Cross-validating Algorithms for Predicting Clinical Outcomes: Applications in Genomic Data

Dr. Annette Molinaro
Assistant Professor, Division of Biostatistics
Yale University, New Haven, CT

Abstract:

Clinicians aim toward a more preventative model of attacking cancer by pinpointing and targeting specific early events in disease development. These early events can be measured as genomic, proteomic, epidemiologic, and/or clinical variables. Such measurements are then used to predict clinical outcomes such as primary occurrence, recurrence, metastasis, or mortality.

Recursive partitioning seeks to explain the individual contributions of various covariates as well as their interactions for the purposes of predicting outcomes, either continuous or categorical. Potential algorithms such as Classification and Regression Trees (CART) and DSA-Partitioning aggressively search highly-complex covariate spaces.

There are several important considerations when using such algorithms. The first is to not overfit the data. The second consideration is the stability of the resulting predictor. Algorithms such as CART are sensitive to data fluctuations and, thus, given a perturbation will potentially build a different predictor than that built on the original data. A third consideration is variable importance. In this talk, such considerations will be discussed and results comparing both algorithms presented.