Estimating the Expected Error of Empirical Minimizers for Model Selection

Tobias Scheffer
Technische Universität Berlin, FR 5-8,
Franklinstr. 28/29, 10587 Berlin, Germany
scheffer@cs.tu-berlin.de

Thorsten Joachims
Uni Dortmund, LS 8 Informatik
44221 Dortmund, Germany
thorsten@ls8.informatik.uni-dortmund.de

Abstract

Model selection [*e.g.,* 1] is considered the problem of choosing a hypothesis language which provides an optimal balance between low empirical error and high structural complexity. In this Abstract, we discuss the intuition of a new, very efficient approach to model selection. Our approach is inherently Bayesian [*e.g.,* 2], but instead of using priors on target functions or hypotheses, we talk about priors on *error values* – which leads us to a new mathematical characterization of the expected true error. In the setting of classification learning, a learner is given a sample, drawn according to an unknown distribution of labeled instances, and returns the empirical minimizer (the hypothesis with the least empirical error) which has a certain (unknown) true error. If this process is carried out repeatedly, the true error of the empirical minimizer will vary from run to run as the empirical minimizer depends on the (randomly drawn) sample. This induces a *distribution* of true errors of empirical minimizers, over the possible samples drawn according to the unknown distribution. If this distribution would be known, one could easily derive the *expected* true error of the empirical minimizer of a model by integrating over this distribution. This would immediately lead to an *optimal* model selection algorithm: Enumerate the models, calculate the expected error of each model by integrating over the error distribution, and select the model with the least expected error. PAC theory [3] and the VC framework provide worst-case *bounds* on the chance of drawing a sample such that the true error of the minimizer exceeds some ε – "worst-case" meaning that they hold for *any* distribution of instances and any concept in a given class. By contrast, we focus on how to *determine* this distribution for a *fixed, given learning problem* (under some specified assumptions). Unlike the worst-case bound (which depends only on the size, or VC-dimension of the hypothesis space) the actual error distribution depends on the hypothesis space and the unknown distribution of labeled instances itself. However, we can prove that, under a certain assumption of independence of hypotheses, the distribution of true errors – and hence the expected true error – can be expressed as a function of the distribution of empirical errors of uniformly drawn hypotheses (which can be thought of as a prior on error values). The latter distribution (which is *always* one-dimensional) can be *estimated* from a *fixed-sized* initial portion of the training data and a fixed-sized set of randomly drawn hypotheses. This estimate of the distribution now leads us to an *estimate* of the expected true error of the empirical minimizer of the model – which, in turn, leads to a highly efficient model selection algorithm. We study the behavior of this approach in several controlled experiments. Our results show that the accuracy of the error estimate is at least comparable to the accuracy of the estimate obtained by 10-fold cross-validation – provided the prior on error values can be estimated using at least 50 examples. But while 10-CV requires ten invocations of the learner per model, the time which our algorithm requires to assess each model is constant in the size of the model. We also study the robustness of our algorithm against violations of our independence assumptions. We can observe a bias in our predictions when the hypotheses space is of size four or less. When the hypothesis space is of size 40 or more, the dependencies are so diluted that the violations of our assumptions are negligible and do not incur a significant error. The full paper is available at **http://ki.cs.tu-berlin.de/~scheffer/papers/eed-report.ps**.

References

[1] M. Kearns, Y. Mansour, A. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. In *Machine Learning* 27: 7–50. 1997.

[2] J. Rissanen. Minimum-description-length principle. In *Ann. Statist.* 6: 461–464. 1985.

[3] L. G. Valiant. A Theory of the Learnable. In *Comm. ACM* 27. 1984.