Challenges in Privacy-Preserving Data Analysis*

Kamalika Chaudhuri

Abstract

Machine learning algorithms increasingly work with sensitive information on individuals, and hence the problem of privacy-preserving data analysis – how to design data analysis algorithms that operate on the sensitive data of individuals while still guaranteeing the privacy of individuals in the data– has achieved great practical importance. In this talk, we address two problems in privacy-preserving data analysis.

First, we address the problem of privacy-preserving classification, and present an efficient classifier which is private in the differential privacy model of Dwork et al. Our classifier works in the ERM (empirical loss minimization) framework, and includes privacy preserving logistic regression and privacy preserving support vector machines.

We next address the question of learning from sensitive correlated data, such as private information on users connected together in a social network, and measurements of physical activity of a single user across time. Unfortunately differential privacy cannot adequately address privacy challenges in this kind of data, and as such, these challenges have been largely ignored by existing literature. We consider a recent generalization of differential privacy, called Pufferfish, that can be used to address privacy in correlated data, and present new privacy mechanisms in this framework. Based on joint work with Claire Monteleoni (George Washington University), Anand Sarwate (Rutgers), Yizhen Wang (UCSD) and Shuang Song (UCSD).

*Machine Learning External Seminar, Gatsby Unit, March 2, 2016.