

## Analytical Informatics: Chemistry 696D

### Part I. A Statistical Foundation for Informatics

Goal: Understanding of tools of informatics, major applications, limitations of machine analysis. Grasp of underlying statistical concepts. Develop ability to use high level tools like R.

*All the math you need and the justification for the work to learn it...*

#### Lecture 1 – What is analytical informatics

- Information theoretic view of analytical chemistry
- Need for better tools to deal with the data glut
- Big ideas in machine learning
- Applications of various types of machine learning
- Real problems in analytical chemistry addressed by machine learning
- The problem with patterns and naïve use of tools
- Principle of parsimony with a bioinformatic example

#### Lecture 2 – Populations, features and classes

- Review of probabilities in analytical chemistry
  - Mixtures of molecules
  - Observable ‘features’ from molecules
  - Example using Lipinski’s ‘Rule of 5’
- Introduction to Statistical Decision Theory
- The Bayes Decision Rule
- Error and cost of error
- Quantitative idea of “Concept”: classes
- Examples of populations and classes in analytical chemistry

#### Lecture 3 – Pattern Recognition: “Class” Description and Distinction

- Features in multiple dimensions
- Review of descriptive statistics
- Statistics in multiple dimensions
- Decision boundaries in multiple dimensions

#### Lecture 4 – Using R for Data Analysis:

- Practical statistical tools: The R environment
- The ideas behind R
- Basic operation of the R GUI
- Matrix operations in R
- Random variables and Linear Transformations in R
- Generating graphs and visualizing results
- Generating data, plotting data, reading files
- Useful transformations for creating simulated data

Lecture 5 – Regression of a Response Matrix  
Linear Models and Least Squares  
Simple classifier based on regression  
Contour plotting a decision surface  
Linear regression of a polynomial equation  
Overfitting: an introduction to bias and variance

Lecture 6 – Local methods and simple comparisons  
Nearest neighbor methods and Parzen windows  
knn functions in R  
Bias and variance measurements: classification error

Lecture 7 – Multiple classes and mixtures (Hand out Exam 1)  
Multiple classes, multiple boundaries  
Computational methods with knn and multiple classes  
Mixtures  
Types of mixtures  
Bias-variance issues with mixtures

Lecture 8 – Chemical Application: Guest Lecture on Raman Classification (D. Zhang)  
Data preparation  
Feature extraction  
Classification methods

Part II: Applied Machine Learning and Pattern Recognition

Goal: Obtain a firm understanding of the principles, features, benefits and risks of automated data analysis via machine learning. Gain experience performing common analyses using R. Learn how to tell when something is right.

*A survey of methods that can be applied to analytical chemistry data*

Lecture 9 – Linear discriminant analysis  
Canonical variates  
LDA as an alternative to Least Squares  
Multiple dimensions, dimensional reduction with LDA

Lecture 10 - Logistic Regression and Separating hyperplanes  
Logistic models  
Using GLM objects to perform logistic classification  
Comparing Logistic Regression and LDA  
Separating hyperplanes & perceptrons

Lecture 11 – Feature extraction, data preparation and dimensions

- Local methods in high dimensions
- Curse of dimensionality
- The dangers of overfitting
- Dimension reduction through feature extraction
- Simple preprocessing steps

Lecture 12 – Feature extraction in instrumental analytical chemistry

- Chromatography
- Mass Spectrometry
- NMR
- Absorption spectroscopy (UV-Vis)
- IR and similar spectra
- Important practical method: “Finding Peaks”

Lecture 13 – Neural Networks

- Limitations of perceptrons
- Multi-layer neural networks
- Fitting
- Starting, stopping, early stopping
- General problems with neural networks

Lecture 14 – Support Vector Machines

- Support vectors
- Kernels which enlarging the feature space
- SVMs and the curse of dimensionality

Lecture 15 – Recursive Partitioning

- Guest Lecturer: Richard Higgs

Lecture 16 – Ensemble methods

- Guest Lecturer: Richard Higgs

Lecture 17 – Unsupervised learning

- Clustering overview
- K-means
- Hierarchical

### Part III: Technology for Chemical Data Mining and Predictive Modeling

Goal: Learn how to apply machine learning to problems in chemical analysis. Learn how to use databases, scripting languages and specialty analysis tools to build predictive models in analytical chemistry. Gain experience working with databases, writing integrative programs, selecting and testing machine learning methods.

#### *Technology for implementing analytical informatics*

##### Lecture 18 – Technology: Computer Systems for Informatics

Elements: Platforms, Languages and Databases

Computer platforms (Windows and Linux)

Languages for informatics (C/C++, Java, Perl, VB)

Data management: SQL databases (MySQL, Oracle, MS Access)

Web servers & dynamic content

##### Lecture 19 – Programming in Perl

Introduction

Crash course in Perl

Getting Perl, CPAN, Documentation

Control structures

Variables and memory management

Arrays

Regular Expressions

Basic I/O (Reading and writing data)

##### Lecture 20 – More Programming in Perl

More regular expressions

Hash tables

References

Complex data structures

Subroutines and modules

##### Lecture 21 – Chemical data representation

Connection representation

String-based: SMILES

Working with SMILES

Dealing with SMILES using Perl

##### Lecture 22 – XML in Analytical Informatics

Overview of XML

DTD's and the XML-schema

The Chemical Markup Language (CML)

SpectroML, GAML and analytical data formats

Application specific ML's: ProteinLynx and PEDRoML

Lecture 23 – Introduction to Relational Databases

- Relational database concepts
- Tables and fields
- Database design concepts
- Describing databases
- SQL
- MySQL

Lecture 24 – Database programming

- MySQL Permission Tables
- Accessing databases using Perl
- Using Perl to perform SQL queries
- Chemical Database Examples

Lecture 25 – Integrated Systems

- Using Perl to run external programs
- Using Babel to convert molecule file formats
- Using Perl to condition data for an external program (like Babel or R)
- Example MySQL->Babel->MySQL (Perl driven)

Lecture 26 – Analytical and Bioinformatics and an High performance computing & networks

- Bioinformatics in the Machine Learning context
- Proteomics data and data analysis
- High performance compute platforms
- Clusters
- Rocks Cluster Distribution
- Methods of programming parallel systems

Lecture 27 – Case Study Guest Lecture: (E. Stauffer)

- Case study of a high throughput system integrating many elements of analytical informatics.