STATISTICS COLLOQUIUM

TUESDAY, FEBRUARY 28, 2012
TALK: 4:30 P.M. — SCIENCE CENTER RM. 705

"Algorithmic Solutions to Some Statistical Questions"

Gregory Valiant
University of California, Berkeley

ABSTRACT

I will discuss three statistical problems for which the computational perspective offers new insights and provides more efficient approaches to coping with the increasing size of real-world datasets.

The first problem is the basic task of finding correlated vectors. Given a set of n binary vectors, with the promise that some pair is r-correlated but the rest are uncorrelated, can one find the correlated pair faster than simply searching all $O(n^2)$ pairs? We give an algorithm that finds such a pair in time poly $(1/r)*n^1.6$, which is the first sub-quadratic time algorithm for this task with a polynomial dependence on the correlation, and improves upon the prior best of $n^ (2-O(r))$, given by the Locality Sensitive Hashing approach. This result can be leveraged to yield significantly improved algorithms for several important problems in learning theory, including the problems of learning juntas, and learning parities with noise.

The second problem is recovering the parameters of a mixture of Gaussian distributions. Given data drawn from a single Gaussian distribution, the sample mean and covariance of the data trivially yield good estimates of the parameters of the true distribution; if, however, some of the data points are drawn according to one Gaussian, and the rest of the data points are drawn according to a different Gaussian, how can one recover the parameters of each Gaussian component? This problem was first proposed by Pearson in the 1890's, and, in the last decade, has been revisited by computer scientists. In a pair of papers with Adam Kalai and Ankur Moitra, we give the first polynomial-time algorithm for this problem that has provable success guarantees.

The final problem, investigated in a series of papers with Paul Valiant, considers the tasks of estimating a broad class of statistical properties, which includes entropy, various distance metrics between pairs of distributions, and support size. Our estimators are the first proposed estimators for these properties which achieve small error using a sub-linear number of samples, and are based on a novel approach of approximating the portion of the distribution from which one draws no samples. There are several implications of our results, including resolving the sample complexity of the 'distinct elements problem' (given a vector of length n, how many random indices must one query to accurately estimate the number of distinct elements?), and entropy estimation (given samples from a distribution of support n, how many samples are necessary to estimate the entropy to within an additive constant?); we show that for both problems, on the order of n/log n samples is both necessary and sufficient, improving upon the prior lower-bounds, and contradicting the previously widely-held belief that $O(n)$ samples are necessary.