Adaptive methods in classification

Sara van de Geer
University of Leiden

Suppose we have data $\{(X_i, Y_i) : \ i = 1, \ldots, n\}$, where $X_i \in \mathcal{X}$ is a feature of observation i and Y_i is its label. The classification problem is to predict the label Y of a new observation X. This prediction, $\hat{Y}$ say, is a function of the data and of X. We write $\hat{Y} = \hat{f}(X)$, where $\hat{f}$ is a function depending on the data $\{(X_i, Y_i) : \ i = 1, \ldots, n\}$. More generally, we call any function f on the feature space a *classifier*.

In this talk, we will first review some results for the empirical risk minimizer. Suppose we are in the binary case $Y \in \{0, 1\}$. Let $\mathcal{G}$ be a collection of subsets of $\mathcal{X}$, the so-called model class. The empirical risk minimizer $\hat{G}_n$ is the classifier within $\mathcal{G}$ which makes the smallest number of errors in the sample. We will derive the prediction error of this classifier.

The prediction error depends to a large extend on the *richness* of the model class $\mathcal{G}$. The model selection problem is to choose $\mathcal{G}$ in such a way that a balance is obtained between estimation error and approximation error. This is similar to a bias-variance trade-off. We Will show how this problem can be approached using complexity regularization.

It turns out that balancing estimation error and approximation error is particularly difficult in classification, because the estimation error depends heavily on the unknown distribution of the data. As additional problem, empirical risk minimization over a large model class $\mathcal{G}$ is computationally very difficult. To handle both problems, we propose to use the support vector machine loss function, and to employ complexity regularization through a soft thresholding type penalty. We will show that this procedure yields estimators that adapt to model complexity and other properties of the underlying distribution.