9 Adversarial Machine Learning Challenges

Machine learning algorithms provide the ability to quickly adapt and find patterns in large diverse data sources and therefore are a potential asset to application developers in enterprise systems, networks, and security domains. They make analyzing the security implications of these tools a critical task for machine learning researchers and practitioners alike, spawning a new subfield of research into adversarial learning for security-sensitive domains. The work presented in this book advanced the state of the art in this field of study with five primary contributions: a taxonomy for qualifying the security vulnerabilities of a learner, two novel practical attack/defense scenarios for learning in real-world settings, learning algorithms with theoretical guarantees on training-data privacy preservation, and a generalization of a theoretical paradigm for evading detection of a classifier. However, research in adversarial machine learning has only begun to address the field's complex obstacles—many challenges remain. These challenges suggest several new directions for research within both fields of machine learning and computer security. In this chapter we review our contributions and list a number of open problems in the area.

Above all, we investigated both the practical and theoretical aspects of applying machine learning in security domains. To understand potential threats, we analyzed the vulnerability of learning systems to adversarial malfeasance. We studied both attacks designed to optimally affect the learning system and attacks constrained by real-world limitations on the adversary's capabilities and information. We further designed defense strategies, which we showed significantly diminish the effect of these attacks. Our research focused on learning tasks in virus, spam, and network anomaly detection, but also is broadly applicable across many systems and security domains and has far-reaching implications to any system that incorporates learning. Here is a summary of the contributions of each component of this book followed by a discussion of open problems and future directions for research.

Framework for Secure Learning

The first contribution discussed in this book was a framework for assessing risks to a learner within a particular security context (see Table 3.1). The basis for this work is a taxonomy of the characteristics of potential attacks. From this taxonomy (summarized in Table 9.1), we developed security games between an attacker and defender tailored to the particular type of threat posed by the attacker. The structure of these games was primarily determined by whether or not the attacker could influence the training data; either

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press

Table 9.1 Our Taxonomy of Attacks against Machine Learning Systems

Axis	Attack Properties		
Influence	**Causative** – influences training and test data	**Exploratory** – influences test data	
Security violation	**Confidentiality** – goal is to uncover training data	**Integrity** – goal is false negatives (FNs)	**Availability** – goal is false positives (FPs)
Specificity	**Targeted** – influence prediction of particular test instance	**Indiscriminate** – influence prediction of all test instances	

a *Causative* or *Exploratory* attack. The goal of the attacker contributed to the game in two ways. First, it generically specifies the attack function (i.e., whether the attack had an *Integrity*, *Availability*, or *Privacy* goal specifying which class of data points is desirable for the adversary). Second, it specifies whether that goal is focused on a small number of points (a *Targeted* attack) or is agnostic to which errors occur (an *Indiscriminate* attack).

Beyond security games, we augmented the taxonomy by further exploring the contamination mechanism used by the attacker. We proposed a variety of different possible contamination models for an attacker. Each of these models is appropriate in different scenarios, and it is important for an analyst to identify the most appropriate contamination model in the threat assessment. We further demonstrated the use of different contamination models in our subsequent investigation of practical systems.

Causative Attacks against Real-World Learners

The second major contribution we presented was a practical and theoretical evaluation of two risk minimization procedures in two separate security domains (spam filtering and network anomalous flow detection) under different contamination models. Within these settings we not only analyzed attacks against real-world systems but we also suggested defense strategies that substantially mitigate the impact of these attacks.

The first system, which we analyzed in Chapter 5, was the spam filter SpamBayes' learning algorithm. This algorithm is based on a simple probabilistic model for spam and has also been used by other spam filtering systems (BogoFilter, Thunderbird's spam filter, and the learning component of Apache SpamAssassin filter (Apa n.d.)), suggesting that the attacks we developed would also be effective against other spam filters. Similarly, they may also be effective against analogous learning algorithms used in different domains. We demonstrated that the vulnerability of SpamBayes originates from its modeling assumptions that a message's label depends only on the tokens present in the message and that those tokens are conditionally independent. While these modeling assumptions are not an inherent vulnerability, in this setting, conditional independence coupled with the rarity of most tokens and the ability of the adversary to poison large numbers of vulnerable tokens with every attack message makes SpamBayes' learner highly vulnerable to malicious contamination.

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press

Motivated by the taxonomy of attacks against learners, we designed real-world *Causative* attacks against SpamBayes' learner and demonstrated the effectiveness of these attacks using realistic adversarial control over the training process of SpamBayes. Optimal attacks against SpamBayes caused unreasonably high false-positive rates using only a small amount of control of the training process (causing more than 95% misclassification of ham messages when only 1% of the training data is contaminated). The Usenet dictionary attack also effectively used a more realistically limited attack message to cause misclassification of 19% of ham messages with only 1% control over the training messages, rendering SpamBayes unusable in practice. We also showed that an informed adversary can successfully target messages. The focused attack changed the classification of the target message virtually 100% of the time with knowledge of only 30% of the target's tokens. Similarly, a pseudospam attack was able to cause nearly 90% of the target spam messages to be labeled as either *unsure* or *ham* with control of less than 10% of the training data.

To combat attacks against SpamBayes, we designed a data sanitization technique; reject on negative impact. RONI expunges any message from the training set if it has an undue negative impact on a calibrated test filter. This technique proved to be a successful defense against dictionary attacks as it detected and removed all of the malicious messages we injected. However, RONI also has costs: it causes a slight decrease in ham classification, it requires a substantial amount of computation, and it may slow the learning process. Nonetheless, this defense demonstrates that attacks against learners can be detected and prevented.

The second system, which we presented in Chapter 6, was a PCA-based classifier for detecting anomalous traffic in a backbone network using only volume measurements. This anomaly detection system inherited the vulnerabilities of the underlying PCA algorithm; namely, we demonstrated that PCA's sensitivity to outliers can be exploited by contaminating the training data, allowing the adversary to dramatically decrease the detection rate for DoS attacks along a particular target flow.

To counter the PCA-based detector, we studied *Causative Integrity* attacks that poison the training data by adding malicious noise; i.e., spurious traffic sent across the network by compromised nodes that reside within it. This malicious noise was designed to interfere with PCA's subspace estimation procedure. Based on a relaxed objective function, we demonstrated how an adversary can approximate optimal noise using a global view of the traffic patterns in the network. Empirically, we found that by increasing the mean link rate by 10% with globally informed chaff traffic, the FNR increased from 3.67% to 38%—a 10-fold increase in misclassification of DoS attacks. Similarly, by only using local link information the attacker was able to mount a more realistic add-more-if-bigger attack. For this attack, when the mean link rate was increased by 10% with add-more-if-bigger chaff traffic, the FNR increased from 3.67% to 28%—an eight-fold increase in misclassification of DoS attacks. These attacks demonstrate that with sufficient information about network patterns, attackers can mount attacks against the PCA detector that severely compromise its ability to detect future DoS attacks traversing the network it is monitoring.

We also demonstrated that an alternative robust method for subspace estimation can be used to make the resulting DoS detector less susceptible to poisoning attacks. The

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press

alternative detector was constructed using a subspace method for robust PCA developed by Croux et al. and a more robust method for estimating the residual cutoff threshold. Our resulting ANTIDOTE detector was affected by poisoning, but its performance degraded more gracefully than PCA. Under nonpoisoned traffic, ANTIDOTE performed nearly as well as PCA, but for all levels of contamination using add-more-if-bigger chaff traffic, the misclassification rate of ANTIDOTE was approximately half the FNR of the PCA-based solution. Moreover, the average performance of ANTIDOTE was much better than the original detector; it outperforms ordinary PCA for more flows and by a large amount. For multiweek boiling frog attacks, ANTIDOTE also outperformed PCA and caught progressively more attack traffic in each subsequent week.

Privacy-Preserving Learning

In Chapter 7, we explored learning under attacks on *Privacy*. After contributing a brief survey of pivotal breaches that influenced thinking on data privacy, we laid the foundation for differential privacy—a formal semantic property that guarantees that information released does not significantly depend on any individual datum. We reviewed the simplest generic mechanism for establishing differential privacy: the Laplace mechanism that introduces additive noise to nonprivate releases, with a scale that depends on sensitivity of releases to data perturbation. After briefly introducing the support vector machine (SVM), we provided an overview of the objective perturbation approach of Chaudhuri et al. (2011). Instead of optimizing the SVM's convex program, we minimized the same program with a random linear term added to the objective.

We discussed our own output perturbation approach (Rubinstein et al. 2012) in Section 7.4. We applied existing results on SVM algorithmic stability to determine the level of classifier perturbation; i.e., the scale of our Laplace noise. We next formulated the utility of privacy-preserving approximations, as the high-probability pointwise similarity of the approximate response predictions compared to nonprivate classifications. We demonstrated results on the utility of both approaches to differentially private SVMs. We generalized our results from the linear SVM (or SVMs with finite-dimensional feature mappings) to SVMs trained with translation-invariant kernels. These results work even for the RBF kernel that corresponds to an infinite-dimensional feature mapping. To do so, we used a technique from large-scale SVM learning that constructs a low-dimensional random kernel that uniformly approximates the desired translation-invariant kernel. Finally we explored lower bounds, which frame fundamental limits on what can possibly be learned privately while achieving high utility. The mechanisms explored, while endowed with theoretical guarantees on privacy and utility, are easily implemented and practical.

Evasion Attacks

In Chapter 8, we generalized Lowd & Meek's near-optimal evasion framework for quantifying query complexity of classifier evasion to the family of convex-inducing classifiers; i.e., classifiers that partition space into two regions, one of which is convex. For the ℓ_p costs, we demonstrated algorithms that efficiently use polynomially many queries

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press

to find a near-optimal evading instance for any classifier in the convex-inducing classifiers, and we showed that for some ℓ_p costs efficient near-optimal evasion cannot be achieved generally for this family of classifiers. Further, the algorithms we presented achieve near-optimal evasion without reverse engineering the classifier boundary and, in some cases, achieve better asymptotic query complexity than reverse-engineering approaches. Further, we showed that the near-optimal evasion problem is generally easier than reverse engineering the classifier's boundary.

A contribution from this work was an extensive study of membership query algorithms that efficiently accomplish ϵ-$IMAC$ search for convex-inducing classifiers with weighted ℓ_1 costs (see Section 8.2). When the positive class is convex, we demonstrated efficient techniques that outperform the previous reverse-engineering approaches for linear classifiers. When the negative class is convex, we applied the randomized ellipsoid method introduced by Bertsimas & Vempala to achieve efficient ϵ-$IMAC$ search. If the adversary is unaware of which set is convex, it can trivially run both searches to discover an ϵ-$IMAC$ with a combined polynomial query complexity; thus, for ℓ_1 costs, the family of convex-inducing classifiers can be efficiently evaded by an adversary; i.e., this family is ϵ-$IMAC$ searchable.

Further, we also extended the study of convex-inducing classifiers to general ℓ_p costs (see Section 8.3). We showed that $\mathcal{F}^{convex}$ is only ϵ-$IMAC$ searchable for both positive and negative convexity for any $\epsilon > 0$ if $p = 1$. For $0 < p < 1$, the MultiLineSearch algorithms of Section 8.2.1 achieve identical results when the positive set is convex, but the nonconvexity of these ℓ_p costs precludes the use of the randomized ellipsoid method. The ellipsoid method does provide an efficient solution for convex negative sets when $p > 1$ (since these costs are convex). However, for convex positive sets, we showed that for $p > 1$ there is no algorithm that can efficiently find an ϵ-$IMAC$ for all $\epsilon > 0$. Moreover, for $p = 2$, we proved that there is no efficient algorithm for finding an ϵ-$IMAC$ for any fixed value of ϵ.

9.1 Discussion and Open Problems

In the course of our research, we have encountered many challenges and learned important lessons that have given us some insight into the future of the field of adversarial learning in security-sensitive domains. Here we suggest several intriguing research directions for pursuing secure learning. We organize these directions into two topics: *i*) unexplored components of the adversarial game and *ii*) directions for defensive technologies . Finally, we conclude by enumerating the open problems we suggested throughout this book.

9.1.1 Unexplored Components of the Adversarial Game

As suggested in Chapter 3, adversarial learning and attacks against learning algorithms have received a great deal of attention. While many types of attacks have been explored,

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press

there are still many elements of this security problem that are relatively unexplored. We summarize some promising ones for future research.

9.1.1.1 Research Direction: The Role of Measurement and Feature Selection

As discussed in Section 2.2.1, the measurement process and feature selection play an important role in machine learning algorithms that we have not addressed in this book. As suggested in Section 3.1, these components of a learning algorithm are also susceptible to attacks. Some prior work has suggested vulnerabilities based on the features used by a learner (e.g., Mahoney & Chan 2003; Venkataraman et al. 2008; Wagner & Soto 2002), and others have suggested defenses to particular attacks on the feature set (e.g., Globerson & Roweis 2006; Sculley et al. 2006) It has been observed that high dimensionality serves to increase the attack surface of *Exploratory* attacks (Sommer & Paxson 2010; Amsaleg et al. 2016), suggesting that (randomized) feature selection be used as a defensive strategy. In game-theoretic models of *Causative* attacks, high dimensions also have computational consequences on finding equilibrium solutions (Alpcan et al. 2016). However, it has also been observed that traditional approaches to feature reduction can be vulnerable to feature substitution (Li & Vorobeychik 2014). The full role of feature selection remains unknown.

Selecting a set of measurements is a critical decision in any security-sensitive domain. As has been repeatedly demonstrated (e.g., Wagner & Soto 2002) irrelevant features can be leveraged by the adversary to cripple the learner's ability to detect malicious instances with little cost to the attacker. For example, in Chapter 5, we showed that tokens unrelated to the spam concept can be used to poison a spam filter. These vulnerabilities require a concerted effort to construct tamper-resistant features, to identify and eliminate features that have been corrupted, and to establish guidelines for practitioners to meet these needs.

Further, feature selection may play a pivotal role in the future of secure learning. As discussed in Direction 9.1.1.2, these methods can provide some secrecy for the learning algorithm and can eliminate irrelevant features. In doing so, feature selection methods may provide a means to gain an advantage against adversaries, but they may also be attacked. Exploring these possibilities remains a significant research challenge.

9.1.1.2 Research Direction: The Effect of Attacker Capabilities

In Section 1.2, we acknowledge that adversarial learning should adhere to Kerckhoffs' Principle: resilient learning systems should not assume secrecy to provide security. However, to understand under what threat models learnability is possible, it is important to characterize the impact of the adversary's capabilities on attack effectiveness.

Question 9.1 Consider underlying stochastic data. How is learning on such data affected by the attacker's information about the data and learner, as well as the attacker's control over the data? What are appropriate parameterizations of attacker capabilities for characterizing learnability?

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press

As learning algorithms generally find patterns in their training data, it is not necessary to exactly reproduce the training data to discover information about the learned hypothesis. In many cases, to approximate the learned hypothesis, the adversary need only have access to a similar dataset.

As observed by Papernot, McDaniel, Goodfellow, Jha, Celik, & Swami (2016) for a special case, reverse-engineered models can be used as surrogates in successful evasion attacks. To the extent that this approach works in general, reverse engineering can amplify an adversary's ability to launch subsequent misclassification attacks.

Question 9.2 How accurate must a surrogate model be for effective misclassification attacks against a target?

As in the near-optimal evasion framework in Chapter 8, the adversary can procure a great deal of information about the learned hypothesis with little information about the training algorithm and hypothesis space.

One motivation for studying reverse-engineering attacks—outside their use in enabling low-information misclassification attacks—is situations in which the defender wishes to protect commercial-in-confidence information about the learner. Tramèr et al. (2016) develop practical reverse-engineering attacks against cloud-based ML-as-a-service systems, both for cases where the model returns only class labels and where the model returns precise confidence values permitting an approach based on solving systems of (nonlinear) equations.

Question 9.3 In general, how effective is reverse engineering at building surrogate models? What guarantees, in terms of query complexity, are possible?

Perhaps the most obvious ingredient to be protected is the training data used to create the learned hypothesis. Settings discussed throughout this book consider adversaries that (partially) control inputs; even in such settings, differential privacy guarantees (as explored in Chapter 7) hold for arbitrary manipulation of all but a single private training datum.

Feature selection (as presented in Section 2.2.1) could potentially play a role in defending against an adversary by allowing the defender to use dynamic feature selection. In many cases, the goal of the adversary is to construct malicious data instances that are inseparable from innocuous data from the perspective of the learner. However, as the attack occurs, dynamic feature selection could be employed to estimate a new feature mapping $\phi'_{\mathbb{D}}$ that would allow the classifier to continue to separate the classes in spite of the adversary's alterations.

9.1.2 Development of Defensive Technologies

The most important challenge remaining for learning in security-sensitive domains is to develop general-purpose secure learning technologies. In Section 3.3.5, we suggested several promising approaches to defend against learning attacks, and several secure learners have been proposed (e.g., Dalvi et al. 2004; Globerson & Roweis 2006; Wang et al. 2006) However, the development of defenses will inevitably create an arms race, so successful defenses must anticipate potential counterattacks and demonstrate that they

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press

are resilient against reasonable threats. With this in mind, the next step is to explore general defenses against larger classes of attack to exemplify trustworthy secure learning.

9.1.2.1 Research Direction: Game-Theoretic Approaches to Secure Learning

Since suggested by Dalvi et al. (2004), the game-theoretic approach to designing defensive classifiers has rapidly expanded (e.g., Brückner & Scheffer 2009; Kantarcioglu et al. 2009; Biggio et al. 2010; Großhans et al. 2013) In this approach, adversarial learning is treated as a game between a learner (which chooses a model) and an adversary (which chooses data or a data transform). Both players are constrained and seek to optimize an objective function (typically at odds with the other player's objective). These approaches find an optimal model against the adversary and one that is thus robust against attacks. This game-theoretic approach is particularly appealing for secure learning because it incorporates the adversary's objective and limitations directly into the classifier's design through an adversarial cost function. However, this cost function is difficult to specify for a real-world adversary, and using an inaccurate cost function may again lead to inadvertent blind spots in the classifier. This raises interesting questions:

Question 9.4 How can a machine learning practitioner design an accurate cost function for a game-theoretic cost-sensitive learning algorithm? How sensitive are these learners to the adversarial cost? Can the cost itself be learned?

Game-theoretic learning approaches are especially interesting because they directly incorporate the adversary as part of the learning process. In doing so, they make a number of assumptions about the adversary and its capabilities, but the most dangerous assumption made is that the adversary behaves *rationally* according to its interests. While this assumption seems reasonable, it can cause the learning algorithm to be overly reliant on its model of the adversary. For instance, the original adversary-aware classifier proposed by Dalvi et al. attempts to preemptively detect evasive data, but will classify data points as benign if a rational adversary would have altered them; i.e., in this case, the adversary can evade this classifier by simply not changing its behavior. Such strange properties are an undesirable side effect of the assumption that the adversary is rational, which raises another question:

Question 9.5 How reliant are adversary-aware classifiers on the assumption that the adversary will behave rationally? Are there game-theoretic approaches that are less dependent on this assumption?

9.1.2.2 Research Direction: Broader Incorporation of Research Methods

Currently, choosing a learning method for a particular task is usually based on the structure of application data, the speed of the algorithm in training and prediction, and expected accuracy (often assessed on a static dataset). However, as our research has demonstrated, understanding how an algorithm's performance can change in security-sensitive domains is critical for its success and for widespread adoption

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press

in these domains. Designing algorithms to be resilient in these settings is a critical challenge.

Generally, competing against an adversary is a difficult problem and can be computationally intractable. However, the framework of robust statistics as outlined in Section 3.5.4.3 partially addresses the problem of adversarial contamination in training data. This framework provides a number of tools and techniques to construct learners robust against security threats from adversarial contamination. Many classical statistical methods often make strong assumptions that their data is generated by a stationary distribution, but adversaries can defy that assumption. For instance, in Chapter 6, we demonstrated that a robust subspace estimation technique significantly outperformed the original PCA method under adversarial contamination.

Robust statistics augment classical techniques by instead assuming that the data comes from two sources: a known distribution and an unknown adversarial distribution. Under this setting, robust variants exist for parameter estimation, testing, linear models, and other classic statistical techniques. Further, the breakdown point and influence function provide quantitative measurements of robustness, which designers of learning systems can use to evaluate the vulnerability of learners in security-sensitive tasks and select an appropriate algorithm accordingly. However, relatively few learning systems are currently designed explicitly with statistical robustness in mind. We believe, though, that as the field of adversarial learning grows, robustness considerations and techniques will become an increasingly prevalent part of practical learning design. The challenge remains to broadly integrate robust procedures into learning for security-sensitive domains and use them to design learning systems resilient to attacks.

9.1.2.3 Research Direction: Online Learning

An alternative complementary direction for developing defenses in security-sensitive settings is addressed by the game-theoretic expert aggregation setting described in Section 3.6. Recall that in this setting, the learner receives advice from a set of experts and makes a prediction by weighing the experts' advice based on their past performance. Techniques for learning within this framework have been developed to perform well with respect to the best expert in hindsight. A challenge that remains is designing sets of experts that together can better meet a security objective. Namely,

Question 9.6 How can one design a set of experts (learners) so that their aggregate is resilient to attacks in the online learning framework?

Ideally, even if the experts may be individually vulnerable, they are difficult to attack as a group. We informally refer to such a set of experts as being *orthogonal*. Orthogonal learners have several advantages in a security-sensitive environment. They allow us to combine learners designed to capture different aspects of the task. These learners may use different feature sets and different learning algorithms to reduce common vulnerabilities; e.g., making them more difficult to reverse engineer. Finally, online expert aggregation techniques are flexible: existing experts can be altered or

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press

new ones can be added to the system whenever new vulnerabilities in the system are identified.

To properly design a system of orthogonal experts for secure learning, the designer must first assess the vulnerability of several candidate learners. With that analysis, the designer should then choose a base set of learners and sets of features for them to learn on. Finally, as the aggregate predictor matures, the designer should identify new security threats and patch the learners appropriately. This patching could be done by adjusting the algorithms, changing their feature sets, or even adding new learners to the aggregate. Perhaps this process could itself be automated or learned.

9.2 Review of Open Problems

Many exciting challenges remain in the field of adversarial learning in security-sensitive domains. Here we recount the open questions we suggested throughout this manuscript.

Problems from Chapter 6

6.1 What are the worst-case poisoning attacks against the Antidote-subspace detector for large-volume network anomalies? What are game-theoretic equilibrium strategies for the attacker and defender in this setting? How does Antidote's performance compare to these strategies? 163

6.2 Can subspace-based detection approaches be adapted to incorporate the alternative approaches? Can they find both temporal and spatial correlations and use both to detect anomalies? Can subspace-based approaches be adapted to incorporate domain-specific information such as the topology of the network? 164

Problems from Chapter 7

7.1 Can the mechanisms and proof techniques used for differentially private SVM by output perturbation be extended to other kernel methods? 197

7.2 Is there a general connection between algorithmic stability and global sensitivity? 198

7.3 An important open problem is to reduce the gap between upper and lower bounds on the optimal differential privacy of the SVM. 198

Problems from Chapter 8

8.1 Can we find matching upper and lower bounds for evasion algorithms? Is there a deterministic strategy with polynomial query complexity for all convex-inducing classifiers? 232

8.2 Are there families larger than the convex-inducing classifiers that are ϵ-$IMAC$ searchable? Are there families outside of the convex-inducing classifiers for which near-optimal evasion is efficient? 232

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press

Problems from Chapter 9

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press

9.3 Concluding Remarks

The field of adversarial learning in security-sensitive domains is a new and rapidly expanding subdiscipline that holds a number of interesting research topics for researchers in both machine learning and computer security. The research presented in this book has both significantly affected this community and highlighted several important lessons. First, to design effective learning systems, practitioners must follow the principle of proactive design as discussed in Section 1.2. To avoid security pitfalls, designers must develop reasonable threat models for potential adversaries and develop learning systems to meet their desired security requirements. At the same time, machine learning designers should promote the security properties of their algorithms in addition to other traditional metrics of performance.

A second lesson that has reemerged throughout this book is that there are inherent tradeoffs between a learner's performance on regular data and its resilience to attacks. Understanding these tradeoffs is important not only for security applications but also for understanding how learners behave in any non-ideal setting.

Finally, throughout this book, we suggested a number of promising approaches toward secure learning, but a clear picture of what is required for secure learning has yet to emerge. Each of the approaches we discussed are founded in game theory, but have different benefits: the adversary-aware classifiers directly incorporate the threat model into their learning procedure, the robust statistics framework provides procedures that are generally resilient against any form of contamination, and the expert aggregation setting constructs classifiers that can do nearly as well as the best expert in hindsight. However, by themselves, none of these form a complete solution for secure learning. Integrating these different approaches or developing a new approach remains the most important challenge for this field.

https://doi.org/10.1017/9781107338548.009 Published online by Cambridge University Press