Theory and Algorithms for Modern Problems in Machine Learning and an Analysis of Markets

Author: Ashish Rastogi

Advisors: Richard Cole and Mehryar Mohri

Abstract

The unprecedented growth of the Internet over the past decade and of data collection, more generally, has given rise to vast quantities of digital information, ranging from web documents and images, genomic databases to a vast array of business customer information. Consequently, it is of growing importance to develop tools and models that enable us to better understand this data and to design data-driven algorithms that leverage this information. This thesis provides several fundamental theoretical and algorithmic results for tackling such problems with applications to speech recognition, image processing, natural language processing, computational biology and web-based algorithms.

- Probabilistic automata provide an efficient and compact way to model sequence-oriented data such as speech or web documents. Measuring the similarity of such automata provides a way of comparing the objects they model, and is an essential first step in organizing this type of data. We present algorithmic and hardness results for computing various discrepancies (or dissimilarities) between probabilistic automata, including the
relative entropy and the $L_p$ distance; we also give an efficient algorithm to
determine if two probabilistic automata are equivalent. In addition, we
study the complexity of computing the norms of probabilistic automata.

- Widespread success of search engines and information retrieval systems
has led to large scale collection of rating information which is being used
to provide personalized rankings. We examine an alternate formulation
of the ranking problem for search engines motivated by the requirement
that in addition to accurately predicting pairwise ordering, ranking sys-
tems must also preserve the magnitude of the preferences or the dif-
terence between ratings. We present algorithms with sound theoretical
properties, and verify their efficacy through experiments.

- Organizing and querying large amounts of digitized data such as images
and videos is a challenging task because little or no label information is
available. This motivates transduction, a setting in which the learning
algorithm can leverage unlabeled data during training to improve per-
formance. We present novel error bounds for a family of transductive
regression algorithms and validate their usefulness through experiments.

- Finally, price discovery in a market setting can be viewed as an (ongoing)
learning problem. Specifically, the problem is to find and maintain a set
of prices that balance supply and demand, a core topic in economics.
This appears to involve complex implicit and possibly large-scale infor-
mation transfers. We show that finding equilibrium prices, even approx-
imately, in discrete markets is NP-hard and complement the hardness result with a matching polynomial time approximation algorithm. We also give a new way of measuring the quality of an approximation to equilibrium prices that is based on a natural aggregation of the dissatisfaction of individual market participants.