Centre for Systems Informatics Engineering
Department of Systems Engineering and Engineering Management
City University of Hong Kong

Art and Practice of Regression Trees and Forests

Professor Wei-Yin LOH
Department of Statistics, University of Wisconsin

Date: 28th May 2018 (Monday) &
29th May 2018 (Tuesday)
Time: 9:00am to 12:00nn
Venue: P7303, Yeung Kin Man Academic Building (AC1)

Abstract
Regression tree and forest methods have greatly improved in the last decade. Their ease of use, prediction accuracy, execution speed, and interpretability make them essential tools for machine learning and data analysis. The course teaches how to use the tools effectively and efficiently in practice. It follows an example-focused style, with each example chosen to illustrate particular weaknesses of traditional solutions and to show how tree methods overcome them and yield new insights. Examples include a large consumer survey with hundreds of variables and substantial amounts of missing values; cancer and diabetes randomized trials with censored and longitudinal responses for precision medicine; and observational studies of high-school dropouts and Alzheimer's patients. Learning highlights are (1) how trees deal with missing values without requiring imputation, (2) how importance scores help with variable selection, and (3) how to perform post-selection inference with the bootstrap. To encourage hands-on training, the presentation is interwoven with live demos of free software. No commercial software is required. Specific algorithmic techniques are discussed where appropriate but no systematic presentation of entire algorithms is given. Attendees should have experience with linear and logistic regression. Instructions for software and dataset downloads will be given in advance.

About the Speaker
Wei-Yin Loh is professor of statistics at the University of Wisconsin, Madison, and an ASA and IMS Fellow. He has been using and developing classification and regression algorithms and software for more than 30 years. He is the sole developer of the GUIDE algorithm and co-developer of the FACT, QUEST, CRUISE, and LOTUS algorithms. He regularly teaches semester-long undergraduate and graduate courses on the subject at his university and has given one- and multi-day short courses at professional meetings (U.S. Army Applied Statistics Conference 1995, 1999; KDD 1999, 2001; JSM 2007, 2011, 2013, 2015; ICSA Applied Statistics Conference 2015; Midwest Biopharmaceutical Statistics Workshop 2015; ASA Conference on Statistical Practice 2017; Deming Conference 2017; Interface Conference 2013), ASA chapters (Northeastern Illinois Chapter 2014, Washington Statistical Society 2016), biopharma companies (Merck 2007; Abbott 2011; Eli Lilly 2011; Biogen Idec 2014; Gilead Sciences 2016; AbbVie 2016; Takeda 2017; MedImmune 2017), and overseas academic institutions (National University of Singapore 2010, 2014, 2017; East China Normal University 2012; National Tsinghua University, Taiwan, 2012; City University of Hong Kong 2014; Academia Sinica, Taipei, 2017).

Outline

The course may be subtitled "Classification and Regression Trees by Example." Specially selected real datasets are used to motivate and illustrate particular difficulties faced by traditional techniques and how they are overcome and solved in new ways by tree methods. Live demos of free software are interwoven in the presentation to encourage hands-on training. No commercial software is required. The target audience is statisticians, data scientists, and researchers in business, government, industry, and academia. It should be particularly useful for those who need to explore and analyze complex datasets with many variables and missing values and who want to learn to use the free classification and regression tree software.

1. Estimation of population mean income from a consumer expenditure survey. Aims are to show that (i) popular missing value methods such as multiple imputation are grossly inadequate when the amount of missing values is substantial and the variables number in the hundreds, (ii) regression tree and forest models can be constructed easily without missing value imputation, and (iii) tree algorithms can quickly order the predictor variables in terms of their predictive importance. Material based on Loh et al. (2017). Data from Bureau of Labor Statistics.

2. Classification of peptide sequences. Introduce concepts of node impurity in classification tree models with categorical predictors, Compare CART (Breiman et al. 1984), GUIDE (Loh 2002, 2009) and Random forest (Breiman 2001) in their importance scoring of variables. Compare GUIDE with neural networks on predictive accuracy.

3. Birthweight data. Introduce concepts of class priors and misclassification costs. Show how to build classification tree models from data with rare events or highly unbalanced classes. Data from Centers for Disease Control.

4. College tuition data. Build quantile regression tree models to estimate upper percentiles of tuition in U.S. colleges. Also build single regression tree models that simultaneously predict tuition cost and graduation rate. Data from U. S. News & World Report.

5. Hourly wages of high-school dropouts. Show the deficiencies of traditional linear mixed models. Build regression tree models for data with time-varying and longitudinal responses. Material based on Loh and Zheng (2013). Data from Singer and Willet (2003).

6. Alzheimer's disease data. Cluster response trajectories of Alzheimer's patients using patient baseline measurements. Compare results with those obtained with traditional clustering methods. Data from Alzheimer's Disease Neuroimaging Initiative (ADNI).

7. Breast cancer randomized trial. Discuss problems with subgroup identification for differential treatment effects using traditional proportional hazards models. Compare regression tree solutions that adjust for local linear effects of prognostic variables. Material based on Loh et al. (2015, 2017). Data from Schumacher et al. (1994).

8. Type II diabetes randomized trial. Identify subgroups with differential treatment effects for longitudinal response data. Material based on Loh et al. (2016). Data from Eli Lilly.

9. Mortality from cardiovascular disease. Use classification tree models for matching and propensity scoring to estimate effect of hypertensive treatment in observational data. Compare results with those from logistic regression. Data from NIH Framingham Heart Study.

10. Post-selection inference. Regression tree methods have long been considered useful only for exploratory purposes, due to hitherto nonexistent methods of statistical inference. The problem is due to the difficulty of adjusting for the many algorithmic steps employed in the search for splits. This all changed very recently with the development of a bootstrap calibration technique that yields a theoretically justifiable method for construction of confidence intervals for subgroup means in the terminal nodes of a tree. Material based on Loh (1987, 2016, 2017).

Enquiry: lolli.lee@cityu.edu.hk