Classification and Regression Trees and Forests
December 9-10, 2013

Wei-Yin Loh, University of Wisconsin, Madison

Classification and regression trees are essential tools for data mining, machine learning, and statistical data analysis. This year marks the 50th anniversary of the publication of the first journal article on the subject. In a classification or regression tree model, the data and sample space are split into two or more partitions and a simple statistical model is fitted to each of them. The model is intuitive to interpret because the partitions can be displayed as a decision tree. Besides, the models often possess prediction accuracy as good as or better than that of linear discriminant analysis and linear regression. This course reviews the major techniques and discusses their relative strengths, weaknesses, capabilities, and computational requirements. Extensions to ensemble procedures, such as bagging and random forest, are included. Concepts are explained with examples from business, industry, science, and engineering.

Special emphasis is given to the instructor's GUIDE algorithm (http://www.stat.wisc.edu/~loh/guide.html). For classification, GUIDE can construct decision trees with nonparametric models, such as nearest neighbor and kernel discrimination, in the nodes. For regression, GUIDE can use least squares, least median of squares, quantile, Poisson, and proportional hazards loss functions and apply them to univariate, multivariate, longitudinal, and censored response data. If time permits, instruction on the use of GUIDE and other free software will be given.

For a brief review of the subject, see Loh, W.-Y. (2011). "Classification and regression trees." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1, 14-23. (http://www.stat.wisc.edu/~loh/treeprogs/guide/wires11.pdf).

For a list of sample applications in the scientific literature, see http://www.stat.wisc.edu/~loh/apps.html.