UCLA

BIOSTATISTICS SEMINAR

FALL 2016

NEYMAN-PEARSON (NP) CLASSIFICATION ALGORITHMS AND NP RECEIVER OPERATING CHARACTERISTIC (NP-ROC) CURVES

YANG FENG
DEPT OF STATISTICS
COLUMBIA UNIVERSITY

Wednesday, September 28 , 2016
3:30pm - 4:30pm, Dentristry 13-041
Refreshments served at 3:00 PM in room 51-254 CHS

ABSTRACT:

In many binary classification applications such as disease diagnosis and spam detection, practitioners often face great needs to control type I errors (e.g., chances of missing a malignant tumor) under some desired threshold. To address such needs, a natural framework is the Neyman-Pearson (NP) classification paradigm, which installs some upper bound α on population type I errors before minimizing population type II errors. However, common practices that directly control empirical type I errors under α do not satisfy the type I error control objective, as the resulting classifiers are likely to have population type I errors much larger than α. As a result, the NP paradigm has not been probably implemented for many classification scenarios in practice. In this work, we develop the first and general umbrella algorithm that implements the NP paradigm for popular classification methods, including logistic regression, support vector machines, and random forests. For this algorithm, we also suggest the minimum sample size required for controlling the population type I error with high probability. Powered by this umbrella algorithm, we propose a novel evaluation metric for the NP classification methods: the NP receiver operating characteristic (NP-ROC) curve, a variant of the popular receiver operating characteristic (ROC) curve. Despite its conceptual simplicity and wide applicability, the ROC curve lacks information on how to choose a classifier or compare different classifiers whose population type I errors are under some desired threshold with high probability. In contrast, the NP-ROC curve will serve as a new effective tool to evaluate, compare and select binary classifiers aiming for population type I error control. We demonstrate the use and properties of the NP umbrella algorithm and the NP-ROC curve, available in R package nproc, through simulation and real data case studies.