Preface

All the interests of my reason, speculative as well as practical, combine in the three following questions: 1. What can I know? 2. What ought I to do? 3. What may I hope?

(Immanuel Kant, The Critique of Pure Reason)

Dear Reader,

my interest in what is now the topic of this book has begun during my studies at the University of Bristol, continued through my PhD in Bonn, and still lasts. Kernel for structured data are of interest to researchers in machine learning and data mining but also to practitioners in many other fields applying state-of-the-art machine learning techniques. Machine learning and data mining are two research fields concerned with automated learning that have demonstrated considerable success in a wide variety of applications and hence became key enabling technologies in many areas.

Kernel methods are currently one of the most popular class of machine learning algorithms. The best known kernel method is the support vector machine. Kernel methods are distinguished by their theoretically sound foundation in learning theory and their outstanding empirical results. These have first been achieved on domains where the objects of learning can easily be embedded in a Euclidean space.

Real-world machine learning problems, however, are often such that the objects that we want to learn about have no natural representation in attribute-value form. An example of such a problem is to estimate the activity of chemical compounds against an illness. In this application one very natural representation of such chemical compounds is their chemical structure graph. Kernel methods and other conventional machine learning

algorithms can not directly be applied to this sort of problems.

This book is concerned with the extension of kernel methods to structured data. In particular we consider two different representation languages for structured data: logic and graphs. As a logic based representation we chose to use the basic terms of a typed higher-order logic. For graphs we will consider directed and undirected labelled graphs. The distinction between these representation languages is beneficial for the clarity of the description of the kernel function as well as from a computational perspective. Both formalisms, as we will see, lend themselves naturally to certain kinds of application domains. Together they cover most—if not all—kinds of structured data that might occur in real-world applications.

The natural way to extend kernel methods to logical and graph-based representations of structured data is to define a positive definite kernel function on the set of possible object representations. This book presents a systematic approach to define kernel functions for structured data and to apply these kernel functions to large scale real-world machine learning problems. We define and characterise suitable kernel functions for structured data and their computational properties. Our empirical evaluation shows that kernel methods with our kernel functions for structured data substantially outperform conventional methods on a variety of important application domains.

Life is just a long random walk

(Devroye, Györfi, and Lugosi. A Probabilistic Theory of Pattern Recognition)

Many thanks go to Peter Flach for introducing me to the world of machine learning and Colin Campbell for teaching me about support vector machines. Since then, Peter has helped me in an uncountable number of ways—Thanks!

Back to Germany I started my PhD thesis at the University of Bonn and at Fraunhofer IAIS which at the time was called GMD AIS. It goes without saying that my thesis supervisor, Stefan Wrobel, has had the largest impact on my thesis and thus this book. Thanks for your advice, support, help, comments, pleasant conversations, etc. I am also very grateful to the remaining members of my thesis committee and reviewers: Michael Clausen, Stefan Kramer, Wolfgang Alt, and Reinhard Klein. In particular, I want to thank Michael Clausen who read this thesis very carefully and

gave many helpful comments.

Of course, fruitful research is only possible in friendly and pleasant working environments. Michael—thanks for providing such an environment. Tamás—thanks for your sense of humour and our scientific discussions. Hendrik and Francois—thanks for all our discussions and all the fun we had with many red wines and/or white beers. Myriam—thanks for all your help. Though I could easily continue this list of colleagues, I want to keep things short and just thank everybody who works or worked at IAIS (inside and outside the knowledge discovery group) who made my work time more pleasant and easier—Thanks!

One of the parts of my PhD I enjoyed most is travelling. Apart from meeting great people at various conferences, I had the honour of visiting different research labs for collaborations which always left me with some unforgettable experiences. Thanks to everybody in the Freiburg and Leuven machine learning groups—meeting you is always stimulating as well as a lot of fun (and extra thanks for Fluffy). During these travels and various conferences I met a number of people whom I had the pleasure of writing papers with that strongly influenced my thesis. Thanks go to Adam, Alex, Jan, John, Kristian, and Kurt. Just after submitting my thesis I had the opportunity of visiting the statistical machine learning program at the National ICT Australia. Thanks to everybody there for providing a great research environment. Since then, I have visited more groups—thanks to everyone in these groups—and co-authored more paper—thanks to Quoc, Vishy, Ulf, Tobias, Yasemin, Gemma, Hanna, and Jürgen. During the travel free time I enjoy most to work with the people that make up the CAML group: Gihad, Karim, Lana, Mario, and Shankar—Thanks!

The people most important to me must of course also be mentioned: My family and friends. The first and probably most significant influence to this book came, as might be expected, from my parents and sisters—Thanks! Throughout my life (scientific as well as real) I had the pleasure of meeting many more people than I can mention here from all over the world whom I now consider my friends. You know who you are—Thanks!

Thomas Gärtner 18th June 2008