Journal of Machine Learning Research 4 (2003) 755-757 Published 10/03

Introduction to the Special Issue on Learning Theory

Ralf Herbrich RHERB@MICROSOFT.COM
Thore Graepel THOREG@MICROSOFT.COM
Microsoft Research Ltd.
7 J J Thomson Avenue
Cambridge, CB3 0FB, UK

This special issue builds on material presented at the Fifteens Annual Conference on Computational Learning Theory (COLT 2002) and the Sixteenth Annual Conference on Neural Information Processing Systems (NIPS*2002), held in Sydney in July 2002 and in Vancouver in December 2002, respectively. We contacted authors who presented work on the analysis of learning in general, and with a particular focus on kernel methods and Bayesian analysis in particular, and asked them to submit an extended journal paper. The six paper in this special issue were selected from a total of nine.

Computational learning theory has a long history dating back to the works of Valiant (1984) and Vapnik (1982). While the theory was initially mostly concerned with questions about learnability and sample complexity for specific fixed classes of functions, there has been a recent trend in studying data-dependent function classes and specific learning algorithms.

A very powerful class of functions is given by the kernel class (see Schölkopf et al. 1999). Over the last few years, there has been some work on the analysis of the effective complexity of kernel classes based on covering number techniques (Schölkopf et al., 1999, Williamson et al., 2000). In *On the Performance of Kernel Classes*, Shahar Mendelson presents a series of complexity results for these classes based on localised Rademacher averages. In particular, the paper investigates the ERM algorithm which finds the kernel predictor with the smallest training error.

The paper *Path Kernels and Multiplicative Updates* by Eiji Takimoto and Manfred Warmuth introduces an interesting new class of kernels which can be used directly in online algorithms with multiplicative updates. In a nutshell, path kernels, which are defined by a directed graph, implicitly map each instance (edge labels) to all exponentially many product features defined by all the paths from a designated source to a designates sink node. Using path kernels, the authors are able to recover a large number of known multiplicative update algorithms. The examples include efficient algorithms for learning disjunctions and a recent algorithm that predicts as well as the best pruning of a series of parallel digraphs.

Mutliplicative algorithms such as Winnow are known to work robustly in the presence of (exponentially) many irrelevant features (Littlestone, 1988). In his paper *Tracking Linear-threshold Concepts with Winnow*, Chris Mesterharm gives a slightly modified version of the Winnow algorithm in which demotions cannot reduce the feature weight below a pre-specified threshold ε. This allows him to show that not only is the total number of mistakes bounded by the logarithm of the total number of features but—even stronger—by the logarithm of the length of the longest input vector under the 1–norm. Furthermore, he can show that this algorithm is also able to perform well if the target concepts slightly drifts in between trials.

Bayesian approaches to learning have played a significant role in the statistical literature over many years. While they lead to excellent performance in practical applications, there have not been

many precise characterisations of their performance for finite sample size under general conditions (McAllester, 1998, 1999). Ron Meir and Tong Zhang are the first to give a quantitative answer to this question for Bayesian mixture algorithms in their paper *Generalization Error Bounds for Bayesian Mixture Algorithms*. Their results demonstrate that mixture approaches are particularly robust. The paper establishes an interesting link between Rademacher analysis and the performance of Bayesian mixtures. However, their approach emphasises the mixture aspect of Bayesian inference and the bounds derived can be applied directly to non-Bayesian mixture approaches such as Boosting and Bagging, as well.

Studying these algorithms, Gilles Blanchard, Gábor Lugosi and Nicolas Vayatis find performance bounds for regularised boosting classifiers in their paper *On the rate of convergence of regularized boosting classifiers*. They find dimension-independent rates of convergence to the Bayes classifier, thereby providing a potential explanation for the success of these algorithms in practice. They present a detailed analysis for boosting decision stumps, and find a theoretical basis for the heuristic of adding a small amount of noise to the training data.

Finally, we have chosen to include the paper *Concentration Inequalities for the Missing Mass and for Histogram Rule Error* by David McAllester and Luis Ortiz. This paper gives an excellent introduction and historical background on large deviation inequalities which lie at the foundations of statistical learning theory. The paper also presents an application of a newly derived large deviation inequality for bounding the concentration of the missing mass estimator. Having drawn a random sample from a collection of items, the *missing mass* is the total probability of all items not yet drawn. Estimators for the missing mass, such as the Good-Turing estimator Good (1953), have been used extensively in language modelling applications. Interestingly, according to Good (2000), the Good-Turing estimator was developed by Alan Turing during World War II while breaking Enigma codes.

Many people have been involved in making this special issue possible. We would especially like to thank the authors for their excellent contributions, the reviewers for their invaluable feedback and Leslie Pack Kaelbling, David Cohn and Karin Bierig for their help at JMLR.

References

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(16):237–264, 1953.

I. J. Good. Turing's anticipation of empirical bayes in connection with the cryptanalysis of the naval enigma. *Journal of Statistical Computation and Simulation*, 66(2):101–112, 2000.

Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

David A. McAllester. Some PAC Bayesian theorems. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 230–234. ACM Press, 1998.

David A. McAllester. PAC-Bayesian model averaging. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 164–170, 1999.

B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors. *Advances in Kernel Methods—Support Vector Learning*. MIT Press, Cambridge, MA, 1999.

B. Schölkopf, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Kernel-dependent support vector error bounds. In *Ninth International Conference on Artificial Neural Networks*, pages 103–108. IEE, 1999.

Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer, Berlin, 1982.

R. C. Williamson, A. J. Smola, and B. Schölkopf. Entropy numbers of linear function classes. In *Proceedings of the Annual Conference on Computational Learning Theory*, pages 309–319. Morgan Kaufmann Publishers, 2000.