Overfitting and Undercomputing in Machine Learning

Tom Dietterich
Department of Computer Science
Oregon State University
Corvallis, Oregon 97331
tgd@cs.orst.edu

A central problem in machine learning is *supervised learning*—that is, learning from labeled training data. For example, a learning system for medical diagnosis might be trained with examples of patients whose case records (medical tests, clinical observations) and diagnoses were known. The task of the learning system is to infer a function that predicts the diagnosis of a patient from his or her case records. The function to be learned might be represented as a set of rules, a decision tree, a Bayes network, or a neural network.

Learning algorithms essentially operate by searching some space of functions (usually called the hypothesis class) for a function that fits the given data. Because there are usually exponentially many functions, this search cannot actually examine individual hypothesis functions but instead must use some more direct method of constructing the hypothesis functions from the data. This search can usually be formalized by defining an objective function (e.g., number of data points predicted incorrectly) and applying various algorithms to find a function that minimizes this objective function. In virtually all interesting cases, the computational problem of minimizing the objective function is NP-hard. For example, fitting the weights of a neural network or finding the smallest decision tree are both NP-complete problems [1, 4]. Hence, heuristic algorithms such as gradient descent (for neural networks) and greedy search (for decision trees) have been applied with great success.

Of course, the sub-optimality of such heuristic algorithms immediately suggests a reasonable line of research: find algorithms that can search the hypothesis class better. Hence, there is been extensive research in applying second-order methods to fit neural networks and in conducting much more thorough searches in learning decision trees and rule sets. Ironically, when these algorithms were tested on real datasets, it was found that their performance was often worse than simple gradient descent or greedy search [3, 5]. In short: it appears to be better not to optimize!

One of the other important trends in machine learning research has been the establishment and nurturing of connections between various previously-disparate fields including computational learning theory, connectionist learning, symbolic learning, and statistics. The connection to statistics was crucial in resolving this paradox.

The key problem arises from the structure of the machine learning task. A learning algorithm is trained on a set of training data, but then it is applied to make predictions on new data points. The goal is to maximize its predictive accuracy on the new data points—not necessarily its accuracy on the training data. Indeed, if we work too hard to find the very best fit to the training data, there is a risk that we will fit the noise in the data by memorizing various peculiarities of the training data rather than finding a general predictive rule. This phenomenon is usually called "overfitting".

Hence, the objective function that we used in formulating the optimization problem ("minimize error on the training data") is in fact not the correct objective function. A large body of work has addressed this problem by augmenting the objective function with various penalty terms (e.g., regularization methods, minimum-description length methods, generalized cross-validation, etc.) [2]. These terms attempt to predict the off-training-set accuracy from its on-training-set accuracy.

With this fix to the objective function, we can again apply our armamentum of optimization algorithms to solve this new optimization problem (but with the same—or worse—computational complexity problems). Experimental results confirm that this removes the paradox.

However, rather than attempting to solve this corrected objective function, we can view the simple gradient descent and greedy algorithms as implicitly embodying correction terms. In other words, a greedy algorithm can be viewed as sub-optimal if its objective is to find the smallest decision tree that fits the data, but it can be viewed as optimal if its objective is to find a tree that minimizes some combination of decision tree size plus a penalty term that corrects for the difference between the training data and the ultimate test data. In a few cases, it is possible to prove this equivalence, but the support for this view is primarily empirical.

I think it is fascinating that by doing a poor job of solving one optimization problem, we are actually doing a good job of solving a different one. In machine learning, it is optimal to be sub-optimal! This is fortunate, because the original optimization problems were intractable. In the end, we have a polynomial-time algorithm that does the right thing. By "undercomputing" we avoid "overfitting".

This interaction between computational issues (the complexity of various search problems) and statistical issues (the need to control overfitting) is one example of the important interplay between computer science and statistics. I expect that the coming decade will produce more such fruitful interactions.

References

[1] A. Blum and R. L. Rivest. Training a 3-node neural net is NP-Complete. In *Advances in Neural Information Processing Systems I*, pages 494–501. Morgan Kaufmann, 1989.

[2] Michael Kearns, Yishay Mansour, Andrew Y. Ng, and Dana Ron. An experimental and theoretical comparison of model selection methods. In *ACM Conference on Computational Learning Theory*. Morgan Kaufmann, San Francisco, CA, 1995.

[3] J. R. Quinlan and R. M. Cameron-Jones. Oversearching and layered search in empirical learning. In *Proceedings of IJCAI-95*. 1995.

[4] J. R. Quinlan and R. L. Rivest. Inferring decision trees using the Minimum Description Length Principle. *Inform. Comput.*, 80(3):227–248, March 1989. (An early version appeared as MIT LCS Technical report MIT/LCS/TM-339 (September 1987).).

[5] Andreas Weigend. On overfitting and the effective number of hidden units. In P. Smolensky, D. S. Touretzky, J. L. Elman, and A. S. Weigend, editors, *Proceedings of the 1993 Connectionist Models Summer School*, pages 335–342, Hillsdale, NJ, 1994. Lawrence Erlbaum Associates.