ACMS Statistics Seminar

Wesley Tansey
Columbia University
Tuesday, September 18
154 Hurley Hall
3:30– 4:30 PM

Holdout Randomization Tests: Easy and Principled Feature Selection For Black Box Models

A key scientific problem is sifting through many candidate features to find explanatory signals in data. Often, predictive models are used for this task: the model is fit, error on heldout data is measured, and strong performing models are assumed to have discovered some fundamental properties of the system under study. A heuristic method (e.g. tree membership counts in random forests or coefficient magnitudes in lasso models) is then used to rank important features, with top features reported as discoveries. However, such heuristics provide no statistical guarantees and can produce unreliable results. Here we propose the holdout randomization test (HRT) as a principled approach to feature selection. HRTs are model agnostic and produce valid p-values for each feature. Further, they require neither (potentially-costly) refitting of the predictive model nor hand-tuning of hyperparameters. This makes HRTs a natural drop-in replacement for many heuristic procedures commonly used in scientific analysis pipelines.