allow for sufficient practice of the principles discussed. A typical assignment will consist of a set of practice problems, expected to be completed within two weeks. All assignments are posted online, and the homework is expected to be completed by the beginning of class on the due date. All late homework will have a 10% penalty per day late.

Students are allowed and ***encouraged*** to collaborate with each other on homework assignments, however, each student must turn in an *original* piece of work. Homework is your time to learn the material with heavy guidance from the instructor. A random set of problems will be graded on a 3-point scale, with deductions for non-needed computer output (basically, submit only what was used in answering each question).

The points received divided by the total points possible will be your homework percentage for any given week. Each week's homework will count the same weight toward your final grade (i.e. weighted averages will be used).

Homework Grading Scale

Points	Expectation
3	Answer is correct and concise. You demonstrate you know the content area
2	Effort is made, but answer has some errors. You show you are on the right track and trying.
1	Minimal effort for the problem. Do not demonstrate you know the concepts. OR -- Way too much computer output (if applicable)
0	No answer given

Course Project

To successfully complete this course, you must submit a course project. The course project is an important part of this course, accounting for 40% of the total grade. More importantly, the course project is where you will gain practical experience in applying the measurement method taught in this course. As an ultimate goal for the project, consider choosing a topic that will lead to a publishable paper rather than a topic that will just allow you to complete the course.

The requirement for the course project is an empirical research project, accompanied by an APA-style research paper describing it. For the research project, you may work in a small group with other students (at most, three students per group), but each student must turn in an individually written paper. The project is described in further detail on another handout.

Presentation

A portion of the course grade comes from the presentation of an empirical research article on a topic covered in the course. Each student is expected to prepare one 20-30 minute presentation that will coincide with the topic covered during the week.

Course Style and Content

Lecture Format

All lecture notes will be available digitally, with notes available online the morning prior to the lecture. If you have a data set you would like to see included in the in-class examples, I encourage you to submit it to me.

Reading Assignments

To be fully successful in this course, I **strongly** encourage you to read the assigned paper(s)/chapter(s) prior to the course when we will cover the topic. I believe the book is written in a very readable style so that people with only a minimal statistical background can understand (undergraduate statistics – Psych 300), and should become an excellent reference for you in the future.

Furthermore, there are many portions of the book the author mentions as optional. These sections are not optional for this course, we will cover everything.

Topic and Reading List (subject to change as necessary):

Statistical Computing (first two weeks)

- Introduction to SAS (1/24)
  - My handouts.

- Introduction to R (1/26, 1/31, 2/2)
  - Venables, W. N., & Smith, D. M. (2005). *An introduction to R.* Bristol: Network Theory Ltd. ISBN: 0954161742.

Introduction to Clustering/Classification (one class)

- Chapter 1 and 2 (p. 1-34) from Gordon, A. D. (1999). *Classification* (2nd ed.). Boca Raton: Chapman and Hall/CRC. ISBN: 1584880139.

- Chapter 1 (p. 1-37) from Massart, D. L., & Kaufman, L (1983). *The interpretation of analytical chemical data by the use of cluster analysis.* New York: Wiley. ISBN: 0471078611

Cluster Validation (one class)

- Chapter 6 (p. 183-212) from Gordon, A. D. (1999). *Classification* (2nd ed.). Boca Raton: Chapman and Hall/CRC. ISBN: 1584880139.

Discriminant analysis (one week)

- Chapter 11 (p. 581-687) from Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Upper Saddle River: Prentice Hall. ISBN: 0130925535

- Klecka, W. R. (1980). *Discriminant analysis.* Newbury Park: Sage. ISBN: 0803914911.

- Anderson, J. D. (2005). Financial problems and divorce: Do demographic characteristics strengthen the relationship? *Journal of Divorce and Remarriage*, 43, p. 149-161.

Hierarchical clustering methods (two weeks)

- Chapter 12 (p. 668-700) from Johnson, R. A., & Wichern, D. W. (2002). *Applied multivariate statistical analysis*. Upper Saddle River: Prentice Hall. ISBN: 0130925535

- Bailey, K. D. (1994). *Typologies and taxonomies: An introduction to classification techniques.* Newbury Park: Sage. ISBN: 0803952597

- Chapter 3 (p. 75-100) from Massart, D. L., & Kaufman, L (1983). *The interpretation of analytical chemical data by the use of cluster analysis.* New York: Wiley. ISBN: 0471078611

- Gonzalez-Ibanez, A, Aymami, M. N., Jimenez, S., Domenech, J. M., Granero, R., Lourido-Ferreira, M. R. (2004). Assessment of pathological gamblers who use slot machines. *Psychological Reports*, 93, p. 707-716.

K-means clustering methods

- Hartigan, J. A. (1975). *Clustering algorithms*. New York: Wiley. ISBN: 047135645X

- Steinley, D. (2003). Local optima in K-means clustering: what you don't know may hurt you. *Psychological Methods, 8*, 294-304.

- Napoli, J., & Ewing, M. T. (2001). The net generation: An analysis of lifestyles, attitudes, and media habits. *Journal of International Consumer Marketing*, 13, p. 21-34.

Additive Trees

- Corter, J. E. (1996). *Tree models of similarity and association*. Newbury Park: Sage. ISBN: 0803957076.

Taxometrics

- Waller, N. G., & Meehl, P. E. (1997). *Multivariate taxometric procedures : distinguishing types from continua*. Newbury Park: Sage. ISBN: 0761902570.

Latent Class Analysis

- McCutcheon, A. L. (1987). *Latent class analysis.* Newbury Park: Sage. ISBN: 0803927525.

- Dayton, C. M. (1999). *Latent class scaling analysis*. Newbury Park: Sage. ISBN: 0761913238.

- Keller, F., & Kemph, W. (1997). Some latent trait and latent class analyses of the Beck Depression Inventory (BDI). In J. Rost and R. Langeheine (eds.) Applications of Latent Trait and Latent Class Models in the Social Sciences.

Latent Profile Analysis

- Chapter 8 (p. 226-244) from Lazarsfeld, P. F., & Henry, N. W. (1968). Latent structure analysis. New York: Houghton Mifflin.

- O'connor, R. M., & Colder, C. R. (2005). Predicting alcohol patterns in first-year college students through motivational systems and reasons for drinking. *Psychology of Addictive Behaviors*, 19, p. 10-20.

Finite Mixture Models

- Chapter 1 (p. 1-39) and Chapter 3 (p. 81-116) of McLachlan, G. J., & Peel, D. (2000). *Finite mixture models*. New York: Wiley. ISBN: 0471006262.

- Slaney, M., & McRoberts, G. (2003). BabyEars: A recognitions system for affective vocalizations. *Speech Communication,* 39, p. 367-384.

Growth Mixture Models

- Li, F., Duncan, T. E., Duncan, S. C., & Acock, A. (2001). Latent growth modeling of longitudinal data: A finite growth mixture modeling approach. *Structural Equation Modeling*, 8, p. 493-530.

- Wang, C.-P., Hendricks Brown, C., & Bandeen-Roche, K. (2005). Residual diagnostics for growth mixture models: Examining the impact of a preventative intervention on multiple trajectories of aggressive behavior. *Journal of the American Statistical Association,* 100, p. 1054-1076.

- Li, F., Barrera, M., Hops, H, & Fisher, K. J. (2002). The longitudinal influence of peers on the development of alcohol use in late adolescence: A growth mixture analysis. *Journal of Behavioral Medicine*, 25, 293-315.

Models for Cognitive Diagnosis/Skills Assessment

- Templin, J. (under review). Cognitive diagnosis: Concepts and common models.

- de la Torre, J., & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, 69, p. 333-353.

- Templin, J., & Henson, R. (under review). Measurement of psychological disorders using cognitive diagnosis models.

Tentative Course Schedule

Date	Topic
1/24	Course Overview
1/26	Introduction to SAS (Fraser Hall computer lab)
1/31, 2/2	Introduction to R (here in class, augmented by your home computers).
2/2	Introduction to Clustering/Classification
2/7	Cluster Validation
2/9, 2/14	Discriminant Analysis
2/16	**No Class (I am at a conference on measurement)**
2/21, 2/23	Hierarchical Clustering Methods
2/28, 3/2	K-means Clustering Methods
3/7	Additive Trees
3/9	Taxometrics
3/14, 3/16	Latent Class Analysis
	Initial Project Proposal Due Tuesday, 3/7
3/21, 3/23	**No Class (Spring Break)**
3/28, 3/30	Latent Profile Analysis
4/4, 4/13	Model Estimation (EM and MCMC algorithms)
4/6, 4/11	**No Class (I am at NCME)**
4/18, 4/20	Introduction to Finite Mixture Models
	Final Project Proposal Due Tuesday, 4/18
4/25	Finite Mixture Models
4/27	Combining Methods: Growth Mixture Models
5/2, 5/4, 5/9, 5/11	Models for Cognitive Diagnosis