Universal algorithms for learning theory

A. Cohen

(joint work with P. Binev, W. Dahmen, R. DeVore and V. Temlyakov)

We shall present the construction and analysis of a universal estimator for the mathematical problem of supervised learning. In this problem, we observe the data $\mathbf{z} = (z_1, \ldots, z_m) \subset X \times Y$ of m independent random observations $z_i = (x_i, y_i)$, $i = 1, \ldots, m$, identically distributed according to a probability ρ on a product space $X \times Y$. We are interested in estimating the *regression function* $f_\rho(x)$ defined as the conditional expectation of the random variable y at x:

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

with $\rho(y|x)$ the conditional probability measure with respect to x. The estimator $f_{\mathbf{z}}$ is assessed by the measure of the error $\|f_\rho - f_{\mathbf{z}}\|$ in the $L^2(X, \rho_X)$ metric with ρ_X the marginal probability which is unknown. This type of problem is refered to as *distribution-free*, see [**?**] for a general introduction.

Universal means that the estimator does not depend on any a priori assumptions about the regression function to be estimated. Our universal estimator, introduced in [**?**], consists of a least-square fitting procedure using piecewise constant functions on a partition which depends adaptively on the data. The partition is generated by a splitting procedure which somehow differs from those used in CART algorithms [**?**] in the sense that it is based on thresholding empirical quantities which play the role of wavelet coefficients.

It is proven that this estimator performs at the optimal convergence rate for a wide class of priors on the regression function. Namely if the regression function is in a smoothness space of order not exceeding one (a limitation resulting because the estimator uses piecewise constants) then the estimator converges to the regression function (in the least squares sense) with an optimal rate of convergence in terms of the number of samples. The estimator is also numerically feasible and can be implemented on-line.

References

[1] P. Binev, A. Cohen, W. Dahmen, R. DeVore and V. Temlyakov, Universal algorithms for learning theory - Part I : piecewise constant functions, preprint, 2004

[2] D.L. Donoho, CART and best-ortho-basis : a connection, Ann. Stat. **25**, 1870-1911, 1997.

[3] L. Györfy, M. Kohler, A. Krzyzak, A. and H. Walk, A distribution-free theory of nonparametric regression, Springer, Berlin, 2002.