Spectral Clustering with Higher Criticism Thresholding

Jiashun Jin
Department of Statistics
Carnegie Mellon University

Abstract

Consider a two-class clustering problem where we have (X_i, Y_i), $1 \leq i \leq n$, from two possible classes. The X_i's are $p \times 1$ vectors that are observable, and $Y_i \in \{-1, 1\}$ are class labels which are unknown to us and it is of interest to estimate them.

We propose the following approach to spectral clustering:

1. We use Kolmogorov-Smirnov statistic to assess the importance of the features (i.e., genes).
2. Based on the p-values, we perform a feature selection, where the threshold is determined by the recent idea of Higher Criticism Thresholding (HCT).
3. Based on all retained features, we obtain the leading eigenvector of the so-called *dual covariance matrix*, and predict the class labels by the signs of the coordinates of this eigenvector.

We reveal a surprising connection between the HCT and the so-called Signal Noise Ratio (SNR) associated with the post-screening dual covariance matrix. We apply the approach to two gene microarray data sets, where it gives much more satisfactory results than existing clustering methods.

The talk contains two main theoretic results. First, we derive a simple approximation for the tail probability Kolmogorov-Smirnov statistic assuming the underlying data is Gaussian using Loader's results on boundary crossing probabilities. Second, we show that HCT is consistent to the ideal threshold choice (the threshold one would choose if the underlying parameters are known to us) in the most challenging regime where signals are both rare and weak.