Statistical Science
2007, Vol. 22, No. 4, 513–515
DOI: 10.1214/07-STS242A
Main article DOI: 10.1214/07-STS242

Comment: Boosting Algorithms: Regularization, Prediction and Model Fitting

Trevor Hastie

We congratulate the authors (hereafter BH) for an interesting take on the boosting technology, and for developing a modular computational environment in R for exploring their models. Their use of low-degree-of-freedom smoothing splines as a base learner provides an interesting approach to adaptive additive modeling. The notion of "Twin Boosting" is interesting as well; besides the adaptive lasso, we have seen the idea applied more directly for the lasso and Dantzig selector (James, Radchenko and Lv, 2007).

In this discussion we elaborate on the connections between L_2-boosting of a linear model and infinitesimal forward stagewise linear regression. We then take the authors to task on their definition of degrees of freedom.

1. L_2-BOOST AND INFINITESIMAL FORWARD STAGEWISE LINEAR REGRESSION

Motivated by a version of L_2-boosting in Chapter 10 of Hastie, Tibshirani and Friedman (2001), Efron, Hastie, Johnstone and Tibshirani (2004) proposed the LARS algorithm. The intent was to:

- develop a limiting version of L_2-boost in which the step-length ν went to zero;
- show that this limiting version gave paths identical to the lasso, as was hinted in that chapter.

The result was three very similar varieties of the LARS algorithm, namely lasso, LAR and infinitesimal forward stagewise (iFSLR) (package `lars` for R, available from CRAN). iFSLR is indeed the limit of L_2-boost as $\nu \downarrow 0$, with piecewise-linear coefficient profiles, but is not always the same as the lasso.

On a slight technical note, the version of L_2-boost proposed in BH is slightly different from that in Hastie, Tibshirani and Friedman (2001). Compare

$$\text{(1)} \qquad \text{[BH]} \quad \hat{\beta}^{[m]} = \hat{\beta}^{[m-1]} + \nu \cdot \hat{\beta}^{(\hat{\mathcal{S}}_m)},$$

$$\text{(2)} \qquad \text{[HTF]} \quad \hat{\beta}^{[m]} = \hat{\beta}^{[m-1]} + \nu \cdot \operatorname{sign}[\hat{\beta}^{(\hat{\mathcal{S}}_m)}].$$

Despite the difference, they both have the same limit, which is computed exactly for squared-error loss by the `type="forward.stagewise"` option in the package `lars`. As ν gets very small, initially the same coefficient tends to get continuously updated by infinitesimal amounts (hence linearly). Eventually a second variable ties with the first for coefficient updates, which they share in a balanced way while remaining tied. Then a third joins in, and so on. Using simple least-squares computations, the LARS algorithm computes the entire iFSLR path with the same cost as a single multiple-least-squares fit. Note that in this limiting case, we can no longer index the sequence by step-number m as in (1) or (2), but must resort to some other measure, such as the L_1-arc-length of the coefficient profile (Hastie, Taylor, Tibshirani and Walther, 2007).

Lasso and iFSLR are not always the same. In high-dimensional problems with correlated predictors, lasso profiles become wiggly quickly, whereas iFSLR profiles tend to be much smoother and monotone (Hastie et al., 2007). Efron et al. (2004) establish sufficient *positive cone conditions* on the model matrix X which effectively limit the amount of correlation between the variables and guarantee that lasso and iFSLR are the same; in particular, if the lasso profiles are monotone, all three algorithms are identical.

2. DEGREES OF FREEDOM

The authors propose a simple formula for the degrees of freedom for an L_2-boosted model. They construct the hat matrix $\mathcal{B}_m$ that computes the fit at iteration m, and then use $\mathrm{df}(m) = \mathrm{trace}(\mathcal{B}_m)$. They are in effect treating the model at stage m as if it were computed by a predetermined sequence of linear updates. If this were the case, their formula would be spot on, by the accepted definitions for effective degrees of freedom for linear operators (Hastie et al., 2001; Efron et al., 2004). They acknowledge that this is an approximation (since the sequence was not predetermined, but rather adaptively chosen), but do not

Trevor Hastie is Professor, Department of Statistics, Stanford University, Starford, California 94305, USA (e-mail: hastie@stanford.edu).