Estimators for Supervised Learning

Ronald DeVore [*]

Abstract

The following regression problem occurs in various settings of supervised learning. We are given a domain $X \subset I\!R^d$ and an interval $Y = [-M, M] \subset I\!R$. There is a probability measure ρ defined on $Z := X \times Y$ which is unknown to us. We see information about ρ from samples $\mathcal{Z} := \{(x_i, y_i)\}_{i=1}^m$ which are drawn at random with respect to ρ. We are interested in learning the function $f_\rho(x) := \int\limits_Y y\, d\rho(y|x)$ which is the expected value of y given x.

One way of constructing an estimator for f_ρ is to choose a space of functions $\mathcal{S}$ (called the hypothesis space in learning theory) and find the function $f_{\mathcal{Z}}$ from $\mathcal{S}$ which minimizes the empirical least squares error:

$$f_{\mathcal{Z}} := \underset{f \in \mathcal{S}}{argmin} \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2. \tag{0.1}$$

Cucker and Smale have shown how to estimate the performance of such an estimator by using the Kolmogorov entropy of the set $\mathcal{S}$ in L_∞. They then build estimators in the following way. Given a priori knowledge that f_ρ is in a ball $B(W)$ of a smoothness class W, they choose $\mathcal{S}$ depending on this knowledge and then obtain the estimator $f_{\mathcal{Z}}$. A typical example of their theory gives that if $f_\rho \in B(W)$ with $W = W^r(L_\infty(X))$ then choosing $\mathcal{S}$ as a ball in $W^r(L_\infty(X))$ one obtains

$$E(\|f_\rho - f_{\mathcal{Z}}\|_{L_2(\rho)}) \leq cm^{-\frac{r}{2r+2}}, \quad m = 1, 2, \dots. \tag{0.2}$$

This talk will center on several ways to improve on these estimators. We show that using tools from approximation theory such as linear or nonlinear widths then we can construct estimators where the right side of (0.2) is replaced by $cm^{-\frac{r}{2r+1}}$ and that this decay rate is optimal. Secondly, the class $W^r(L_\infty(X))$ can be replaced by much larger classes such as the Sobolev classes $W^r(L_p(X))$, $p > d$. Perhaps most importantly we show how to construct estimators that do not require any a priori assumptions on f_ρ but still perform optimaly for a large range of smoothness classes. These latter estimators are built using nonlinear methods of approximation such as adaptive partitioning and n term wavelet approximation.

This work is in collaboration with Peter Binev, Albert Cohen, Wolfgang Dahmen, Gerard Kerkyacharian, Dominique Picard, and Vladimir Temlyakov.

[*] Dept. of Mathematics, University of South Carolina, Columbia, SC 29208