A survey on adversarial machine learning

Adversarial machine learning is a nascent field in machine learning and cyber security research. Over the last decade the application of machine learning techniques to cyber security systems has become more common. As with any cyber security system attackers have tried to exploit the weaknesses of these systems to carry out attacks. For this part of the assignment we surveyed a few papers on adversarial machine learning that have been published in recent years. These papers range from survey papers that deal with classifying the types of attacks possible on the general class of machine learning systems to papers that deal with the vulnerabilities of individual systems and developing more robust systems.
We start by a short introduction to the field with a survey of 'The Security of Machine Learning' (which was discussed during a lecture). This paper begins with a classification of possible attacks on machine learning based security systems. Attacks have been classified along 3 dimensions - (i)Causative (Altering training data) or Exploratory (Exploiting existing weakness), (ii)Integrity (Aimed at false negatives) or Availability (DoS or False positives) and (iii)Targeted (Aimed at misclassification of particular input) or Indiscriminate. They present this taxonomy along with examples of specific attacks on a spam classifier. These attacks are then formally reintroduced as an adversarial game where the attacker has access to different distributions. This taxonomy is also used to classify known methods of attack and previous literature in the field into classes. They then present possible defences against a few classes of attacks.

The second paper we survey is 'Open problems in the security of machine learning'. This paper focusses on the directions of research in the field of secure machine learning. They broadly classify the directions of research into (i)Finding bounds on adversarial influence - limits on what an attacker can or cannot do, (ii)Value of adversarial capabilities - how the success of the attack depends on the information and influence available to the attacker and (iii)Technologies for secure learning - building machine learning systems that perform well in adversarial environments.
Finding bounds on adversarial influence involves developing possible attacks on existing learning based systems and analysing the amount of information and adversarial effort required to introduce or exploit the required discrepancy in the learning system. This involves works like good word attacks for spam classifiers, data poisoning attacks on PCA based anomaly detectors etc. It also involves identifying which classifiers are robust to noisy training data and reverse engineering for exploratory attacks.
The second field of work is analysing the natural threat models for learners in the deployed systems and the extent to which learners can tolerate adversarial influence. It also deals with the trade-offs between learners generalization performance, the complexity of the learner and the effect of adversarial influence.
Developing technologies for secure learning systems involves identifying the potential threat models and developing systems resilient to these threats. If the attacker has access to the training data this identification and removal of the malicious data by outlier removal methods falls under this class. Developing statistically robust learners with guaranteed performance

or using mixtures of experts to mitigate the attackers' influence fall under the purview of this class.

The next paper we surveyed is 'Poisoning Attacks against Support Vector Machines'. This paper deals with a family of attacks on a SVM classifier that inject specially crafted malicious data into the training samples to influence the performance of the classifier. The attacker is assumed to be aware of the training data distribution. The problem of introducing malicious data points is formulated as an optimization measure, subject to the condition that that the best possible solution to the SVM will be obtained. Since this optimization is shown to be non-convex, a simple gradient ascent algorithm is used to generate the attack points. Since the method depends only on the gradients of the dot products of the points, it works for kernelized SVM as well.
For the problem formulation, it is assumed that the attacker draws a validation dataset and assesses the performance of the SVM trained on the dataset + 1 injected data point on it. It can then be used to compute the gradient of the cost function on the validation set with respect to the injected point which is found to depend only on the dot product of the input data points and therefore can easily be extended to kernelized SVM. They choose a random point from the existing training set as a starting point and then apply gradient ascent until convergence to obtain the attack point.
The performance of the system is then evaluated on the MNIST image dataset for the problems 7 vs 1, 9 vs 8 and 4 vs 0. They choose a training dataset of 100 points and a validation set of 500 points. A single attack point increases the initial error rate of 2-5% to 15-20%. Considering that this attack is not scalable to larger datasets due to it's computation intensive method for obtaining attack points (each gradient step involves training a SVM), it present a good approximate upper bound on the vulnerability of SVMs to carefully crafted attack points in the training dataset.

The last paper surveyed is 'ANTIDOTE: Understanding and Defending against Poisoning of Anomaly Detectors'. It deals with evaluating the performance of PCA-subspace based anomaly detection method under adversarial conditions and a robust statistics based anomaly detection method is proposed. It is assumed that the adversary knows that the ISP is using a PCA-subspace based anomaly detection method and is trying to modify the training set so that the learned set of PCA vectors is distorted. The attacker adds chaff (additional traffic) to the network in order to modify the training phase of the learner. It is well known that the median is a much more robust statistic than the mean. Similarly, a mean absolute deviation based anomaly detector is proposed which is robust to the introduction of additional traffic.
The idea behind a PCA-subspace based anomaly detector is the assumption that normal traffic flows exist in a subspace much smaller than the dimension of the entire flow. This is useful for detecting volume anomalies i.e. huge amounts of traffic with very small components in the normal traffic subspace. The paper assumes different levels of attacker information (none, local, global) and the time horizon of poisoning (short term, boiling frog) and evaluates the performance of the 2 detectors for all these scenarios. They show that the ANTIDOTE method has a much higher breakdown point compared to conventional PCA measures.

With the dramatic increase in security data, machine learning based methods are here to stay and attackers will target their weaknesses. The study of adversarial machine learning and developing methods robust to malicious data and difficult to evade is the need of the hour.