Thesis: Learning Finite-State Machines: Algorithmic and Statistical Aspects
Author: Borja de Balle Pigem
Supervisors: Jorge Castro Rabal and Ricard Gavaldà Mestre
Institution: Universitat Politècnica de Catalunya (Barcelona, Spain)
Defense Date: July 12, 2013

Abstract

Data in the form of sequences of dependent observations are often encountered in applied domains. Examples include language and speech processing, computational biology, time series analysis, and sequential decision-making problems. In all those scenarios, being able to extract models from existing datasets is a key step towards leveraging data into knowledge and informed decisions. In order to build these models, a comprehensive set of analytic tools capable of working with sequential data in a transparent and straightfoward manner is needed.

Discriminative tasks with sequential data can already be solved using standard machine learning tools empowered with feature representations and similarity measures for sequences. On the other hand, learning generative models typically requires using ad-hoc models for each different type of data. The main purpose of this thesis is to study several machine learning problems about generative models for sequential data. Our presentation emphasizes several theoretical aspects of these problems, in particular: the constraints imposed by the choice of a data access model, giving algorithms whose running time is polynomial in all the relevant parameters, and proving statistical bounds on the accuracy of learned models. In some cases we are also able to extend our techniques from generative to discriminative models, in which case one obtains discriminative algorithms working directly with sequential representations of the data.

The thesis is divided into two parts according to the different techniques used for deriving learning algorithms. The models considered in both parts have in common that they can be defined in terms of probabilistic finite-state machines. Most of these models could also be described in terms of latent variable models satisfying some markovian assumption. However, instead of using the graphical model perspective, we choose to present our results from the point of view of automata theory. This has several advantages, among which the most salient is perhaps that it gives a common theme to the two parts of the thesis. Using a common automata-theoretic framework helps us unveal some connections between the two approaches we consider which might have otherwise been blurred.

The first part of the thesis studies algorithms for learning the probabilistic analog of Deterministic Finite Automata (DFA) and several generalizations of them. This class of machines provide fairly expressive generative models for sequences, which in addition have very interesting algorithmic properties. Algorithms for learning machines with deterministic state transitions rely on inductively finding a correct transition structure by either merging or splitting one state at a time. These state-merging algorithms can be interpreted as a divisive clustering scheme where the “dependency graph” between clusters is not necessarily a tree.

One of our contributions in this area is to give a characterization of these algorithms in terms of statistical queries. The key ingredient of such algorithms is a statistical test to determine whether two states generate the same distribution. The complexity of this testing problem is usually measured by the distinguishability of the target machine. Using

our characterization, we can prove the first lower bound on the complexity of state-merging algorithms with an explicit dependency on the distinguishability of the target.

Moving away from the statistical query setting and into a more realistic scenario, we consider an on-line setting where examples are presented to the algorithm one at a time. We give a state-merging algorithm for this problem satisfying the stringent algorithmic constraints of the data streams computing paradigm. Our algorithm comes with strict PAC learning guarantees, and is also able to detect and adapt to changes in the nature of the data over time.

A whole chapter is devoted to a systematic study of the statistical tests for distribution similarity lying at the heart of state-merging algorithms. In particular, we give ways to improve the memory required to implement such tests. We also look into the problem of obtaining practically efficient tests with finite-sample guarantees based on boostrapped confidence intervals. This tests yield much faster convergence in many situations, and can be combined with a sketching technique to reduce memory usage and test accurately in on-line situations.

The last chapter of this part studies a wider class of models to which the state-merging paradigm can be extended. Essentially, these are generalizations of previous models and algorithms to the case where each symbol in the sequence can be decomposed into two parts. Applications of this method include continuous-time Markovian models and stochastic transducers on pairs of aligned sequences. Our techniques give PAC learning guarantees for all these models, and provide a general recipe for obtaining new learning algorithms for other generalized models provided two basic primitives (similarity testing and density estimation) are available.

Results presented in this part of the thesis have been published in conferences [1, 2, 3] and journals [4, 5]. Some of the tools used for obtaining these results are concentration inequalities and sketching algorithms.

The second part of the thesis presents our contributions to the rapidly growing body of spectral learning algorithms. In automata-theoretic terms, these algorithms allow us to learn probabilistic machines with non-deterministic transitions – as opposed to the models with deterministic transitions considered in the first part. The traditional solution for learning these models relies on some instantation of the Expectation-Maximization (EM) algorithm, which is known to converge only to a local maximum of the sample likelihood function. In contrast, under some simple separation assumption, spectral methods yield consistent learning algorithms for a large class of models with non-deterministic transitions. Some virtues of this type of algorithms include the possibility of proving finite-sample error bounds, and enormous savings on computation time with respect to iterative methods like EM.

In this thesis we give the first application of this method for learning conditional distributions over pairs of aligned sequences defined by probabilistic finite-state transducers with finite-sample bounds. This constitutes a non-trivial extension of previous spectral algorithms for learning hidden Markov models to a case with input and output sequences. We also extend previous proofs to show that the spectral algorithm can learn the class of all probabilistic automata. This result extends the class of models for which learning algorithms with strong finite-sample bounds can be obtained using these techniques.

The last two chapters present works combining spectral learning with methods from convex optimization and matrix completion. Respectively, these yield an alternative interpretation of spectral learning and an extension to cases with missing data. The former also opens the door to new versions of the algorithm that incorporate arbitrary convex regularizers representing prior knowledge about the target machine. In the latter case we used a novel joint stability analysis of matrix completion and spectral learning to prove the first generalization bound for spectral algorithms that holds in the non-realizable case.

Some of the results presented in this part of the thesis have appeared in the conference papers [6, 7, 8, 9] and the journal paper [10]. Our work in this area has been motivated by connections between spectral learning, classic automata theory, and statistical learning; tools from these three areas are extensively used in the proofs.

Distinctions

Publications derived from this thesis have received the following honors and awards:

- Best Paper Award for [7] at EACL 2012
- Best Student Paper Award for [3] at ICGI 2012
- Honorable Mention for [9] on the Outstanding Student Paper Awards at NIPS 2012
- Paper [9] selected for oral presentation at NIPS 2012 (5.4% of accepted papers)

Funding

The author received competitive funding from the following sources during the course of the research presented in this thesis:

- FI Doctoral Fellowship, Catalan Government
  (January 2009 – July 2009)
- FPU Doctoral Fellowship, Spanish Ministry of Education
  (August 2009 – January 2013)
- FPU Travel Award (for visiting NYU), Spanish Ministry of Education
  (December 2011 – June 2012)

References

[1] B. Balle, J. Castro, and R. Gavaldà. Learning PDFA with Asynchronous Transitions. *International Colloquium on Grammatical Inference (ICGI)*, 2010.

[2] B. Balle, J. Castro, and R. Gavaldà. A Lower Bound for Learning Distributions Generated by Probabilistic Automata. *Algorithmic Learning Theory (ALT)*, 2010.

[3] B. Balle, J. Castro, and R. Gavaldà. Bootstrapping and Learning PDFA in Data Streams. *International Colloquium on Grammatical Inference (ICGI)*, 2012.

[4] B. Balle, J. Castro, and R. Gavaldà. Learning Probabilistic Automata: A Study In State Distinguishability. *Theoretical Computer Science*, 473:46–60, 2013.

[5] B. Balle, J. Castro, and R. Gavaldà. Adaptively Learning Probabilistic Deterministic Automata from Data Streams. *Machine Learning (DOI: 10.1007/s10994-013-5408-x)*, 2013.

[6] B. Balle, A. Quattoni, and X. Carreras. A Spectral Learning Algorithm for Finite State Transducers. *European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD)*, 2011.

[7] F. M. Luque, A. Quattoni, B. Balle, and X. Carreras. Spectral Learning for Non-Deterministic Dependency Parsing. *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2012.

[8] B. Balle, A. Quattoni, and X. Carreras. Local loss optimization in operator models: A new insight into spectral learning. *International Conference on Machine Learning (ICML)*, 2012.

[9] B. Balle and M. Mohri. Spectral Learning of General Weighted Automata via Constrained Matrix Completion. *Neural Information Processing Systems Conference (NIPS)*, 2012.

[10] B. Balle, X. Carreras F. M. Luque, and A. Quattoni. Spectral Learning of Weighted Automata: A Forward-Backward Perspective. *Machine Learning (DOI: 10.1007/s10994-013-5416-x)*, 2013.