On Regression in Extreme Regions

Stephan Clémençon[1], Nathan Huet[1], Anne Sabourin[2]
[1] LTCI, Télécom Paris, Institut polytechnique de Paris
[2] Université Paris Cité, CNRS, MAP5 F-75006 Paris, France

In the classic regression problem, the value of a real-valued random variable Y is to be predicted based on the observation of a random vector X, taking its values in $\mathbb{R}^d$ with $d \geq 1$ say. The statistical learning problem consists in building a predictive function $\hat{f} : \mathbb{R}^d \to \mathbb{R}$ based on independent copies of the pair (X, Y) so that Y is approximated by $\hat{f}(X)$ with minimum error in the mean-squared sense. Motivated by various applications, ranging from environmental sciences to finance or insurance, special attention is paid here to the case of extreme (*i.e.* very large) observations X. Because of their rarity, they contribute in a negligible manner to the (empirical) error and the predictive performance of empirical quadratic risk minimizers can be consequently very poor in extreme regions. In this paper, we develop a general framework for regression in the extremes. It is assumed that (X, Y) distribution belongs to a non parametric class of heavy-tailed probability distributions. It is then shown that an asymptotic notion of risk can be tailored to summarize appropriately predictive performance in extreme regions of the input space. It is also proved that minimization of an empirical and non asymptotic version of this 'extreme risk', based on a fraction of the largest observations solely, yields regression functions with good generalization capacity. In addition, numerical results providing strong empirical evidence of the relevance of the approach proposed are displayed.