STATISTICS COLLOQUIUM
MONDAY, NOVEMBER 9, 2009
TALK: 4:00 PM — SCIENCE CENTER, ROOM 309
RECEPTION: 5:15 PM — SCIENCE CENTER, 7TH FLOOR

Higher Criticism Thresholding: Optimal Feature Selection when Useful Features are Rare and Weak

Jiashun Jin
Department of Statistics
Carnegie Mellon University

Motivated by many ambitious modern applications – genomics and proteomics are examples, we consider a two-class linear classification in high-dimensional, low-sample size setting (a.k.a. p n). We consider the case where among a large number of features (dimensions), 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 [2]). The success of linear classification hinges on how to select a small subset of useful features.

We select features by thresholding feature *z*-scores. The threshold is set by the recent innovation of *higher criticism* (HC) [1, 2]: 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 $\left(i/n - p_{(i)}\right) / \sqrt{p_{(i)}\left(1-p_{(i)}\right)}$. HCT has many interesting features as follows.

Asymptotic optimality in threshold selection. 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.

Optimal partition of the phase diagram. Our asymptotic analysis frames the notion of two-dimensional *phase space*, a two-dimensional diagram with coordinates quantifying “rare” and “weak” in the RW model. The phase space can be partitioned into two regions – one where ideal threshold classification is successful, and one where the features are so rare and so weak that it must fail. Surprisingly, the regions where HCT succeeds and fails partition the phase diagram in the exact same way. In comparison, many popular threshold choices (e.g. that by controlling the False Discover Rate) don't have the same partition of regions in the phase diagram, and are therefore suboptimal.

Outperforms popular threshold choice methods. We show that HCT behaves very differently from other analytical principles popular today (e.g. False Discovery Rate control or Sure Screening). We also show that 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.

References
[1] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. Ann. Statist. 32 962-994.
[2] Donoho, D. and Jin, J. (2008). Higher Criticism thresholding: optimal feature selection when useful features and rare and weak, Proc. Natl. Acad. Sci., 105 (39), 14790-14795.