GRAMMATICAL INFERENCE

The problem of inducing, learning or inferring grammars has been studied for decades, but only in recent years has grammatical inference emerged as an independent field with connections to many scientific disciplines, including bio-informatics, computational linguistics and pattern recognition. This book meets the need for a comprehensive and unified summary of the basic techniques and results, suitable for researchers working in these various areas.

In Part I, the objects of use for grammatical inference are studied in detail: strings and their topology, automata and grammars, whether probabilistic or not. Part II carefully explores the main questions in the field: what does learning mean? How can we associate complexity theory with learning? In Part III the author describes a number of techniques and algorithms that allow us to learn from text, from an informant, or through interaction with the environment. These concern automata, grammars, rewriting systems, pattern languages and transducers.

COLIN DE LA HIGUERA is Professor of Computer Science at the University of Nantes.

GRAMMATICAL INFERENCE

Learning Automata and Grammars

Colin de la Higuera
Université de Nantes

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore,
São Paulo, Delhi, Dubai, Tokyo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521763165

First published 2010

Printed in the United Kingdom at the University Press, Cambridge

A catalogue record for this publication is available from the British Library

ISBN 978-0-521-76316-5 Hardback

Contents

Preface

> Young men should prove theorems, old men should write books.
>
> ***Godfrey H. Hardy***

> There is nothing to writing. All you do is sit down at a typewriter and bleed.
>
> ***Ernest Hemingway***

A zillion grammatical inference years ago, some researchers in grammatical inference thought of writing a book about their favourite topic. If there was no agreement about the notations, the important algorithms, the central theorems or the fact that the chapter about learning from text had to come before or after the one dealing with learning from an informant, there were no protests when the intended title was proposed: *the art of inferring grammars*. The choice of the word *art* is meaningful: like in other areas of machine learning, what counted were the ideas, the fact that one was able to do something complicated like actually building an automaton from strings, and that it somehow fitted the intuition that biology and images (some typical examples) could be explained through language. This 'artistic' book was never written, and since then the field has matured.

When writing this book, I hoped to contribute to the idea that the field of grammatical inference has now established itself as a scientific area of research. But I also felt I would be happy if the reader could grasp those appealing *artistic* aspects of the field.

The artistic essence of grammatical inference was not the only problem needing to be tackled; other questions also required answers...

Why not call this book 'grammar induction'?

When one wants to search for the different published material about the topic of this book, one finds it associated with many different fields, and more surprisingly under a number of names, such as 'grammar learning', 'automata inference', 'grammar identification', but principally 'grammar induction' and 'grammatical inference'. Even if this is not formalised anywhere, I believe that 'grammar induction' is about finding a grammar that can explain the data, whereas grammatical inference relies on the fact that there is a (true or only

possible) target grammar, and that the quality of the process has to be measured relatively to this target.

Practically this may seem to make little difference, as in both cases what probably will happen is that a set of strings will be given to an algorithm, and a grammar will have to be produced. But whereas in grammar induction this is the actual task, in grammatical inference this is still a goal but more a way of measuring the quality of the learning method, the algorithm or the learning setting.

In other words, in the case of grammar induction what really matters is the data and the relationship between the data and the induced grammar, whereas in grammatical inference the actual learning process is what is central and is being examined and measured, not just the result of the process.

Is this about learning languages or learning grammars?

Even if historically the task has been associated with that of learning languages (and typically with that of children acquiring their first language), we will concentrate on learning by a machine, and therefore (as in any computational task) a representation of the languages will be essential. Given this first point, a simple study of formal language theory shows us that not all representations of languages are equivalent. This will justify the choice of presenting the results with a specific representation (or grammar) system in mind.

Why use terms from formal language theory, like finite automata, when you could use more generic terms like finite state machines or alternative terms like hidden Markov models?

Actually, there is a long list of terms referring to the same sort of object: one also finds Mealy and Moore machines and (weighted) finite state transducers. In some cases these terms have a flavour close to the applications they are intended for; in others the names are inheritances of theoretical research fields. Our choice is justified by the fact that there are always some computable transformations allowing us to transform these objects into deterministic or non-deterministic finite automata, with or without probabilities. In the special case of the transducers, separate theory also exists.

By defining everything in terms of automata theory, we aim to use a formalism that is solidly established and sufficiently general for researchers to be able to adapt the results and techniques presented here to the alternative theories.

Why not introduce the definitions just before using them?

A generally good idea is to only introduce the definitions one needs at the moment one needs them. But I decided against this in certain cases; indeed, the theories underlying the

algorithms and techniques of grammatical inference deserve, at least in certain cases, to be presented independently:

- Chapter 3 introduces concepts related to strings, and also to distances or kernels over strings. If the former are well known by formal language theory specialists, this is not always the case with researchers from other fields interested in grammatical inference. It seemed interesting to have these definitions in one separate chapter, in order for these definitions to be compared and depend on each other.
- Most of the material presented in Chapter 4 about grammars and automata can be found in well-known textbooks, even if providing uniform notations is not customary. When considering probabilistic finite state automata, this is even worse as the definitions are far from trivial and the theory is not straightforward. We have also chosen to give here the main algorithms needed to deal with these objects: parsing, estimating, distances, etc.
- The reader just looking for an algorithm might go directly to the chapter corresponding to the class of grammar or the type of presentation he is dealing with. But to understand the convergence issues associated with the algorithm he wants to use, he might want to read Chapter 7 about learning models, and more specifically be interested in the inclusion of complexity questions in these models.
- The panorama concerning probabilities and grammatical inference would have been very difficult to understand if we had not devoted Chapter 10 to this issue: whether we use the distributions to measure classification errors or to actually learn the grammars that generate them, we try in this chapter to give a uniform view of the field, of its definitions and of its difficulties.
- For very different reasons, Chapter 6, about combinatorics, groups a number of more or less well-known results concerning automata, deterministic or non-deterministic, probabilistic or not. These are usually hardness results and rely on special constructions that we give in this separate chapter.

Did the author invent all these things?

No, of course not. Many scientists have contributed the key ideas and algorithms in this book. Some are clearly identifiable but others are not, the idea having 'been around for a while' or being the synthesis of thoughts of several people. Moreover, in such a field as this one, where papers have been written in a great diversity of areas and where up to now no text book had been produced, it is difficult, if not impossible, to know who has the paternity of what idea. I have tried to *render unto Caesar the things which are Caesar's*, and have consulted widely in order to find the correct authorships, but there are probably going to be researchers who will feel that I have wrongfully attributed their result to someone else. This I hope to have avoided as much as possible and I wish to express my apologies for not having found the correct sources. A specific choice to increase readability has been to leave the citations outside the main text. These are discussed at the end of each chapter.

Is the reader right to say that he or she has seen these proofs before?

Yes, of course he or she is right. In some rare cases, our knowledge today of the objects and the problems allows us to propose alternative proofs that somehow fitted in better. In other

cases (as duly acknowledged as possible), the original proof seemed unbettered. And for many reasons, most important of which is usually just the fact that the proof does add that useful information, I chose to include them, with just the necessary changes of notation.

These do not seem to be the standard notations. Why didn't the author use the standard notations?

In such a new field as grammatical inference there is no such thing as standard notations. Grammatical inference has borrowed notations from many fields: machine learning, formal language theory, pattern recognition and computational linguistics. These are even, in many cases, conflicting. Moreover the authors of the many papers all have different backgrounds and this diversity is reflected in the variety of notations and terms that are used. Choices had to be made for the book to be readable. I have used notations that could adapt smoothly to the different grammatical inference settings reported in this book. A chief goal has been to make algorithmic ideas from one case reusable in another. It is nevertheless fair to say that this has a price (for example where introducing automata with two types of final states) and specialists, used to working with their own notations, may be inclined to disagree with these proposals.

Is this the right moment to write a book on grammatical inference?

That is what I believe. It certainly is not too early, since most of the key questions were asked by Noam Chomsky or Ray Solomonoff 50 years ago! Mark Gold's results date back another 40 years, and Dana Angluin's first contributions to the field correspond to work done 30 years ago. The series of conferences known as ICGI have now been running for more than 15 years. The theme today is present in a number of conferences and the main algorithms are used in many fields. A noticeable thing is that the field is broadening: more complex classes of grammars are being learnt, new learning paradigms are being explored and there are many new applications each year. But no basic books with the main definitions, theorems and algorithms exist. This is an attempt to help register some of the main results (up to now) to avoid them being rediscovered over and over.

Why exercises?

Even if this is not a text book, since grammatical inference is not being taught in a curriculum, there are lectures in machine learning, text mining, and other courses where formal languages and generalisation are taught; there have also been attempts to hold lectures on the topic in summer schools and doctoral schools; moreover, I have received sufficient encouragement to initiate a list of exercises in order to be able to teach in the near future. The exercises are intended for this purpose. They are also here to give a flavour of a few of the interesting questions in the field.

What about trees and graphs?

The original goal was to cover extensively the field of grammatical inference. This of course meant discussing in detail tree automata and grammars, giving the main adaptation of classical string algorithms to the case of trees, and even dealing with those works specific to trees. As work progressed it became clear that learning tree automata and grammars was going to involve at least as much material as with strings. The conclusion was reached to only sketch the specificities here, leaving the matter largely untouched, with everything to be written. This of course is not justified by the importance of the question, but only by the editorial difficulty and the necessity to stop somewhere. Of course, after trees will come the question of graphs...

Is this the final book?

No, yet even more preposterously, I hope it to be an initial book. One that allows fellow researchers to want to write their positions in new books expressing the variety of points of view of a community made up of colleagues with such different interests, whether in machine learning, computational biology, statistics, linguistics, speech recognition, web applications, algorithmics, formal language theory, pattern analysis...

Acknowledgements

It all started in July 2005, in a meeting with David Tranah who easily convinced me that I really wanted to write a book on grammatical inference. During the next few years he was to encourage me in many ways and somehow remind me that this was something I really wanted to do.

During the years it took to write it, I had an unexpected (by me) number of problems I would not have been able to solve without the expertise of Thierry Murgue, who helped me solve a number of technical questions, but also provided his knowledge about probabilistic finite state machines.

A lot of what I know of the topic of grammatical inference is due to the patient efforts of Laurent Miclet and Jose Oncina, with whom, over the years, I have interacted and learnt.

On various occasions I have had to teach these things or organise workshops related to these topics, and it has always been a pleasure to prepare these with colleagues like Tim Oates, Pieter Adriaans, Henning Fernau, Menno van Zaanen and the aforementioned.

Most of the material I am presenting here comes either from work done by others, or, when I have been actually involved, from work done with students and a number of collaborators. Let me thank Leo Becerra Bonache, Rafael Carrasco, Francisco Casacuberta, Pierre Dupont, Rémi Eyraud, Jean-Christophe Janodet, Luisa Micó, Frédéric Tantini, Franck Thollard, Enrique Vidal and a number of students including Cristina Bibire, Anuchit Jittpattanakul and Émilie Samuel.

I am grateful also to several of the people thanked above for reading, with a lot of care, the different chapters in many of the preliminary versions I prepared. But I also received help in this from Hasan Akram, Stefan Gulan, François Jacquenet, Anna Kasprzik, Satoshi Kobayachi, Etsuji Tomita and Takashi Yokomori. Of course, I take entire blame for all the remaining errors.

Most of the work was done while I held a position at Saint-Étienne University, first in the EURISE team and later in the Laboratoire Hubert Curien. The conditions were great and I am pleased to use this occasion to thank all my colleagues there.

When one has not written a book, the customary acknowledgement to the author's spouse and children seems meaningless. When one has spent endless evenings and weekends being between entirely absent and half-absent from the family activities and spirit, it reaches its full meaning. Lindsey, Boris and Vikki, all my thanks.