Geometric tools for high-dimensional data analysis

Ann Lee
Yale University

Thursday, February 17, 2005
L02 Carson Hall, 4:00 pm
(Tea 3:30 pm Math Lounge)

Abstract

In many applied fields—such as image analysis, information technology and biology—one has to analyze noisy, but structured data, in very high dimensions (> 1000 or even 10,000), often with a small number of samples. This "large d, small N" regime presents challenges for data analysis and calls for efficient dimension reduction tools that take the inherent geometry of natural data into account. In the first part of my talk, I will describe a multi-scale orthogonal basis that can be used for feature extraction of smooth data (such as images and spectral measurements) as well as non-smooth data (such as DNA micro arrays and word-document arrays). I will then, in the second half of the talk, describe a general methodology for organizing high-dimensional data sets by embedding the data into Euclidean space via a non-linear diffusion map. Examples will be taken from image analysis, word-document clustering and spectroscopy.

This talk should be accessible to graduate students.