36-724: Applied Bayesian and Computational Methods
Spring 2006

MWF, 11:30--12:20, Porter 225B
http://www.stat.cmu.edu/~brian/724

Course Information

Instructor:
Brian Junker, Statistics
132E Baker Hall
268-8874
brian@stat.cmu.edu

Grader:
Daniel Ting
Office TBA
Phone TBA
dting@stat.cmu.edu

Office Hours:
Immediately after class
(or by appointment).

Office Hours:
TBA
(or by appointment).

Prerequisites

The background you need for this course would be found in the following Statistics Department courses: 36-705 Intermediate Statistics, 36-711 Statistical Computing, 36-707 Applied Regression. If you have not had these three courses or close equivalents please let me help you decide if it makes sense to stay in the course.

Texts


There is much more material in these texts that we can cover in the course. So we will pick and choose material from this text as we need it in the first half of this course. I will also be taking material from the following texts (especially Gentle and Venables & Ripley), and from other sources:

Course Description and Course Objectives

This is a one-semester course in computationally intensive methods for data analysis and inference, featuring applied Bayesian methods. It extends a former mini-semester course, Applied Bayesian Statistics, to include some modern computational methods for learning from multivariate data. We will spend the first half of the semester on Bayesian data analysis, ending approximately at Spring Break. After Spring Break we will shift our focus to computationally intensive methods for modeling and exploring structure in data.

This course emphasizes a traditional applied statistics perspective: We will examine methods that can be customized with rich background knowledge about a data set to develop models or reveal structure that allow us to make specific inferences or predictions about the phenomena generating the data.

- **Bayesian Data Analysis**

  In this part of the course my goal is to teach you a basic approach to setting up, fitting and evaluating hierarchical Bayesian models for practical, applied data analysis. The featured computational method is Markov Chain Monte Carlo (MCMC) but we will also touch on other methods.

  Here are the topics we will consider, with the chapters from Gelman from which we will extract the information we need:

  - Basic ideas; “simple” Bayesian models; conjugate priors; “simple” simulation [Ch’s 1-3, handouts].
  - More computational methods (Newton-Raphson, E-M, MCMC) [Parts of Ch 4 and Part III, handouts]
  - Hierarchical Models [Ch’s 5 and 9]
  - Model Checking/Comparison/Selection [Ch 6, handouts]
  - Examples [Taken from Parts IV and V; or other sources such as Congdon (2001, 2003)].

- **Computational Data Analysis**

  In this part of the course my goal is to survey a variety of approaches to discovering structure in, or summarizing, large data sets. In contrast to the computationally intensive statistical inference illustrated in the Bayesian part of the course, these methods are closer to computational inference. Some topics we might consider, and the places to read about them, include:

  - Model Inference [Hastie et al. Ch’s 7–8, & some ideas from Ch’s 2–4]
  - Classification: Discriminant Analysis, Bayes/plug-in rules, Support Vector Machines, k-NN [Hastie, Ch’s 4 & 12; Gentle Ch’s 5 & 10, V&R Ch 12]
  - Tree-Based Methods. [Hastie, Ch’s 9–10; Gentle Ch 11, V&R Ch 9]
  - Clustering: Computational and Model-based methods [Hastie, Ch’s 13-14; Gentle Ch’s 5 & 10, V&R Ch 11]
  - Visualization: Principal Components, Factor Analysis, Projection Pursuit, Multidimensional Scaling [Hastie Ch 14; Gentle Ch’s 5 & 10, Venables and Ripley (V&R) Ch 11]

Let me know if there are other topics that might also be of interest.

We will cover a lot of ground in this course, and so we cannot do anything in detail. The idea is to give you a taste of each topic, so that when you need to do these kinds of analysis in the future, you will have some idea what details are important, and where to look to find out about them.
Computing

An essential component of this course is computing. One might think that this applies only to the “Applied Computational Statistics” part of the course, but in fact modern applied Bayesian statistics depends on intensively computational methods as well, such as iterative simulation (Monte Carlo) and optimization algorithms.

The point of departure for computing work in this course will be the general-purpose statistics packages Splus or R, that you became familiar with in 36-711. Some homework should be written up in \LaTeX. In addition to this core, we will also explore other software for Bayesian estimation and analysis, such as BUGS (LINUX) and/or WinBUGS (Windows); as well as visualization software such as Xgobi and vtk. Although we will not delve into them much, other more general-purpose packages—like Matlab, Maple, Minitab, SAS, Xfig, GIMP and Gnu-plot—as well as lower-level programming in C/C++, Perl, etc., also have important roles to play in computationally intensive applied statistics.

Most of these tools are available in some form on the LINUX workstations in the Statistics Department; in-class demo’s will often use the same software in Windows. If you have your own PC (Windows, Mac or LINUX), you will find it convenient to install some of these tools on it. If you are coming from outside the department, please look for similar tools on Andrew LINUX or Windows workstations.

Homework, Tests and Grading

I will give assignments approximately every week or two. Homework will be a mix of developing and exploring methodology, and practicing using software to analyze data. You may work with other students on homework problems or refer to other sources if you would like. The computations and writeup of your assignment, however, must be your own. Please annotate your work with brief, clear sentences explaining your approach and interpreting your results (you are not expected to write full-blown data analysis reports however).

There will be a midterm project and a final project, instead of exams. For these two projects I ask that you discuss your work only with me and/or the TA. Once again, briefly explaining your work and interpreting your results will be an important part of your writeups.

Homeworks together count for 70% of your grade. The two projects count 15% each. The homework will be graded “supportively”: you get credit (and feedback where appropriate) for giving each part a strong, coherent try. The projects will be graded more “evaluatively”.

Unless specifically requested, never submit raw computer output pages. It is best to use \LaTeX, incorporating computer output fragments with verbatim and figures with \texttt{epsfig}; or if you prefer to do the same thing with another word processor that is fine too. As a last resort, cut out the appropriate parts of the output and neatly tape it onto your homework paper. Please label all output, plots, variables, etc., appropriately.

Data Sets, Web Page, Communications

- As far as possible, all materials for this course will be made available on the course web page, \url{http://www.stat.cmu.edu/~brian/724}.
- The surest way to contact us anytime outside of class is via email:
  - Instructor: brian@stat.cmu.edu
  - TA: dting@stat.cmu.edu

Also please feel free to drop by our offices or schedule special appointments with either of us.
Schedule

This course is scheduled to meet Mondays, Wednesdays and Fridays at 11:30. However, the following days are “special”. I will schedule makeup lectures or guest lecturers for some of my out of town dates below.

Monday January 16 is Martin Luther King day. There are no classes after 12:30. However we will meet as usual at 11:30.

Monday January 30 I cannot be in class this day. I will probably schedule a makeup for Tues Jan 31 or Thur Feb 2.

Friday February 10 I will be out of town. Class is cancelled but we will schedule a makeup class, most likely Tue Feb 7 or Wed Feb 8.

Friday March 10 is Mid-semester Break. There are no classes that day.

March 13–17 is Spring Break; there are no classes.

Friday March 31 I will be out of town. I may schedule a makeup lecture earlier in the week.

Fri Apr 7 and MWF Apr 10–14 I may be out of town for several of these lectures. As soon as I know I will announce cancellations, makeups and/or guest speakers.

Friday April 21 is Spring Carnival. There are no classes that day.