Amherst College Department of Mathematics and Statistics

Connecticut Valley Mathematics and Statistics Colloquium

Rachel Ward

University of Texas at Austin

Mathematics for k-means clustering and beyond

Abstract: Under the hood of any modern algorithm for "big data" analysis is a step where the data is clustered into a smaller number of groups. The most widely-used clustering objective is the k-means algorithm, which aims to partition a set of n points into k clusters in such a way that each observation belongs to the cluster with the nearest mean, and such that the sum of squared distances from each point to its nearest mean is minimal. In general, this is a hard optimization problem, requiring an exhaustive search over all possible partitions of the data into k clusters in order to find the optimal clustering. At the same time, fast heuristic algorithms are often applied in many data processing applications, despite having few guarantees on the clusters they produce. In this talk, we will introduce new algorithms and techniques for solving the k-means optimization problem, based on semidefinite relaxation, and present geometric conditions on a set of data such that the algorithm is guaranteed to find the optimal k-means clustering for the data. The new algorithm has surprising connections to matrix factorization and manifold learning, and we conclude by discussing several open problems.

Friday, November 10, 2017

4:30 pm

206 Seeley Mudd Building

Refreshments — 4:00 pm
208 Seeley Mudd
RSVP by Friday, November 3, 2017
to awtorrey@amherst.edu

Dinner — 5:45 pm
30 Boltwood
Choice of limited menu
Reservations required