Speaker: Gregory Valiant, Stanford Computer Science

Title: Estimating Unexplained Variance and Learnability in the Sublinear Data Regime

Abstract:
We consider the problem of estimating how well a model class is capable of fitting a distribution of labeled data. We show that it is often possible to accurately estimate this “learnability” even when given an amount of data that is too small to reliably learn any model that fits the distribution. Our first result applies to the setting where the data is drawn from a $d$-dimensional distribution with isotropic covariance, and the label of each datapoint is a linear function of the data plus independent noise of unknown variance. In this setting, we show that the variance of the noise can be approximated to error $\epsilon$ given $O(\sqrt{d}/\epsilon)$ samples. Note that even if there is no noise, a sample size linear in the dimension, $d$, is required to learn any function correlated with the underlying linear model. We also obtain analogous results for several natural generalizations of this setting, including estimating the unexplained variance in the setting where the label may be an arbitrary (unknown) function of the point, and settings where the goal is to estimate the classification accuracy of the best linear/logistic model. Finally, we extend our estimation approach to the important setting where the data distribution has an (unknown) arbitrary covariance matrix. We will conclude with several experimental results, illustrating the practical viability of our approaches on synthetic and real data.

This is joint work with Weihao Kong.