BIOS 740: STATISTICAL LEARNING AND HIGH-DIMENSIONAL DATA

Spring, 2011

• COURSE DESCRIPTION (3 credit hours)
The course gives an introductory overview of statistical learning methods with or without high-dimensional data. It consists of three major components: learning methods, learning theory and methods for high-dimensional data. The first part on learning methods provides a complete review of supervised learning methods (discriminant analysis, kernel methods, nearest neighborhood, tree methods, neural network, support vector machine, random forest, and boosting methods) and unsupervised learning methods (principal component analysis, factor analysis, cluster analysis, multidimensional scaling, self organizing map). R-functions and real data demo are used for illustration. The second part on learning theory provides some foundational theory for statistical learning methods and it covers theories of Bayesian error, concentration inequalities, VC-theory, risk bound and etc. The last part of the course focus on current statistical methods with high-dimensional data including dimension reduction, variable selection and multiple testing, along with a number of real applications.

• MEETING TIME 11:00-12:15, Monday and Wednesday @ Room MCG2305

• CLASS WEBSITE http:\\ www.bios.unc.edu\ ∼dzeng\Bios740.html

• LECTURE NOTES AND TEXTBOOKS
  − Lecture notes (can be downloaded from the class website)
  − The Elements of Statistical Learning: Data, Mining, Inference, and Prediction, by Hastie et al. (downloadable from their website)
  − A Probabilistic Theory of Pattern Recognition, by Devroye et al.
  − Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond, by Schlkopf and Smola
  − Principles and Theory for Data Mining and Machine Learning, by Clarke et al.

• INSTRUCTOR
  Dr. Donglin Zeng
  Office: 3103B McGavran-Greenberg Building
  Email: dzeng@bios.unc.edu
  Phone: (919)966-7273
  Office hours: 1-3 Friday

• GRADING SYSTEM
  There will be no exams for this course. Homework assignments will be given occasionally. A final project, which can be research-oriented work, or paper review, or real data application, will be required. Final grades will be based on the performance of homework and the final project including the final presentation. For the final project, students are encouraged to meet the instructor to discuss choices around the middle of the semester.

• TOPICS TO BE COVERED
  1. Supervised Learning (10 lectures)
     − Statistical decision theory
     − Direct learning: parametric methods
– Direct learning: Semi-Nonparametric methods
– Direct learning: Nonparametric methods
– Indirect learning

2. Unsupervised Learning (2 lectures)
   – Principal component analysis
   – Latent component analysis
   – Multidimensional scaling
   – Cluster analysis

3. Learning Theory (10 lectures)
   – Bayes error
   – Consistency of direct learning methods
   – Consistency of indirect learning methods
   – Convergence rates
   – Classification error estimation
   – Concentration inequalities

4. High-dimensional Data (6 lectures)
   – Dimension reduction
   – Variable selection
   – Multiple testing
   – Application-specific methods

• OTHER INFORMATION

– Teaching tool will be mainly based on the use of the projector, sometimes with the help of chalkboards or handouts.
– I will be out of town March 21-30 so we need to make up 4 classes in April. My initial plan is to have them on Fridays of April but this is subject to change.
– The classes on April 20, 25, 27 and May 2 are reserved for students to present their final projects.