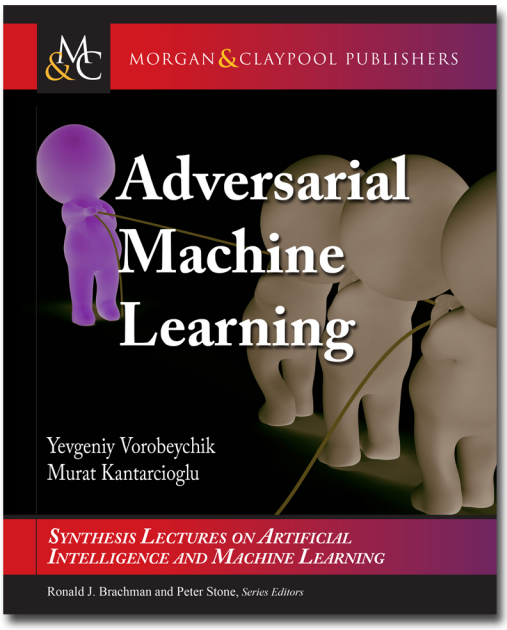
The field of adversarial machine learning has emerged to study vulnerabilities of machine learning approaches in adversarial settings and to develop techniques to make learning robust to adversarial manipulation.



Adversarial Machine Learning

Yevgeniy Vorobeychik, *Washington University in Saint Louis*
Murat Kantarcioglu, *University of Texas, Dallas*

Paperback ISBN: 9781681733951 • eBook ISBN: 9781681733968
Hardcover ISBN: 9781681733975 • August, 2018 • 169 pages
Paperback: $69.95 • eBook: $55.96 • Combo: $87.44
Hardcover: $89.95 • Hardcover Combo: $112.44

The increasing abundance of large high-quality datasets, combined with significant technical advances over the last several decades have made machine learning into a major tool employed across a broad array of tasks including vision, language, finance, and security. However, success has been accompanied with important new challenges: many applications of machine learning are adversarial in nature. Some are adversarial because they are safety critical, such as autonomous driving. An adversary in these applications can be a malicious party aimed at causing congestion or accidents, or may even model unusual situations that expose vulnerabilities in the prediction engine. Other applications are adversarial because their task and/or the data they use are. For example, an important class of problems in security involves detection, such as malware, spam, and intrusion detection. The use of machine learning for detecting malicious entities creates an incentive among adversaries to evade detection by changing their behavior or the content of malicious objects they develop.

The field of adversarial machine learning has emerged to study vulnerabilities of machine learning approaches in adversarial settings and to develop techniques to make learning robust to adversarial manipulation. This book provides a technical overview of this field. After reviewing machine learning concepts and approaches, as well as common use cases of these in adversarial settings, we present a general categorization of attacks on machine learning. We then address two major categories of attacks and associated defenses: decision-time attacks, in which an adversary changes the nature of instances seen by a learned model at the time of prediction in order to cause errors, and poisoning or training time attacks, in which the actual training dataset is maliciously modified. In our final chapter devoted to technical content, we discuss recent techniques for attacks on deep learning, as well as approaches for improving robustness of deep neural networks. We conclude with a discussion of several important issues in the area of adversarial learning that in our view warrant further research.

Given the increasing interest in the area of adversarial machine learning, we hope this book provides readers with the tools necessary to successfully engage in research and practice of machine learning in adversarial settings.

CONTENTS