

# **Boosting**

## **Foundations and Algorithms**

## **Adaptive Computation and Machine Learning**

Thomas Dietterich, Editor

Christopher Bishop, David Heckerman, Michael Jordan, and Michael Kearns, Associate Editors

A complete list of the books published in this series may be found at the back of the book.

**Boosting**

**Foundations and Algorithms**

**Robert E. Schapire**  
**Yoav Freund**

**The MIT Press**  
**Cambridge, Massachusetts**  
**London, England**

© 2012 Massachusetts Institute of Technology

All rights reserved. No part of this book may be reproduced in any form by any electronic or mechanical means (including photocopying, recording, or information storage and retrieval) without permission in writing from the publisher.

For information about special quality discounts, please email [special\\_sales@mitpress.mit.edu](mailto:special_sales@mitpress.mit.edu)

This book was set in Times Roman by Westchester Book Composition.  
Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data

Schapire, Robert E.

Boosting : foundations and algorithms / Robert E. Schapire and Yoav Freund.

p. cm.—(Adaptive computation and machine learning series)

Includes bibliographical references and index.

ISBN 978-0-262-01718-3 (hardcover : alk. paper)

1. Boosting (Algorithms) 2. Supervised learning (Machine learning) I. Freund, Yoav. II. Title.

Q325.75.S33 2012

006.3'1—dc23

2011038972

10 9 8 7 6 5 4 3 2

*To our families*

On the cover: A randomized depiction of the potential function  $\Phi_t(s)$  used in the boost-by-majority algorithm, as given in equation (13.30). Each pixel, identified with an integer pair  $(t, s)$ , was randomly colored blue with probability  $\Phi_t(s)$ , and was otherwise colored yellow (with colors inverted where lettering appears). The round  $t$  runs horizontally from  $T = 1225$  at the far left down to 0 at the far right, and position  $s$  runs vertically from  $-225$  at the top to 35 at the bottom. An edge of  $\gamma = 0.06$  was used. [Cover design by Molly Seamans and the authors.]

# Contents

Series Foreword	xi
Preface	xiii
<b>1 Introduction and Overview</b>	<b>1</b>
1.1 Classification Problems and Machine Learning	2
1.2 Boosting	4
1.3 Resistance to Overfitting and the Margins Theory	14
1.4 Foundations and Algorithms	17
<i>Summary</i>	19
<i>Bibliographic Notes</i>	19
<i>Exercises</i>	20
<b>I CORE ANALYSIS</b>	<b>21</b>
<b>2 Foundations of Machine Learning</b>	<b>23</b>
2.1 A Direct Approach to Machine Learning	24
2.2 General Methods of Analysis	30
2.3 A Foundation for the Study of Boosting Algorithms	43
<i>Summary</i>	49
<i>Bibliographic Notes</i>	49
<i>Exercises</i>	50
<b>3 Using AdaBoost to Minimize Training Error</b>	<b>53</b>
3.1 A Bound on AdaBoost's Training Error	54
3.2 A Sufficient Condition for Weak Learnability	56
3.3 Relation to Chernoff Bounds	60
3.4 Using and Designing Base Learning Algorithms	62
<i>Summary</i>	70
<i>Bibliographic Notes</i>	71
<i>Exercises</i>	71

<b>4</b>	<b>Direct Bounds on the Generalization Error</b>	<b>75</b>
4.1	Using VC Theory to Bound the Generalization Error	75
4.2	Compression-Based Bounds	83
4.3	The Equivalence of Strong and Weak Learnability	86
	<i>Summary</i>	88
	<i>Bibliographic Notes</i>	89
	<i>Exercises</i>	89
<b>5</b>	<b>The Margins Explanation for Boosting’s Effectiveness</b>	<b>93</b>
5.1	Margin as a Measure of Confidence	94
5.2	A Margins-Based Analysis of the Generalization Error	97
5.3	Analysis Based on Rademacher Complexity	106
5.4	The Effect of Boosting on Margin Distributions	111
5.5	Bias, Variance, and Stability	117
5.6	Relation to Support-Vector Machines	122
5.7	Practical Applications of Margins	128
	<i>Summary</i>	132
	<i>Bibliographic Notes</i>	132
	<i>Exercises</i>	134
<b>II</b>	<b>FUNDAMENTAL PERSPECTIVES</b>	<b>139</b>
<b>6</b>	<b>Game Theory, Online Learning, and Boosting</b>	<b>141</b>
6.1	Game Theory	142
6.2	Learning in Repeated Game Playing	145
6.3	Online Prediction	153
6.4	Boosting	157
6.5	Application to a “Mind-Reading” Game	163
	<i>Summary</i>	169
	<i>Bibliographic Notes</i>	169
	<i>Exercises</i>	170
<b>7</b>	<b>Loss Minimization and Generalizations of Boosting</b>	<b>175</b>
7.1	AdaBoost’s Loss Function	177
7.2	Coordinate Descent	179
7.3	Loss Minimization Cannot Explain Generalization	184
7.4	Functional Gradient Descent	188
7.5	Logistic Regression and Conditional Probabilities	194
7.6	Regularization	202
7.7	Applications to Data-Limited Learning	211
	<i>Summary</i>	219
	<i>Bibliographic Notes</i>	219
	<i>Exercises</i>	220



<b>8</b>	<b>Boosting, Convex Optimization, and Information Geometry</b>	<b>227</b>
8.1	Iterative Projection Algorithms	228
8.2	Proving the Convergence of AdaBoost	243
8.3	Unification with Logistic Regression	252
8.4	Application to Species Distribution Modeling	255
	<i>Summary</i>	260
	<i>Bibliographic Notes</i>	262
	<i>Exercises</i>	263
<b>III</b>	<b>ALGORITHMIC EXTENSIONS</b>	<b>269</b>
<b>9</b>	<b>Using Confidence-Rated Weak Predictions</b>	<b>271</b>
9.1	The Framework	273
9.2	General Methods for Algorithm Design	275
9.3	Learning Rule-Sets	287
9.4	Alternating Decision Trees	290
	<i>Summary</i>	296
	<i>Bibliographic Notes</i>	297
	<i>Exercises</i>	297
<b>10</b>	<b>Multiclass Classification Problems</b>	<b>303</b>
10.1	A Direct Extension to the Multiclass Case	305
10.2	The One-against-All Reduction and Multi-label Classification	310
10.3	Application to Semantic Classification	316
10.4	General Reductions Using Output Codes	320
	<i>Summary</i>	333
	<i>Bibliographic Notes</i>	333
	<i>Exercises</i>	334
<b>11</b>	<b>Learning to Rank</b>	<b>341</b>
11.1	A Formal Framework for Ranking Problems	342
11.2	A Boosting Algorithm for the Ranking Task	345
11.3	Methods for Improving Efficiency	351
11.4	Multiclass, Multi-label Classification	361
11.5	Applications	364
	<i>Summary</i>	367
	<i>Bibliographic Notes</i>	369
	<i>Exercises</i>	369

<b>IV</b>	<b>ADVANCED THEORY</b>	<b>375</b>
<b>12</b>	<b>Attaining the Best Possible Accuracy</b>	<b>377</b>
	12.1 Optimality in Classification and Risk Minimization	378
	12.2 Approaching the Optimal Risk	382
	12.3 How Minimizing Risk Can Lead to Poor Accuracy	398
	<i>Summary</i>	406
	<i>Bibliographic Notes</i>	406
	<i>Exercises</i>	407
<b>13</b>	<b>Optimally Efficient Boosting</b>	<b>415</b>
	13.1 The Boost-by-Majority Algorithm	416
	13.2 Optimal Generalization Error	432
	13.3 Relation to AdaBoost	448
	<i>Summary</i>	453
	<i>Bibliographic Notes</i>	453
	<i>Exercises</i>	453
<b>14</b>	<b>Boosting in Continuous Time</b>	<b>459</b>
	14.1 Adaptiveness in the Limit of Continuous Time	460
	14.2 BrownBoost	468
	14.3 AdaBoost as a Special Case of BrownBoost	476
	14.4 Experiments with Noisy Data	483
	<i>Summary</i>	485
	<i>Bibliographic Notes</i>	486
	<i>Exercises</i>	486
	<b>Appendix: Some Notation, Definitions, and Mathematical Background</b>	<b>491</b>
	A.1 General Notation	491
	A.2 Norms	492
	A.3 Maxima, Minima, Suprema, and Infima	493
	A.4 Limits	493
	A.5 Continuity, Closed Sets, and Compactness	494
	A.6 Derivatives, Gradients, and Taylor's Theorem	495
	A.7 Convexity	496
	A.8 The Method of Lagrange Multipliers	497
	A.9 Some Distributions and the Central Limit Theorem	498
	 Bibliography	 501
	Index of Algorithms, Figures, and Tables	511
	Subject and Author Index	513