Adversarial Machine Learning

Written by leading researchers, this complete introduction brings together all the theory and tools needed for building robust machine learning in adversarial environments. Discover how machine learning systems can adapt when an adversary actively poisons data to manipulate statistical inference, learn the latest practical techniques for investigating system security and performing robust data analysis, and gain insight into new approaches for designing effective countermeasures against the latest wave of cyberattacks. Privacy-preserving mechanisms and near-optimal evasion of classifiers are discussed in detail, and in-depth case studies on email spam and network security highlight successful attacks on traditional machine learning algorithms. Providing a thorough overview of the current state of the art in the field and possible future directions, this groundbreaking work is essential reading for researchers, practitioners, and students in computer security and machine learning and for those wanting to learn about the next stage of the cybersecurity arms race.

Anthony D. Joseph is a Chancellor's Professor in the Department of Electrical Engineering and Computer Sciences at the University of California, Berkeley. He was formerly the Director of Intel Labs Berkeley.

Blaine Nelson is a Software Engineer in the Counter-Abuse Technology (CAT) team at Google. He previously worked at the University of Potsdam and the University of Tübingen.

Benjamin I. P. Rubinstein is an Associate Professor in Computing and Information Systems at the University of Melbourne. He has previously worked at Microsoft Research, Google Research, Yahoo! Research, Intel Labs Berkeley, and IBM Research.

J. D. Tygar is a Professor at the University of California, Berkeley, and he has worked widely in the field of computer security. At Berkeley, he holds appointments in both the Department of Electrical Engineering and Computer Sciences and the School of Information.

Cambridge University Press
978-1-107-04346-6 — Adversarial Machine Learning
Anthony D. Joseph , Blaine Nelson , Benjamin I. P. Rubinstein , J. D. Tygar
Frontmatter
More Information

"Data Science practitioners tend to be unaware of how easy it is for adversaries to manipulate and misuse adaptive machine learning systems. This book demonstrates the severity of the problem by providing a taxonomy of attacks and studies of adversarial learning. It analyzes older attacks as well as recently discovered surprising weaknesses in deep learning systems. A variety of defenses are discussed for different learning systems and attack types that could help researchers and developers design systems that are more robust to attacks."

Richard Lippmann, *Lincoln Laboratory, MIT*

"This is a timely book. Right time and right book, written with an authoritative but inclusive style. Machine learning is becoming ubiquitous. But for people to trust it, they first need to understand how reliable it is."

Fabio Roli, *University of Cagliari*

Adversarial Machine Learning

ANTHONY D. JOSEPH
University of California, Berkeley

BLAINE NELSON
Google

BENJAMIN I. P. RUBINSTEIN
University of Melbourne

J. D. TYGAR
University of California, Berkeley

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre,
New Delhi - 110025, India

79 Anson Road, #06-04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107043466
DOI:10.1017/9781107338548

First published 2019

Printed and bound in Great Britain by Clays Ltd, Elcograf S.p.A.

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

Library of Congress Cataloging-in-Publication data
Names: Joseph, Anthony D., author. | Nelson, Blaine, author. | Rubinstein, Benjamin I. P., author. |
Tygar, J. D., author.
Title: Adversarial machine learning / Anthony D. Joseph, University of California, Berkeley, Blaine Nelson,
Google, Benjamin I.P. Rubinstein, University of Melbourne, J.D. Tygar, University of California, Berkeley.
Description: Cambridge, United Kingdom ; New York, NY : Cambridge University Press, 2019. |
Includes bibliographical references and index.
Identifiers: LCCN 2017026016 | ISBN 9781107043466 (hardback)
Subjects: LCSH: Machine learning. | Computer security. | BISAC: COMPUTERS / Security / General.
Classification: LCC Q325.5 .J69 2017 | DDC 006.3/1 – dc23
LC record available at https://lccn.loc.gov/2017026016

ISBN 978-1-107-04346-6 Hardback

Contents

Cambridge University Press
978-1-107-04346-6 — Adversarial Machine Learning
Anthony D. Joseph , Blaine Nelson , Benjamin I. P. Rubinstein , J. D. Tygar
Frontmatter
More Information

Symbols

$A(\cdot)$: The adversary's cost function on $\mathcal{X}$ (see Section 8.1.1). See 203–209, 211, 212, 214, 216, 219–221, 228, 231, 234, 235

$\mathbb{D}$: A set of data points (see also: dataset). See 23–26, 183

N : The number of data points in the training dataset used by a learning algorithm; i.e., $N \triangleq \left|\mathbb{D}^{(\text{train})}\right|$. See 21, 23, 25–27, 36, 38–40, 46, 47, 50, 54, 56, 183, 184, 256

$\mathbb{D}^{(\text{train})}$: A dataset used by a training algorithm to construct or select a classifier (see also: dataset). See 21, 25, 26, 36, 39, 40, 48, 50, 107, 120, 128

$\mathbb{D}^{(\text{eval})}$: A dataset used to evaluate a classifier (see also: dataset). See 21, 22, 25, 27, 36, 39, 40, 46, 48, 50, 51, 128

$\triangleq$: Symbol used to provide a definition. See 23, 24, 26, 57, 58, 60, 107, 108, 137, 139, 141, 142, 149, 153, 204, 206, 256–259, 265, 276, 278, 279, 281, 282

ϵ-$IMAC$: The set of objects in $\mathcal{X}_f^-$ within a cost of $1+\epsilon$ of the MAC, or any of the members of this set (see also: $MAC(f, A)$). See 204–206, 209–214, 216, 219–221, 225, 229, 231–235, 237, 251

$f(\cdot)$: The classifier function or hypothesis learned by a training procedure $H^{(N)}$ from the dataset $\mathbb{D}^{(\text{train})}$ (see also: classifier). See 21, 24–27, 39, 40, 48–51, 54, 71, 74, 102, 120, 139, 174, 176–183, 189, 195, 196, 202–207, 209, 210, 212, 215, 217–220, 234–237, 251

L_ϵ : The number of steps required by a binary search to achieve ϵ-optimality (see Section 8.1.3). See 205, 210–212, 214–218, 220, 221, 225, 228, 232

$MAC(f, A)$: The largest lower bound on the adversary's cost A over $\mathcal{X}_f^-$ (see also: Equation 8.2). See 204, 206, 207, 210, 212–214, 216, 219, 221, 229, 230, 234, 235

$\mathfrak{N}$: The set of natural numbers, $\{1, 2, 3, \ldots\}$. See 77–79, 81, 83, 88, 137, 256, 257, 276

$\mathfrak{N}_0$: The set of all whole numbers, $\{0, 1, 2, \ldots\}$. See 73, 77–79, 81, 82, 86, 256

$\|\cdot\|$: A non-negative function defined on a vector space that is positive homogeneous and obeys the triangle inequality (see also: norm). See 145, 147, 149, 152, 153, 203, 226, 227, 257

ℓ_p **($p > 0$)** : A norm on a multidimensional real-value space defined in Appendix A.1 by Equation (A.1) and denoted by $\|\cdot\|_p$. See 11, 18, 200, 203, 204, 208, 210–214, 216–218, 220, 221, 223, 225–234, 244, 245, 260, 261, 264, 265

$m_{\mathbb{C}}(\cdot)$: A function that defines a distance metric for a convex set $\mathbb{C}$ relative to some central element $\mathbf{x}^{(c)}$ in the interior of $\mathbb{C}$ (see also: Minkowski metric). See 210, 211

$N^{(h)}$: The total number of ham messages in the training dataset. See 107, 108, 111, 277–280

$n_j^{(h)}$: The number of occurences of the j^{th} token in training ham messages. See 107, 108, 111, 277–280

$N^{(s)}$: The total number of spam messages in the training dataset. See 107, 108, 111, 119, 277–280

$n_j^{(s)}$: The number of occurrences of the j^{th} token in training spam messages. See 107, 108, 111, 119, 277–280

$\mathbf{Q}$: The matrix of network flow data. See 137, 138, 152

$\mathbf{R}$: The routing matrix that describes the links used to route each OD flow. See 138, 142, 152

$\Re$: The set of all real numbers. See 23–25, 27, 142, 256–259

$\Re_{0+}$: The set of all real numbers greater than or equal to zero. See 26, 203, 211, 256, 276

$\Re_{+}$: The set of all real numbers greater than zero. See 27, 216, 256, 257

$\Re^D$: The D-dimensional real-valued space. See 24, 139, 142, 143, 147, 202, 216, 226, 257, 259

$\mathbf{x}$: A data point from the input space $\mathcal{X}$ (see also data point). See 22–24, 138, 139, 141, 145, 147, 148, 200, 203–206, 208–211, 213, 223, 224, 226, 255

$\mathbf{x}^A$: A (malicious) data point that the adversary would like to sneak past the detector. See 70, 203–206, 209–215, 217, 218, 221–223, 225, 226, 231, 233–235, 261, 264, 265

$\mathcal{X}$: The input space of the data (see also: input space). See 22–25, 49, 202, 203, 206, 208, 210, 211, 224, 233, 235, 236, 259, 260

D : The dimensionality of the input space $\mathcal{X}$. See 22, 23, 202–205, 212–218, 220, 223–232, 235, 237, 259

$\mathcal{X}_f^-$: The negative class for the deterministic classifier f (see also: negative class). See 203–206, 208, 210–213, 219, 221–224, 226, 231, 233, 235, 236

$\mathcal{X}_f^+$: The positive class for the deterministic classifier f (see also: positive class). See 203, 205, 210–214, 216, 218, 233, 234

y : A label from the response space $\mathcal{Y}$ (see also: label). See 23, 26, 27, 107

$\mathcal{Y}$: The response space of the data (see also response space). See 23–27, 59, 203

$\mathfrak{Z}$: The set of all integers. See 23, 256, 258

Acknowledgments

We gratefully acknowledge the contributions and assistance of our colleagues in making this book possible, who include but are not limited to Sadia Afroz, Scott Alfeld, Tansu Alpcan, Rekha Bachwani, Marco Barreno, Adam Barth, Peter Bartlett, Battista Biggio, Chris Cai, Fuching Jack Chi, David Fifield, Laurent El Ghaoui, Barbara Goto, Rachel Greenstadt, Yi Han, Ling Huang, Michael Jordan, Alex Kantchelian, Hideaki Kawabata, Marius Kloft, Pavel Laskov, Shing-hon Lau, Chris Leckie, Steven Lee, Justin Ma, Steve Martin, Brad Miller, Satish Rao, Fabio Roli, Udam Saini, Tobias Scheffer, Russell Sears, Anil Sewani, Arunesh Sinha, Dawn Song, Nedim Šrndić, Charles Sutton, Nina Taft, Anthony Tran, Michael Tschantz, Kai Xai, Takumi Yamamoto, and Qi Zhong. We additionally thank Matthias Bussas and Marius Kloft for their careful proofreading of Chapter 4 and the staff at Cambridge University Press including Heather Brolly and Julie Lancashire for their help in preparing this manuscript. We would also like to thank the many colleagues with whom we have had fruitful discussions at the Dagstuhl Perspectives Workshop on Machine Learning Methods for Computer Security (Joseph, Laskov, Roli, Tygar, & Nelson 2013) and at the ACM Workshop on Artificial Intelligence and Security (AISec) and other workshops and conferences.

The authors are currently at the University of California, Berkeley, the University of Melbourne, and Google. We thank these institutions. While we were writing this book, some of the authors were at Universität Tübingen, Universität Potsdam, Università di Cagliari, IBM Research, and Microsoft Research, and we also thank those institutions. We offer special thanks to our support staff, including Angie Abbatecola, Kattt Atchley, Carlyn Chinen, Barbara Goto, Damon Hinson, Michaela Iglesia, Shane Knapp, Jey Kottalam, Jon Kuroda, Lena Lau-Stewart, Christian Legg, and Boban Zarkovich.

We are grateful for the financial sponsors of this research. We received U.S. government funding from the Air Force Office of Scientific Research, Homeland Security Advanced Research Projects Agency, National Science Foundation, and State Department DRL, and in some cases through UC Berkeley laboratories (DETERlab and TRUST). Some authors received additional support from the Alexander von Humboldt Foundation, the Australian Research Council (DE160100584), the Center for Long-Term Cybersecurity, the Future of Life Institute, Oak Ridge National Laboratory, and the Open Technology Fund. The opinions expressed in this book are solely those of the authors and do not necessarily reflect the views of any funder.

The authors could not have written this book without the support, encouragement, and patience of their friends and families.