Abstract

This thesis studies the integration of a particular family of machine learning methods into the analysis of high dimensional data sets from microarray and mass spectrometry technologies. More precisely, the class of regularized least squares methods and kernel-based machine learning methods are considered. Kernel-based methods integrate techniques from convex optimization, functional analysis and statistical learning theory into a powerful yet simple framework, and have enjoyed in the last decade great success over a wide range of application tasks in bioinformatics, financial engineering, text mining, image processing and time series prediction.

At the beginning of this thesis fundamentals and principles on regularized learning and kernel-based methods are reviewed. After this, we concentrate on support vector machines (SVM) methods and, particularly, the Least-Squares Support Vector Machines (LS-SVM) formulations. Based on the structure of the LS-SVM classifiers, we propose a novel algorithm for variable selection that makes use of low rank matrices to update the LS-SVM parameters. This method provides efficient simplifications for linear kernels that can scale up to thousands of variables and it is therefore applicable in microarray analysis. Subsequently, we provide extensions of this framework to polynomial kernels using additive structures. Existing methods discard the use of resampling techniques due to its computational complexity. We provide, however, efficient implementations of the leave-one-out (LOO) estimator which is in turn the mechanism to assess the relevance of the variables.

Later on we turn our point of interest on the application of regularized learning methods to Mass Spectral Imaging (MSI) data. We present an approach to learn sparse predictive models that integrate structural information from MSI data. The goal is to develop (semi)-supervised models that use the labeled portions of the tissue to help predict the anatomical labels for the unlabeled regions. The medical objective of such models would be, for instance, to provide the pathologist with insight in interpreting molecular tissue content of areas that do not lend themselves for straightforward human classification. Particularly we address issues regarding inherent ordering of the model variables and the spatial relationships of the data points. Furthermore, to overcome the lack of labeled data points we exploit the prior assumption that nearby spectra are likely to have the same label. Numerically, the proposed model is shown to be equivalent to a LASSO formulation and therefore can be efficiently solved via the LARS (Least Angle Regression) algorithm. Moreover, each component in our optimization problem clearly embodies the structural information of MSI data.

As a last application of kernel methods in this thesis, we deal with the problem of clustering genes from microarray data. To this end, we explore the use of a spectral clustering formulation that incorporates an extension for out-of-sample points. In our approach, an informative small subset genes is first selected via entropy maximization, and subsequently the clustering model is used to infer cluster memberships for the remaining genes.