

22QA727 Data Mining

Instructor: Yan YU
Office: 527 Lindner Hall
Office Hours: T TH 12:30pm-1:30pm
And by appointment
Class Time: T TH 4:30-5:45pm, Lindner 216, Spring 2009
Email: Yan.Yu@uc.edu
Phone: 556-7147
Web Page: <http://www.blackboard.uc.edu>

Course materials including syllabus, lecture notes, reading assignments, case homework, data sets, SAS and R programs, and course handouts will be posted on the course web in blackboard.

Course Objectives: To provide students with a hands-on data analysis experience using various statistical methods and major statistical software (SAS and R) to analyze complex real world data in business and industry.

Course Format: Classes will be provided in three forms: lecture, case study, and project discussion/presentation. In case study, students will be led through practical problems addressed by data analysis techniques. The aim is to provide a detailed view on how to manage complex real world data; how to convert real problems into models so that statistical software can be used appropriately; and how to interpret and diagnose the model fitting. Project discussion and presentation will enable students to share and compare ideas with one another and to receive specific guidance from the instructor.

Recommended Text:

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

This is an advanced book to read, but it is the best statistical data mining text available. The course will cover the less theoretical material in the text. Lectures should be easier to follow than the text. Supplemental readings and notes will be posted.

Berry, M. and Linoff (2004), *Data Mining Techniques*, 2nd edition, John Wiley & Sons.

This book is recommended as a supplemental reading from practitioners' view. It is very applied without technical details.

Prerequisites: QA722 or equivalent; Basic Statistical Computing skills.

Grading:

Exam (In-Class, Close Book)	40%
Cases	30%
Project	30%

Exam and Honor Policy

Exam is **in-class, close-book and close-notes**. You can bring one page (standard 8 1/2 x 11 paper) **hand-written** cheat sheet, which will be turned in along the exam. Exam should be the sole work of each student. Anyone cheating or assisting another during an exam will be given a 0 for that exam and possibly a grade of F for the class. College procedures will be followed and the graduate dean will be notified.

Computing Resources:

We will use SAS and R. You can access SAS and R in the second floor computer lab (215 and 202). Other computer labs on campus, e.g., Engineering, Education, etc. also have SAS installed. SAS help files are available on *SAS ONLINE* that you may access at the URL <http://www.uc.edu/sashtml>. You can download R for free through the URL <http://www.r-project.org/>.

Group Work Structure of the Course: After the first class, each student will join a work group. A work group will consist of four students. This work group will be maintained for the length of the quarter. The work group will cooperate in all work given during the quarter including case homework and final project. All members of a group will share grades on any submitted work. All members are to contribute equitably to the shared workload, carrying a fair weight for the burden. At the end of the quarter, members of each group will be asked to evaluate the contribution of the other work group peers on the basis of a number of criteria such as intellectual contribution, attendance at group meetings, mentoring and sharing knowledge, writing up the results, presentation, and running relevant SAS and R codes. The peer score will reflect, in some sense, an average over all the assigned work as well as an average of the above criteria. Thus, a student in a work group who may have contributed much on one assignment, may not have contributed the majority of the work on another, yet still such work may be considered by other members to be meritorious “on the average”.

Project and the Fourth Credit Hour: Note that QA 727 is a **4 credit hour** course. It is my intention to add material to this course that will justify the additional hour of credit.

Purpose of project: One goal of the project is to provide you with more experience using data mining tools on practical problems. A second goal is to help you become a self-directed learner; this is the type of learning that you will be doing in the future. It would be most interesting if you have some new methodology learned from this course for some interesting business problems. For some MSQA students, this project could serve as serious starting for your MS project.

Project Teams: You should work in a team of four students. You should form a team yourself. If you cannot find teammates, then let me know and I will help you find a team. If you insist, you may work alone.

Types of projects: Almost any type of project is acceptable. However, I expect that most projects will be either one of or a combination of the following types:

- Applying tools that you have already learned in this course to a data set not used before (or at most, a sample of a LARGE data that we have discussed).
- A report describing one of the data mining tools that are discussed in the textbook but either are not covered in the lectures or will be covered only briefly at the end of the course. Data mining tools that could be studied in your report would be any *except* linear and logistic regression, decision trees, and neural networks. For example, data mining tools that *could* be studied include
 - splines, MARS, generalized additive models
 - MART
 - kernel estimation and classification
 - basis expansion and regularization
 - bagging
 - boosting
 - linear discriminant analysis
 - hierarchical mixtures of experts
 - support vector machines

For a project of this type, in addition to the textbook you should use several other books or articles as source materials. It should involve some programming of the data mining tool to study and testing it on, say, one of the course data sets;

- A Monte Carlo simulation of a data mining tool, where large data sets of known structure are simulated and data mining tools are tested to see how well they can detect the known structure.

What is required? Each team must write a project proposal, find the necessary data, carry out the project, and write a project report.

The report should be at most 20 double spaced pages with one inch margins and 12 point font and should contain:

- Title page with authors and abstract
- Introduction telling what the project is about, what your team has accomplished, and a brief statement of results and conclusions.
- One or more sections describing the project
- Conclusions
- Bibliography

Tables and figures can be interspersed in the text or at the end of the report. All tables and figures should be numbered and referred to by number. The report should not contain raw computer output. Rather, any computer output should be in a table or figure. If necessary, append brief SAS and R codes in the appendix section.

Project Grading: The project is worth 30% of the course grade. Grades will be based on:

- Interesting Application, Creativity.
- How much new materials you have learned.
- Clarity and conciseness of the report.
- Correctness.
- Powerpoint Presentation.
- Intragroup Evaluation.

Academic Integrity: The project should be the sole work of the students on the team. None of the work on the project should have been used as part of any other course, independent study project, master's project, or other type of project for academic credit. *Exception:* Slight overlap between this project and another project for academic credit might be permissible if discussed in advance with the instructor.

Project Proposal: Ideally a proposal can serve part of the first **5-10** pages of your final project report, depending on how much preliminary analysis you have conducted. Proposal should usually include at least 1) Background information of the project. The objective you want to achieve. 2) A detailed data section: discussion of data source and nature of the variables involved in the analysis. 3) Preliminary analysis. That is, some exploratory analysis of the data set, summary, plots, and maybe some kind of linear regression fit to check the feasibility of the problem as well as get a better idea of how this data looks. 4) Proposed work from now till the final project to be turned on. E.g., GLM, GAM, stepwise, CART, NNET etc. Model comparison, and cross validation. 5) List detailed references if it's suitable. 6) List possible simulation study design if it applies. Of course, the length as well as the content should largely depend on the problem each individual group is facing. A sample proposal is posted on the course web.

Due Dates

Project proposal: **05/05/09**, at the beginning of class.

Find a topic, necessary data, summarize what you plan to do in the project.

Project Final report: 5pm **06/10/09**, maximum of twenty typed, double spaced pages with one inch margins and 12 point font.

Tentative Schedule

Date	Lecture
Week 1 03/31 04/02	Overview/ Data Mining; Cases
Week 2 04/07 04/09	Statistical Software (R); Review of Linear Regression, Variable Selection; Case Study of Linear Models
Week 3 04/14 04/16	Generalized Linear Models (e.g. logistic regression)
Week 4 04/21 04/23	Generalized Additive Models (GAM) Case 1 due, presentation
Week 5 04/28 04/30	Classification and Regression Trees (CART)
Week 6 05/05 05/07	Proposal due, presentation
Week 7 05/12 05/14	Case Study of CART, Applications in direct marketing Case 2 due, presentation
Week 8 05/19 05/21	Neural Network Exam
Week 9 05/26 05/28	Case Study of Neural Network, Support Vector Machine (SVM) Summary, Project Presentation
Week 10 06/02 06/04	Project Discussion, Presentation Final Project Report and Evaluation due at 5pm 06/10