LEARNING TO CLASSIFY TEXT USING SUPPORT VECTOR MACHINES

THE KLUWER INTERNATIONAL SERIES IN ENGINEERING AND COMPUTER SCIENCE

LEARNING TO CLASSIFY TEXT USING SUPPORT VECTOR MACHINES

by

Thorsten Joachims
Cornell University, U.S.A.

Dissertation, Universität Dortmund
Fachbereich Informatik
February 2001

SPRINGER SCIENCE+BUSINESS MEDIA, LLC

Library of Congress Cataloging-in-Publication Data

Joachims, Thorsten.
Learning to classify text using support vector machines: methods, theory, and algorithms / by Thorsten Joachims.
p. cm. – (Kluwer international series in engineering and computer science; SECS 668)
Originally presented as the author's thesis (Universität Dortmund) under the title: "The maximum-margin approach to learning text classifiers—methods, theory, and algorithms."
Includes bibliographical references and index.
ISBN 978-1-4613-5298-3 ISBN 978-1-4615-0907-3 (eBook)
DOI 10.1007/978-1-4615-0907-3
1. Text processing (Computer science) 2. Machine learning. I. Title. II. Series.

QA76.9.T48 J63 2002
005—dc21

2002022127

Originally published by Kluwer Academic Publishers in 2002
Softcover reprint of the hardcover 1st edition 2002

Printed on acid-free paper.

Contents

Foreword

"A good theory yields the best practice" – this saying characterizes the work in this book by Thorsten Joachims. Practitioners need to understand whether and how well their problem can be solved, and which approaches are likely to be most effective. This book reports on machine learning research that has produced both strong experimental results and a theory of text classification that can inform the practitioner interested in applying similar learning methods.

The problem addressed in this book is that of text classification: automatically classifying text documents based on their content. The appearance of billions of online documents in the Internet has created a need for automated methods for classification, where rapid changes and growth of the collection of documents prohibit the use of manual techniques. Automatic text classification is useful for many services such as search, filtering, and routing relevant email to the appropriate addressees. The problem is characterized by very high dimensional data – every word in the document is treated as an attribute – and by little training data - there are typically fewer training documents than there are attributes! This book describes a model of automatic text classification that characterizes the difficulty of a given instance of the problem. One models an instance of the problem by counting the words that occur in the documents each of the classes under consideration. Learnability results refer to this model, and predict how hard it will be to train a classifier for this instance of the problem. Although the proofs of the learnability of TCat concepts are subtle, their application to real data sets is straightforward, as illustrated in this book.

While many learning algorithms have been studied for text, one of the most effective is the Support Vector Machine (SVM) which is the focus of this book. Support vector machines (SVM's) were suggested some time ago by Vladimir Vapnik, but Joachims' work is among the first to seriously explore their use for classifying text. In fact, Joachims created an implementation of SVM's, called SVM^{light}, which has been used by many researchers worldwide, because it is an efficient and easy to use implementation of SVMs that is well suited

to text classification. In this book, Joachims explains and explores the use of SVM's for text classification, reporting experimental results applying SVM's to a variety of real-world text classification tasks. He goes on to explore the use of SVM's for transduction, in which the unlabeled examples that are to be classified are used as part of the procedure for learning a good decision boundary, and demonstrates experimentally that this transduction process significantly improves classification accuracy on several text classification problems.

In addition to developing and experimenting with practical algorithms for text classification, the book goes on to develop components of a general theory of text classification. What is the question one would like such a theory to answer? Perhaps the most important question is "can we predict the accuracy that will be achieved by our trained classifier, as a function of the number of training examples it is provided?" Joachims provides in this book the first steps toward a general theory to answer this question for the problem of text classification. This theory builds on the known statistical distributions of words as they occur in natural language, and posits several other properties of text and text classification problems. Within this set of assumptions, Joachims succeeds in relating accuracy to the number of training examples. In fact, the theory is framed in terms of parameters of the text corpus, such as parameters of the word distribution that can be efficiently measured for any text corpus, then used to characterize the properties of the corpus that influence this relation between accuracy and number of training examples. As with most theories, this one formulates the theoretical problem as a somewhat simplified version of the actual problem. Nevertheless, Joachims demonstrates that the theory successfully predicts the relative difficulty of text learning over three different real-world data sets and classification problems.

In short, the work in this book represents an important step in our understanding of text learning. It describes a state-of-the-art machine learning algorithm for practical text classification, with a freeware implementation accessible over the web. In addition, it provides the first theoretical characterization of text classification learning problems, relating the expected error of a classifier to measurable properties of the text data, plus the number of training examples provided to the learner. This research, combining solid experimental work with highly relevant theory, represents an important contribution to our understanding of text learning, and represents a model for future work combining theory and practice. We believe that students, lecturers, computational theorists, and practitioners will enjoy reading the book as much as we enjoyed accompanying its formation.

Prof. Tom Mitchell
Carnegie Mellon University

Prof. Katharina Morik
Universität Dortmund

Preface

Text classification, or the task of automatically assigning semantic categories to natural language text, has become one of the key methods for organizing online information. Since hand-coding such classification rules is costly or even impractical, most modern approaches employ machine learning techniques to automatically learn text classifiers from examples. However, none of these conventional approaches combines good prediction performance, theoretical understanding, and efficient training algorithms.

Based on ideas from Support Vector Machines (SVMs), this book presents a new approach to learning text classifiers from examples. It provides not only learning methods that empirically have state-of-the-art predictive performance, but also a theory that connects the properties of text-classification tasks with the generalization accuracy of the learning methods, as well as algorithms that efficiently implement the methods.

In particular, the results show that the SVM approach to learning text classifiers is highly effective without greedy heuristic components. To explain these empirical findings, this book analyzes the statistical properties of text-classification tasks and presents a theoretical learning model that leads to bounds on the expected error rate of an SVM. The bounds are based on improved results about leave-one-out estimators for SVMs. These results also lead to a new group of performance estimators for SVMs, called $\xi\alpha$-estimators, and to an improved algorithm for computing the leave-one-out error of an SVM. While all results mentioned so far were for the inductive setting, this book also introduces the idea of transduction to text classification. It shows how and why exploiting the location of the test points during learning can improve predictive performance. To make the SVM approach to learning text classifiers applicable in practice, this book also derives new algorithms for training SVMs. For both the inductive and the transductive setting, these algorithms substantially extend the scalability of SVMs to large-scale problems.

While the work presented in this book is driven by the application, the techniques it develops are not limited to text classification, but can be applied in other contexts as well. In addition, not only can individual techniques be transferred to other tasks, this work as a whole can be useful as a case study for how to approach high-dimensional learning tasks.

This book is based on my dissertation with the title *"The Maximum-Margin Approach to Learning Text Classifiers — Methods, Theory, and Algorithms"*, defended in February 2001 at the Fachbereich Informatik, Universität Dortmund, Germany. From the beginning, it was not written purely for the thesis committee, but with a broader audience in mind. So, the book is designed to be self-contained and I believe it can be useful for all readers interested in text classification — both for research, as well as for product development.

After an introduction to Support Vector Machines and an overview of the state-of-the-art in text classification, this book is divided into three parts: theory, methods, and algorithms. Each part is self-contained, facilitating selective reading. Furthermore, all methods and algorithms proposed in this book are implemented in the software SVM^{light}, making it easy to replicate and extend the results presented in the following.

I hope you will find this book useful and enjoy reading it.

THORSTEN JOACHIMS

Acknowledgments

Without each of the following people, my dissertation and this book would have turned out differently. Or maybe, it would never have been started nor finished.

First, I would like to thank the two people who had the largest immediate influence on this work — Prof. Katharina Morik and Prof. Tom Mitchell. Prof. Morik was the person who got me excited about machine learning. It was her enthusiasm that made and still makes me want to understand the nature of learning. In particular, I thank her for teaching me about good and important research and for the guidance that shaped my dissertation without constraining it. Prof. Mitchell was the person who introduced me to text learning during my time at Carnegie Mellon University. So the particular topic of my dissertation is due to him. I am very grateful for his advice and for the motivation he managed to get across despite the distance. I also thank Prof. Fuhr for his comments and his suggestions about this work.

Particular thanks go to Dr. Phoebe Sengers. She kept me happy and appropriately focused throughout this work. Despite her aversion towards Greek letters, she read this whole document twice, found many logical inconsistencies, and corrected my English.

Many of my colleagues deserve my thanks for helpful discussions, in particular Ralf Klinkenberg, Stefan Rüping, and Peter Brockhausen from the AI Unit. With the DFG Collaborative Research Center on Complexity Reduction in Multivariate Data (SFB475) supporting this work, I got much help from Thomas Fender, Ursula Sondhauß, Prof. Gather, and Prof. Weihs from the statistics department. While more difficult due to the distance, discussions and collaboration with Dr. Tobias Scheffer, Prof. Sebastian Thrun, Dr. Andrew McCallum, Dr. Dunja Mladenić, Marko Grobelnik, Dr. Justin Boyan, Dr. John Platt, Dr. Susan Dumais, Dr. Dayne Freitag, Dr. Carlotta Domeniconi, Dr. Maria Wolters, Dr. Mehran Sahami, Prof. Mark Craven, Prof. Lyle Ungar, Dr. Alex Smola, John Langford, Sean Slattery, and Rosie Jones substantially

influenced this work. Of particular inspiration were the discussions with Prof. Vladimir Vapnik during his visit in Dortmund.

Last but not least, I would like to thank my parents Bernhard Joachims and Hiltrud Joachims, as well as my sister Martina Joachims. They supported me throughout my whole life in every respect and made me interested in science and learning about the world — sorry, Mom, it turned out not to be medicine.

Notation

$\vec{x}_i$	input patterns
y_i	target values (classes)
X	feature space
n	number of training examples
k	number of test examples
N	dimensionality of the input space
$\mathcal{L}$	learner
$\mathcal{H}$	hypothesis space
h	hypothesis from the hypothesis space
$h_{\mathcal{L}}$	hypothesis that the learner $\mathcal{L}$ returns
$R(h)$	(expected) risk of the hypothesis h
$R_{emp}(h)$	empirical risk of the hypothesis h on a training sample
L	loss function
$L_{0/1}$	0/1-loss function
$\vec{w}$	weight vector of a hyperplane $< \vec{w}, b >$
b	constant offset (or threshold) of a hyperplane $< \vec{w}, b >$
δ	margin of a hyperplane
R	diameter of a ball containing the data, usually approximated by $\max \|\|\vec{x}\|\|_2$
α_i	Lagrange multiplier
$\vec{\alpha}$	vector of all Lagrange multipliers
ξ_i	slack variables
$(\vec{x}_1 \cdot \vec{x}_2)$	dot product between vectors $\vec{x}_1$ and $\vec{x}_2$
$\mathcal{K}$	Mercer kernel
Q	Hessian of the quadratic program
Err	error rate
Rec	recall
$Prec$	precision
F_β	F_β-measure
$PRBEP$	precision/Recall breakeven point
$PRAVG$	arithmetic average of precision and recall
$\vec{x}^T$	transpose of the vector $\vec{x}$
$\Re$	the set of real numbers
$\mathbf{N}$	the set of natural numbers
$\|X\|$	cardinality of set X
$abs(a)$	absolute value of a
$\|.\|_1$	L_1-norm , $\|\|\vec{x}\|\|_1 := \sum abs(x_i)$
$\|.\|_2$ or $\|.\|$	L_2-norm (Euclidian distance), $\|\|\vec{x}\|\| := \sqrt{(\vec{x} \cdot \vec{x})}$
$exp(a)$	2.7182818^a
ln	logarithm to base 2.7182818