THE UNIVERSITY   OF HONG KONG

Department of Mathematics

Numerical Mathematics and Applied Analysis Group Seminar (NMAA)

---

ccSVM: correcting Support Vector Machines for confounding factors in biological data classification

Dr. Limin LI
Xian Jiaotong University

on Friday, December 16, 2011 at 11:00am
in Room 309, Run Run Shaw Building, HKU

Abstract

Classifying biological data into different groups is a central task of bioinformatics: for instance, to predict the function of a gene or protein, the disease state of a patient or the phenotype of an individual based on its genotype. Support Vector Machines are a wide spread approach for classifying biological data, due to their high accuracy, their ability to deal with structured data such as strings, and the ease to integrate various types of data. However, it is unclear how to correct for confounding factors such as population structure, age or gender or experimental conditions in Support Vector Machine classification. In this talk, I will present a Support Vector Machine classifier that can correct the prediction for observed confounding factors. This is achieved by minimizing the statistical dependence between the classifier and the confounding factors. We prove that this formulation can be transformed into a standard Support Vector Machine with rescaled input data. In our experiments, our confounder correcting SVM (ccSVM) improves tumor diagnosis based on samples from different labs, tuberculosis diagnosis in patients of varying age, ethnicity and gender, and phenotype prediction in the presence of population structure and outperforms state-of-the-art methods in terms of prediction accuracy.

All are welcome