

Higher Criticism Thresholding: Optimal Feature Selection When Useful Features Are Rare and Weak

by

Jiashun Jin

Carnegie Mellon University

Department of Statistics

Carnegie Mellon University

Pittsburgh, PA, 15123 `jiashun@stat.cmu.edu`

Abstract

We consider a two-class linear classification in high-dimension low-sample size setting, where among a large number of features, only a small fraction of them is useful. The useful features are unknown to us, and each of them contributes weakly to the classification decision—we call this setting the rare/weak model (RW Model). We select features by thresholding feature z -scores. The threshold is set by the recent innovation of *higher criticism* (HC): Let π_i denote the p -value associated to the i -th z -score and $\pi_{(i)}$ denote the i -th order statistic of the collection of p -values, the HC threshold (HCT) is the order statistic of the z -score corresponding to index i which maximizes the ratio $(i/p - \pi_{(i)})/\sqrt{\pi_{(i)}(1 - \pi_{(i)})}$.

We formalize an asymptotic framework for studying the RW model, considering a sequence of problems with increasingly many features and relatively fewer observations. We show that along this sequence, the limiting performance of HCT is essentially just as good as the limiting performance of ideal thresholding—the optimal thresholding one would use when underlying parameters are known. HCT behaves very differently from other analytical principles popular today (e.g. False Discovery Rate control or Sure Screening). Also, HCT is dramatically faster and more stable than cross validation thresholding. Comparison to recent classification methods (including the Least Shrunken Centroids and False Discovery Rate Thresholding) will be drawn both with simulated data and real data in cancer classification.

This is joint work with David Donoho.