Distributed Machine Learning: Communication, Efficiency, and Privacy

Avrim Blum, Carnegie Mellon University

Abstract

Suppose you want to learn a rule to distinguish positive and negative examples of some concept. But you have only the positive examples and a different research group has all the negative examples. How much communication do you need to learn a good classifier? In this talk we examine this question and its generalizations, as well as related issues such as privacy.

Broadly, we consider a framework where data is distributed among several locations, and our goal is to learn a low-error classifier over the joint distribution using as little communication, and as few rounds of communication, as possible. It turns out that in addition to VC-dimension and covering number, quantities such as the teaching-dimension and mistake-bound of a class play an important role in determining communication requirements. Moreover, boosting can be performed in a generic manner in the distributed setting to achieve communication with only logarithmic dependence on the final error rate for any concept class. We also present tight results for a number of common specific classes including conjunctions, parity functions, and decision lists, as well as non-tight results with intriguing open questions for the case of linear separators. We additionally present an analysis of privacy, considering both differential privacy and a notion of distributional privacy that is especially appealing in this context.

This is joint work with Maria-Florina Balcan, Shai Fine, and Yishay Mansour.