

CHM 696D Analytical Informatics

Spring 2003

Randall K. Julian, Brown 4185, rkj@lilly.com (317) 276-5868

Monday and Wednesday, 4:30 – 5:45 with office hours by appointment.

CHM 696D covers the theory and application of informatics methods in analytical and bioanalytical chemistry.

The amount and complexity of data generated by instrumental methods has grown tremendously over the past decade. High throughput methods in analytical chemistry and biology seem to be the key to solving many difficult problems. These same methods, however, have created new problems for the analytical chemist: “What do you do with all this data”, and “How can you make any sense of it all”. This course intends to provide a guide through the myriad of approaches to answering these questions.

There have been many suggestions on how to draw inferences from large volumes of complex data. Computer science calls this work artificial intelligence, informatics and data mining. In engineering, the area is known as signal processing and pattern recognition. In applied statistics, work in machine learning, clustering and classification have all had impact. Historically, when these methods are used in chemistry, they are lumped under the name “chemometrics”.

In this course, we will sort through many of the tools and techniques for dealing with data and develop ways to address real-world questions in analytical chemistry. Students will learn the strengths and weaknesses of various methods, and learn to judge methods by quantitative measures. The course will introduce chemistry students to the work of specialists in statistics, engineering and computer science in a clear and direct way. Another goal of the course is practicality – student should be able to immediately apply the ideas to their own research. We will cover supervised machine learning and pattern recognition starting with multivariate regression and discrimination and ending with neural networks and support-vector machines. Along the way, signal processing, feature extraction and reduction methods including wavelets and splines will be explained.

Throughout the course, we will discuss ways to implement these ideas for real-world problems. We will discuss relevant technology including computer languages, relational databases, and methods for representing chemical structures. Computational hardware including LINUX clusters will also be discussed.

The text for the course is: "The Elements of Statistical Learning: Data Mining, Inference and Prediction", by Hastie, Tibshirani and Friedman, (2001). Springer-Verlag. (ISBN 0-387-95284-5). Also recommended is "Pattern Classification" (2nd ed, 2000) Duda, Hart and Stork, Wiley, (ISBN 0-471-05669-3).

The course will make extensive use of “R” (a GNU version of the S statistical software environment), as well as the Perl programming language and the MySQL database server. Students will receive instructions on these free packages. There are many other math/stats packages and programming environments, which might be suitable for use in this course.

Grading in the course will be based on three exams (125 pts each). The first two will be 7-day take home exams, the third will be a 10-day take home.

Exam 1: Out Monday February 10, due Monday February 17.

Exam 2: Out Monday March 24, due Monday March 31.

Exam 3: Out Monday April 21, due Wednesday April 30.

The course is mostly self-contained, however students are expected to solve problems using computer-based tools.