Universal Algorithms for Machine Learning

Wolfgang Dahmen, RWTH Aachen

This talk draws on joint work with A. Barron, P. Binev, A. Cohen and R. DeVore. In the context of supervised learning it is mainly concerned with estimating regression functions from observed data in the following setting. We suppose that ρ is an unknown measure on a product space $Z := X \times Y$, where X is a bounded domain of $\mathbb{R}^d$ and $Y = \mathbb{R}$. Given m independent random observations $z_i = (x_i, y_i)$, $i = 1, \ldots, m$, identically distributed according to ρ, we are interested in estimating the *regression function* $f_\rho(x)$ defined as the conditional expectation of the random variable y at x:

$$f_\rho(x) := \int\limits_Y y d\rho(y|x)$$

with $\rho(y|x)$ the conditional probability measure on Y with respect to x.

One of the goals of learning is to provide estimates under minimal restrictions on the measure ρ since this measure is generally unknown. However, the typical working assumption is that this probability measure is supported on $X \times [-M, M]$, i.e $|y| \leq M$ almost surely, which is the case in many applications. This type of regression problem is referred to as *random design* or *distribution-free*.

Since f_ρ is the minimizer of the risk functional

$$\mathcal{E}(f) := \int\limits_Z (y - f(x))^2 d\rho,$$

over $f \in L_2(X, \rho_X)$ (the space of all functions from X to Y which are square integrable with respect to ρ_X) we seek an estimator $f_{\mathbf{z}}$ for f_ρ based on the data set $\mathbf{z} = \{z_1, \ldots, z_m\} \subset Z^m$ such that the quantity $\|f_{\mathbf{z}} - f_\rho\|_{L_2(X,\rho_X)}$ tends to zero either in *probability* or in *expectation* at a possibley high rate as the sample size m increases.

The estimators should be *universal* in the sense that the algorithm does *not use* any a-priori assumption that the regression function f_ρ belongs to some specific function class while the convergence rates are optimal or "good" whenever f_ρ happens to belong to (some range of such) function classes permitting these convergence rates.

We discuss several concepts for constructing such estimators as well as typical obstructions caused, for instance, by high dimensionality. In this latter context some recent results on greedy estimators are outlined.