Including prior knowledge in shrinkage classifiers for genomic data

Jean-Philippe Vert (Mines ParisTech - Institut Curie)

April 30, 2010

Estimating predictive models from high-dimensional and structured genomic data, such as gene expression of CGH data, measured on a small number of samples is one of the most challenging statistical problems raised by current needs in post-genomics. Popular tools in statistics and machine learning to address this issue are shrinkage estimators, which minimize an empirical risk regularized by a penalty term, and which include for example support vector machines or the LASSO. In this talk I will discuss new penalty functions for shrinkage estimators, including generalizations of the LASSO which lead to particular sparsity patterns, and which can be seen as a way to include problem-specific prior information in the estimator. I will illustrate the approach by several examples such as the classification of gene expression data using gene networks as prior knowledge, or the classification and detection of frequent breakpoints in CGH profiles.