Discussion of the Paper "Additive Logistic Regression: A Statistical View of Boosting" by Jerome Friedman, Trevor Hastie and Robert Tibshirani

Yoav Freund **Robert E. Schapire**
AT&T Labs – Research
180 Park Avenue
Florham Park, NJ 07932-0971 USA
{yoav, schapire}@research.att.com

January 28, 2000

The main and important contribution of this paper is in establishing a connection between boosting, a newcomer to the statistics scene, and additive models.

One of the main properties of boosting that has made it interesting to statisticians and others is its relative (but not complete) immunity to overfitting. As pointed out by the authors, the current paper does not address this issue. Leo Breiman [1] tried to explain this behaviour in terms of bias and variance. In our paper with Bartlett and Lee [4], we gave an explanation in terms of the "margins" of the training examples and the VC-dimension of the base class. Breiman, as well as the current paper, point out that our bounds are very rough and yield bounds that are not useful in practice. While this is clearly true at this time, it is also true that the analysis given by Breiman and by this paper yield no provable bounds whatsoever. It is completely unclear whether this analysis can be used to predict the performance of classification rules outside of the training sample.

At the root of this argument about boosting is a much more fundamental argument about the type of prior assumptions that one should make when embarking on the task of inducing a classification rule from data. The assumption that seems to underlie the use of maximum likelihood in the

current paper is that data are generated by a distribution from a *pre-specified class*. In this case, this is the class of distributions in which the relationship between the features and the labels is described by a log-linear function. In comparison, the assumption that we make in our analysis is that the data are generated from some *arbitrary* distribution in an i.i.d. fashion. Clearly, our assumption is the weaker one and this leads to a theory that is more generally applicable.

From a related but more practical point of view, one main issue when applying boosting or boosting-like techniques in practice is how to choose the base class. The approach taken in this paper is that this choice is made based on our prior beliefs regarding the type of log-linear dependencies that might exist between the features and the label. On the other hand, in the boosting approach, we make an assumption about what kind of rules might have slight but significant correlations with the label. This is the essence of the "weak learning" assumption upon which the theory of boosting is founded.

In the current paper, boosting is analyzed mostly in the context of decision stumps and decision trees. The argument seems to be that while in most real-world cases decision stumps are powerful enough, in some less common cases the type of dependencies that exist in the data require a more powerful base class, such as two- or three-level decision trees. A rather different approach to the combination of decision trees and boosting was recently proposed by Freund and Mason [3]. They represent decision trees as sums of very simple functions and use boosting to simultaneously learn both the decision rules and the way to average them.

Another important issue discussed in this paper is the performance of boosting methods on data which are generated by classes that have a significant overlap, in other words, classification problems in which even the Bayes optimal prediction rule has a significant error. It has been observed by several authors, including those of the current paper, that AdaBoost is not an optimal method in this case. The problem seems to be that AdaBoost over-emphasizes the atypical examples which eventually results in inferior rules. In the current paper, the authors suggest "GentleBoost" as a better method than AdaBoost for this case. The reason that this might be a better method is that it gives less emphasis to misclassified examples. The increase in the weight of the example is quadratic in the negative margin, rather than exponential.

However, one can argue that this alteration of AdaBoost, while being a step in the right direction, is not large enough. In fact, one can argue

that once an example has a very large negative margin it is best to assume that it is an outlier that should be completely removed from the training set. A new boosting algorithm based on this radical approach was recently proposed by Freund [2].

References

[1] Leo Breiman. Arcing classifiers. *The Annals of Statistics*, 26(3):801–849, 1998.

[2] Yoav Freund. An adaptive version of the boost by majority algorithm. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, 1999.

[3] Yoav Freund and Llew Mason. The alternating decision tree learning algorithm. In *Machine Learning: Proceedings of the Sixteenth International Conference*, pages 124–133, 1999.

[4] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Machine Learning: Proceedings of the Fourteenth International Conference*, pages 322–330, 1997. To appear, *The Annals of Statistics*.