Graduate Programs in Epidemiology and Public Health
University of Miami, Miller School of Medicine

BST 640 — Modern Numerical Multivariate Methods

Fall 2015
Thursday 10:00-12:30, CRB 995

Course Instructor: Hemant Ishwaran, Ph.D.
http://web.ccs.miami.edu/~hishwaran
hemant.ishwaran@gmail.com
CRB 1058

Course Description:

This course covers multivariate topics from both a classical as well as modern perspective. Topics chosen from: Multivariate Normal Distribution; Bayesian Multivariate Normal Model; Cross-Validation; Prediction Error; Penalization/Regularization; Simulations; Multivariate Regression; Spectral Decomposition; Principal Component Analysis; Canonical Correlation Analysis; Discriminant Analysis; Newton's Method; Steepest Descent; Gradient Boosting; Coordinate Descent; Trees; Bagging; Forests; New Developments for Forests. The R programming language (http://www.r-project.org) will be used extensively throughout the course for computation and statistical analysis.

Prerequisites: EPI 501/502 and one of EPH 605, MTH 525, BST 575, or with permission of instructor.

Course Learning Objectives:

Students will learn to synthesize multivariate statistical concepts with numerical application implemented via the R programming language. Students will learn how to deal with many types of data including large data sets and non-conventional data challenging to standard multivariate methods and they will learn how to apply good numerical practice when analyzing such data. In addition to being exposed to a collection of classical multivariate topics they will also be presented with a modern framework for multivariate optimization applicable to a wide range of statistical problems that they will conceptually master. Students will apply their skills to various case studies and through considering public and synthetic data.

Course Requirements:

Participation in class discussions, completion of homework assignments (which will include problem solving and analyzing real data), and a term project involving an in depth statistical analysis.

Course Material:

R code, datasets, homework assignments, and solutions will be posted on Blackboard. Students must have access to computers with R installed (http://cran.r-project.org). A set of typed notes to include all material to be covered in the course will be made available on Blackboard (these notes are meant only for students enrolled in the course and should not be shared or distributed in any way). These lecture notes will form the basis for the course text.

In addition to the lecture notes, the follow set of optional references may be beneficial:

1. An Introduction to R (`http://cran.r-project.org/doc/manuals/R-intro.html`).
2. Applied Multivariate Statistical Analysis by R.A. Johnson and D.W. Wichern.
3. Modern Multivariate Statistical Techniques by A.J. Izenman.

Grading:

Homework (6-7 assignments) (60%); Term Project (40%).

Homework problem sets will typically be a mixture of theory problems which require analytical solutions as well as real data applications/case studies which will require applying multivariate techniques implemented computationally using the R computing language. Students are permitted to work together but solutions must be written independently. The term project requires the student to do an in depth statistical analysis. For this you must select a multivariate data set (check with me on its suitability) and apply a modern statistical analysis (again, please check with me). In your written report you should give a full description of the contents of the data, your hypothesis, the model you fit, and the results from the fit using R. The report should clearly state an overall conclusion that must be defended by your analysis. *The complete write up, including tables, graphs, and references should consist of no more than 15 pages. Do not attach unedited computer output. Computer code used for the analysis should be sent by email to me.*

Course Schedule:

Class 1-2 ..... R/Matrix Linear Algebra (Chapters 2-3)

1. The R project; R programming
2. Review of matrix and linear algebra
3. Homework #1

Class 2 .....Multivariate Normal Distributions (Chapter 4)

1. Basic properties
2. Testing multivariate normality
3. Examples and applications

Class 3 .....Linear Regression (Chapter 5)

1. Least squares
2. Model selection
3. Prediction error
4. Cross-validation
5. Penalization: bridge estimators, lasso, lars
6. Simulations
7. Homework #2

Class 4 ..... Principal Components (Chapter 7)

1. Spectral decomposition theorem
2. Principal component transformation
3. Sample principal component analysis
4. Singular value decomposition
5. Examples
6. Homework #3

Class 5 ..... Canonical Correlation Analysis (Chapter 8)

1. Mathematical development
2. Sample canonical correlation analysis
3. R functions
4. Examples

Class 5 ..... Discriminant Analysis (Chapter 9)

1. Introduction
2. Maximum likelihood approach
3. Bayes discriminant rule
4. Linear discrimination
5. Quadratic discrimination
6. Supervised case
7. Examples
8. Homework #4

Class 6 ..... Newton's Method (Chapter 10)

1. Introduction, method, examples
2. Multivariable maximization
3. Logistic regression
4. Comparison of numerical approaches; R functions
5. Fisher scoring
6. Discuss term project with Dr. Ishwaran

Class 7 ..... Steepest Descent (Chapter 11)

1. Introduction, illustrations
2. R functions
3. Homework #5

Class 7-8 ..... Gradient Boosting (Chapters 12)

1. Introduction: the AdaBoost algorithm
2. Generic gradient boosting
3. Gradient boosting for linear regression
4. Connections to lasso
5. Gradient boosting for logistic regression
6. Present a draft of term project to Dr. Ishwaran

Class 9 ..... Trees (Chapter 15)

1. Classification trees
2. Regression trees
3. Weighted averaged estimators
4. Survival trees
5. Gradient boosting using trees
6. Homework #6

Class 10 ..... Tree Instability (Chapter 16)

1. Introduction
2. Bagging
3. Why bagging works

Class 10-11 ..... Random Forests (Chapter 17)

1. Breiman's random forests (RF)
2. The RF predictor
3. RF as a nearest neighbour predictor
4. In-bag and out-of-bag prediction and error
5. RF regression and classification
6. Examples
7. Software
8. Homework #7

Class 12 ..... Random Forests Variable Selection (Chapter 18)

1. Variable importance (VIMP)
2. Joint VIMP for interactions
3. Minimal depth variable selection
4. Variable hunting

Class 13 ..... Random Survival Forests (Chapter 19)

1. Introduction
2. Survival ensembles
3. Examples
4. Extensions

Class 14 ..... New Developments of Random Forests

1. Missing data
2. Unsupervised forests
3. Multivariate forests

Classes End ..... Term project due