Data-Driven Diverse Logistic Regression Ensembles

A novel framework for statistical learning is introduced which combines ideas from regularization and ensembling. This framework is applied to learn an ensemble of logistic regression models for high-dimensional binary classification. In the new framework the models in the ensemble are learned simultaneously by optimizing a multi-convex objective function. To enforce diversity between the models the objective function penalizes overlap between the models in the ensemble. Measures of diversity in classifier ensembles are used to show how our method learns the ensemble by exploiting the accuracy-diversity trade-off for ensemble models. In contrast to other ensembling approaches, the resulting ensemble model is fully interpretable as a logistic regression model, asymptotically consistent, and at the same time yields excellent prediction accuracy as demonstrated in an extensive simulation study and gene expression data applications. The models found by the proposed ensemble methodology can also reveal alternative mechanisms that can explain the relationship between the predictors and the response variable.