Statistical/Machine Learning

Semester 1 2020-2021

Lecturer: Saharon Rosset
Schreiber 022
saharon@post.tau.ac.il
Office hrs: By email coordination
Textbook: *Elements of Statistical Learning* by Hastie, Tibshirani & Friedman

Syllabus

The goal of this course is to gain familiarity with the basic ideas and methodologies of statistical (machine) learning. The focus is on supervised learning and predictive modeling, i.e., fitting $y \approx \wedge f(x)$, in regression and classification.
We will start by thinking about some of the simpler, but still highly effective methods, like nearest neighbors and linear regression, and gradually learn about more complex and "modern" methods and their close relationships with the simpler ones.
As time permits, we will also cover one or more industrial "case studies" where we track the process from problem definition, through development of appropriate methodology and its implementation, to deployment of the solution and examination of its success in practice.
The homework and exam will combine hands-on programming and modeling with theoretical analysis. Topics list:

- Introduction (text chap. 1,2): Local vs. global modeling; Overview of statistical considerations: Curse of dimensionality, bias-variance tradeoff; Selection of loss functions; Basis expansions and kernels
- Linear methods for regression and their extensions (text chap. 3): Regularization, shrinkage and principal components regression; Quantile regression
- Linear methods for classification (text chap. 4): Linear discriminant analysis; Logistic regression; Linear support vector machines (SVM)
- Classification and regression trees (text chap. 9.2)
- Model assessment and selection (text chap. 7): Bias-variance decomposition; In-sample error estimates, including C_p and BIC; Cross validation; Bootstrap methods
- Basis expansions, regularization and kernel methods (text chap. 5,6): Splines and polynomials; Reproducing kernel Hilbert spaces and non-linear SVM
- Committee methods in embedded spaces (material from chaps 8-10): Bagging and boosting
- Deep learning and its relation to statistical learning
- Case studies: Customer wallet estimation; Netflix prize competition; maybe others...

Prerequisites

Basic knowledge of mathematical foundations: Calculus; Linear Algebra; Geometry
Undergraduate courses in: Probability; Theoretical Statistics
Statistical programming experience in R is not a prerequisite, but an advantage

Books and resources

Textbook:
Elements of Statistical Learning by Hastie, Tibshirani & Friedman
Book home page (including downloadable PDF of the book, data and errata)

Other recommended books:
Computer Age Statistical Inference by Efron and Hastie
Modern Applied Statistics with Splus by Venables and Ripley
Neural Networks for Pattern Recognition by Bishop
(Several other books on Pattern Recognition contain similar material)
All of Statistics and *All of Nonparametric Statistics* by Wasserman

Online Resources:
Data Mining and Statistics by Jerry Friedman
Statistical Modeling: The Two Cultures by the late, great Leo Breiman
Course on Machine Learning from Stanford's Coursera.
The Netflix Prize competition is now over, but will still play a substantial role in our course.

Grading

There will be about four homework assignments, which will count for about 30% of the final grade, and a final which will count for 70%, which will include take-home exam and possibly an in-class exam. Both the homework and the exam will combine theoretical analysis with hands-on data analysis.
We will also have an optional data modeling competition, whose winners will get a boost in grade and present to the whole class.

Computing

The course will require extensive use of statistical modeling software. It is *strongly* recommended to use R (freely available for PC/Unix/Mac) or its commercial kin Splus.
R Project website also contains extensive documentation.
A basic "getting you started in R" tutorial. Uses the Boston Housing Data (thanks to Giles Hooker).
Modern Applied Statistics with Splus by Venables and Ripley is an excellent source for statistical computing help for R/Splus.