DOI:10.1145/1592761.1592782

Technical Perspective
Machine Learning for Complex Predictions

By John Shawe-Taylor

INTEREST IN MACHINE learning can be traced back to the early days of computer science. Alan Turing himself conjectured that some form of automatic learning would be required to endow a computer with artificial intelligence. The development of machine learning, however, has not always lived up to these ambitious beginnings.

The first flood of interest in machine learning was generated by Rosenblatt's perceptron.[4] The perceptron is a simple thresholded linear classifier that can be trained from an online sequence of examples. Intriguingly, from a theoretical point of view, the number of mistakes during training can be bounded in terms of intuitive geometric properties that measure a problem's difficulty.[3] One such property is the margin, or the distance of the most "borderline" example to the perceptron's decision boundary. First highlighted by the mistake bound in perceptrons, the margin would come to play an important role in the subsequent development of machine learning; however, its importance in Rosenblatt's time was not fully realized. Interest in perceptrons waned after a critical study by Minsky and Papert that catalogued their limitations.[2]

Machine learning developed in other directions during the 1970s–1980s with such innovations as decision trees, rule-learning methods for expert systems, and self-organizing maps. However, in the late 1980s, there was a resurgence of interest in architectures that used perceptrons as basic building blocks. In particular, the limitations highlighted by Minsky and Papert were overcome by multilayer networks in which perceptron-like nodes appeared as simple computing elements. These so-called neural networks were trained by a biologically inspired backpropagation algorithm[5] that performed gradient descent on error-based cost functions. This line of work not only raised hopes of creating machines that were able to learn, but also understanding the basic mechanisms behind biological learning. After a period of intense activity, however, history seemed to repeat itself with early enthusiasm overreaching actual accomplishments.

The study of support vector machines (SVMs) initiated a new approach to machine learning that focused more on statistical foundations[7] and less on biological plausibility. SVMs are trained by minimizing a convex cost function that measures the margin of correct classification. This is the same margin that measures the mistake complexity in perceptrons, but in SVMs the optimization is justified by a rigorous statistical analysis. This approach to binary classification has proven effective in a wide range of applications from text classification ("Is this article related to my search query?") to bioinformatics ("Do these microarray profiles indicate cancerous cells?"). Indeed, a 1999 paper on SVMs by Thorsten Joachims[1] recently won the field's award for having the most significant and lasting impact.

SVMs were originally formulated for problems in binary classification where the goal is simply to distinguish objects in two different categories. In the following paper, the authors significantly extend the applicability of this approach to machine learning. The paper considers problems that cannot easily be reduced to simple classification—that is, where the goal is to predict a more complex object than a single binary outcome. Such problems arise in many applications of machine learning, including search engine ranking and part-of-speech tagging.

For a long time, researchers attempted to solve such problems by using simple-minded reductions to binary classification problems. These reductions failed to exploit any information about the structure of predicted objects. For example, conventional SVMs can be applied to the problem of part-of-speech tagging by considering each word of a sentence based on the surrounding words. This approach ignores the fact that the tags of nearby words are mutually constraining—what is really needed is a self-consistent sequence of tags. The approach described here incorporates this type of structure through representations similar to those of probabilistic graphical models, such as Markov random fields.[6] Moreover, the models in this approach are trained by optimizing measures of discriminative performance that are of most interest to practitioners.

The authors describe a principled framework for predicting structured objects with SVMs. They show that the required optimization involves an exponentially large number of constraints in the problem size. The combinatorial explosion reflects the large number of possible misclassifications when predicting structured objects. Remarkably, the authors show that despite the apparently unmanageable number of constraints, an ϵ-accurate solution can be found by a cutting-plane algorithm that only considers $O(1/\epsilon)$ constraints. The algorithm is described here, along with examples of real-world applications in information retrieval and protein sequence alignment. The impressive variety of applications considered in the paper is yet another reminder that the scope of machine learning expands with each new generation of researchers. C

References

1. Joachims, T. Transductive inference for text classification using support vector machines. In *Proceedings of the 16th International Conference on Machine Learning* (Bled, Slovenia, 1999), 200–209.
2. Minsky, M.L. and Papert, S.A. *Perceptrons.* MIT Press, Cambridge, MA, 1969.
3. Novikoff, A.B. On convergence proofs on perceptrons. *Symposium on the Mathematical Theory of Automata.* Polytechnic Institute of Brooklyn (1962), 615–622.
4. Rosenblatt, F. The Perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review 65*, 6 (1958), 386–408.
5. Rumelhart, D.E., Hinton, G.E., and Williams, R.J. Learning representations by backpropagating errors. *Nature 323* (1986), 533–536.
6. Taskar, B., Guestrin, C., and Koller, D. Maxmargin markov networks. *Advances in Neural Information Processing Systems.* S. Thrun, L. Saul, and B. Schölkopf, Eds. MIT Press, Cambridge, MA, 2004, 25–32.
7. Vapnik, V. *Statistical Learning Theory.* John Wiley, 1998.

John Shawe-Taylor is a professor at University College London, where he is director of the Centre for Computational Statistics and Machine Learning.