Automata Learning and its Applications

Dana Ron

Thesis Supervisor: Professor Naftali Tishby

Abstract

This thesis is a study of automata learning. Most of the work presented here is in the framework of Computational Learning Theory and hence emphasizes the theoretical aspects of learning algorithms and their rigorous analysis. However, a substantial part of this work was directly motivated by practical applications in human-machine interactions such as the statistical modeling of natural languages and handwriting. Thus, in addition to the formal description and analysis of learning algorithms for both deterministic and probabilistic automata, several applications of these algorithms are given. The thesis consists of two parts: one on learning deterministic automata and one on learning probabilistic automata, where the latter are automata that generate distributions. Since there are severe limitations on our ability to learn efficiently both deterministic and probabilistic automata, a common thread that passes through these two parts is the search for natural subclasses of automata that can be learned efficiently. In addition we emphasize the importance of learning algorithms that work in a *noisy* environment in which the observations of the learner might be corrupted by some form of random noise.

We begin the first part by describing efficient algorithms for learning deterministic finite automata, where our approach is primarily distinguished by two features: (1) the adoption of an average-case setting to model the "typical" labeling of a finite automaton, while retaining a worst-case model for the underlying graph of the automaton, along with (2) a *passive* learning model in which the learner is not provided with the means to experiment with the machine, but rather must learn solely by observing the automaton's output behavior on a random input sequence. The main contribution of this work is in presenting the first efficient algorithms for learning non-trivial classes of automata in an entirely passive learning model.

We continue with an efficient algorithm for *actively* learning an environment that can be described by a deterministic automaton whose underlying graph has certain topological properties. The learner may perform a single walk of its choice on the automaton and observe the outputs of the states passed. It must then be able to predict correctly the output sequence corresponding to *any* future walk. This work was partly motivated by a game theoretical problem of playing repeated games against a computationally bounded

adversary. We also show that a variant of this algorithm is robust to random noise. Previous work in this model either assumed access to a very powerful oracle, or made other limiting assumptions on the target automaton and the information that the learner is given.

We close the first part with a description of an algorithm for learning automata under the uniform distribution from a *fallible expert*. The expert answers all queries on inputs of our choice, but often gives incorrect answers. The choice of queries which are answered incorrectly, is random, but the expert is persistent, i.e., if queried more than once on the same input, it will always return the same answer. The learning algorithm constructs a hypothesis that not only concisely holds the correct knowledge of the expert, but actually surpasses the expert by using the structural properties of the concept in order to correct the expert's errors.

In the second part of this thesis, we give evidence to the hardness of learning general probabilistic automata. We show that under an assumption about the difficulty of learning parity functions with classification noise in the PAC model (a problem closely related to the longstanding coding theory problem of decoding a random linear code), the class of distributions defined by probabilistic finite automata is not efficiently learnable. However, we describe efficient algorithms for learning two subclasses of probabilistic automata: *variable memory* probabilistic automata, and *acyclic state-distinguishable* probabilistic automata. Research on both subclasses was directly motivated by practical applications of modeling various natural sequences, and we present the following applications of the two respective learning algorithms.

We applied the algorithm for learning variable memory probabilistic automata in order to construct: (1) a model of the English language which we used to correct corrupted text; (2) a simple probabilistic model for E.*coli* DNA. We applied the algorithm for learning acyclic probabilistic automata in order to construct: (1) models for cursively written letters which were used for segmenting labeled data; (2) multiple-pronunciation models for spoken words.

Moreover, the two subclasses we consider (and their respective learning algorithms) complement each other. Whereas the class of variable memory probabilistic automata captures the long range, stationary, statistical properties of the natural source, the class of acyclic probabilistic automata captures the short sequence statistics. Together, these algorithm constitute a powerful language modeling scheme. This scheme was applied to cursive handwriting recognition and may be applicable to similar problems such as speech recognition.