

Course Description

Approaches to finding the unexpected in data: data mining, pattern recognition and understanding. Emphasis is on data-centered, non-inferential statistics, for large or high-dimensional data, and topical problems. Simple graphical methods, as well as classical and computer-intensive methods applied in an exploratory manner.

Course Objectives

This course is designed to provide students with the experience of working with statistical methods for analyzing new, complex data, and arriving at reliable and detailed summaries of the data. Students will learn to identify a scientific problem underlying selected data examples and address this problem in the analysis and conclusion. The internet will be used to scavenge for software to conduct analyses and literature related to the scientific problem.

Course Outline

Week 1 (Jan 10-14)	What is data mining? Introduction to data analysis software R.
Weeks 2-3 (Jan 19-28)	Case study 1: Tipping behavior
Weeks 4-5-6 (Feb 2-25)	Case study 2: Olive oils
Weeks 7-8-9 (Mar 2-28)	Case study 3: Hurricanes
Weeks 10-11-12 (Apr 1-18)	Case study 4
Weeks 13-14-15 (Apr 20-27)	project presentations

Tentative Case Study Schedule

Level	Case Study	Methods
Simple	← Tipping Behavior →	Summary statistics; Categorical variable plots, continuous variable plots, re- structuring variables, linear regression.
	↓	

Not so simple ← Italian Olive Oils → Classification: LDA, QDA, neural networks, support vector machines, cross-validation.

↓
Less than complex ← Hurricanes → Time/Space dependence: variograms, time series plots; missing values.

↓
Complex ← ? → Unsupervised classification: hierarchical, k -means, model-based clustering; pattern extraction.

Methods

- Data: Manipulating flat files, working with data stored in a database.
- Data cleaning: re-formulating variables to extract different types of information, handling missing values, fixing errors in data, transformations.
- Interactive and dynamic exploratory graphics.
- Classical statistical procedures: regression, GLM, PCA, hierarchical and k-means clustering.

- Computationally intensive data-centered procedures: Smoothing, bootstrap, projection pursuit, neural networks, support vector machines, trees and forests.
- Presentation graphics: how to nicely/formally present the information uncovered to a new audience.

Software

- R: A free data analysis software package available from www.R-project.org. This is a very powerful data analysis package that has many additional packages for specialist tasks, and reasonably good presentation graphics. It operates on a variety of operating systems and hardware.
- GGobi: A free direct manipulation and dynamic graphics package available from www.ggobi.org. Operates on Windows and linux computers.
- mysql: A free database software package available from www.mysql.org.