Permutation Tests for Classification

Sayan Mukherjee

Institute for Genome Sciences and Policy
Institute of Statistics and Decision Sciences
Duke University
Durham, NC 27708, USA

Permutation procedures have been used going back to when Student (W. S. Gosset) came up with the t-statistic while working at Guinness and have been used in both Bayesian and frequentist settings. We describe a permutation procedure used extensively in classification problems in computational biology and medical imaging. We empirically study the procedure on simulated data and real examples from neuroimaging studies and DNA microarray analysis. A theoretical analysis is also suggested to assess the asymptotic behavior of the test. A byproduct of the analysis is a uniform central limit theorem for permutation procedures. An interesting observation is that concentration of the permutation procedure is controlled by a Rademacher average which also controls the concentration of empirical errors to expected errors.
joint work with Polina Golland, Feng Liang, and Dmitry Panchenko