A theoretical comparison of regularized classifiers

Sara van de Geer[a]

[a]Mathematical Institute, University of Leiden

In binary classification, the problem is to predict a label $Y \in \{\pm 1\}$ given a feature $X \in \mathcal{X}$. A classifier is of the form $\text{sign}(f(X))$, where $f : \mathcal{X} \to \mathbf{R}$ is some function of X. In fact, we will also call f itself a classifier. Using a training set $\{(X_i, Y_i)\}_{i=1}^n$ of i.i.d. copies of (X, Y), the aim is to construct a classifier $\hat{f}_n$ with small prediction error. To this end one may consider various model classes $\mathcal{F}$ as candidate classifiers. various loss functions and also various complexity penalties.

We will study the case where $\mathcal{F}$ is a subset of a linear space, say

$$\mathcal{F} \subset \{f = \sum_{j=1}^{m} \alpha_j \psi_j : \alpha \in \mathbf{R}^m\},$$

where $\psi_j : \mathcal{X} \to \mathbf{R}$ $(j = 1, \ldots, m)$ are given base functions. Examples of base functions are those corresponding to a kernel representation, the base functions may be $\{\pm 1\}$-valued base classifiers (in the case of averaging classifiers), or they may form an orthogonal system. Examples of loss functions are: exponential, logit, or hinge loss (support vector machines). Examples of penalties are L_2 norms (e.g. induced by a kernel), ℓ_1 norms on the coefficients, penalties based on the dimensionality or other measures of complexity.

We will put these various choices in a single framework, and derive inequalities for the excess risk of the classifier $\hat{f}_n$. Our results depend on the margin behavior, in particular on the margin parameter κ as introduced by Tsybakov (Ann. Statist. 2004). We illustrate that a comparison of exponential loss or logit loss with hinge loss, depends on the (unknown) smoothness of the regression $\eta(X) = P(Y = 1|X)$. For kernel vector machines, an optimal tuning of the smoothing parameters may require knowing the margin parameter κ. We also give an example where hinge loss yields a classifier that is both adaptive to κ as well as to the "smoothness" of the boundary of Bayes classifier.