Open Problems in the Security of Learning

allen

2020 年 7 月 11 日

1 摘要

本文提出三个机器学习与安全的方向。本文通过提出问题的方式，将已有的研究成果和未来研究的方向进行了总结和展望。本文的参考文献比较有价值。本文发表于2008年，从后续的论文中可以发现，很多研究确实按照这个文章的思路进行。

* *we suggest that finding bounds on adversarial influence is important to understand the limits of what an attacker can and cannot do to a learning system.*

* we investigate the value of adversarial capabilities the success of an attack depends largely on what types of information and influence the attacker has.

* we propose directions in technologies for secure learning and suggest lines of investigation into secure techniques for learning in adversarial environments.

2 介绍

2.1 背景

机器学习的研究，目前朝着两个方向发展。

* statistical machine learning has entered the mainstream as a broadly useful technique for building applications. In adaptive systems, machine learning enjoys several advantages over handcrafted rules and other approaches: it can infer hidden patterns in data, it can adapt quickly to new signals and behaviors, and it can provide statistical soundness.

* the need to protect systems against malicious adversaries continues to increase across computing applications. Rising levels of hostile behavior have plagued application domains such as email, web search, pay-per-click advertisements, file sharing, instant messaging, mobile phone communications,and others. As the motivation for attacks becomes increasingly fiscal, attackers employ more sophisticated methods and the computing landscape grows ever more treacherous.

3 攻击者影响的上下限

3.1 对于学习结果（ learner Performance）的影响

问题1：我们能否给出学习者在现实世界中受到最小影响时犯的最大错误数？例如：异常检测，垃圾邮件过滤，蠕虫签名生成，钓鱼检测，点击欺诈检测等问题。在这一前提下是否可以降低误差范围？本文在这一节，给出了已有的研究结果。这一类问题是机器学习领域分类器误差的常见问题，如何提高分类器模型的准确率。

3.2 对于逆向工程（ Reverse Engineering State）的影响

问题2：我们能否量化更为复杂的分类器和回归器的攻击复杂性，以及对抗函数的成本？有监督的机器学习模型可以分为两类：分类和回归。我们使用准确率、精确率和召回率对分类器的性能进行评价，使用均方误差、可释方差和R方差对回归器进行评价。问题在于在安全学习中，我们如何在已知攻击者干扰了数据样本之后，对起进行评分。同时得出进行安全防护的成本。问题3：有没有，或者说那一些分类器可以被证明是无法逆向工程的？攻击者污染样本时，一般会通过逆向工程分析训练模型，然后针对性的提供污染样本。我们是否可以训练无法逆向的机器学习模型。

4 对抗者能力的价值（量化攻击威胁）

我们的学习系统如何理解攻击者机器学习的模型是我们进行研究的第一步。攻击者机器学习的模型可以通过攻击目标和攻击能力，攻击能力可以解释为信息能力和控制能力。第二步，我们需要对攻击者攻击能力进行量化。信息能力包括攻击者对于学习者信息的了解，比如学习者的学习特征、学习算法、当前决策函数、策略的培训和再培训以及良性数据的生成过程等。控制能力包括攻击者对于学习系统的训练和测试数据的控制能力。

4.1 确定对手的能力

问题4：在已经部署的系统中，学习者使用的自然威胁模型是什么？相对于这些模型，遇到攻击者的信息和控制时，学习者有多安全？自然威胁模型可以使用传统的机器学习方法进行分析。遇到攻击者学习者的安全性并不高。这一节，作者阐述了当时的研究现状。

4.2 描述可容忍的能力

问题5：在完全的机器学习供方中，学习者能够容忍那些攻击者的能力？进一步的，我们如何描述这种可容忍性？实际情况中，学习者和攻击者都是基于已有的信息进行模型的训练。模型完成之后再对训练的样本进行调整，进而修改模型。问题6：我们如何定量的描述学习者对于无害数据、假设数据以及攻击者数据之间的差别？

4.3 防卫者信息

问题7：是否存在能够不被防卫者发现的攻击者（显著影响学习系统的攻击者）？我们如何去判断攻击者对于防卫者的影响较为显著？能否证明某些攻击者是十分隐蔽的？如果存在我们如何限制他？攻击者为了隐藏攻击来源或者是防止防卫者发现一般会逃避检测。

5 安全学习技术

我们需要设计一个对敌对的污染数据具有弹性防护能力的学习代理环境。问题8：在真实的威胁环境中，是否存在一个对攻击具有弹性防护能力的学习系统？我们如何构建这个系统？首先，我们需要构建一个攻击者模型，能够针对学习系统进行攻击。如同第三节描述的攻击者的能力。然后，我们需要考虑对手可能采取的行动、学习者应对的行动以及对手和学习者最终的结果，并设计一种算法来应对已经预期到的安全威胁，这种算法需要有一定的鲁棒性和容忍性。最后，我们需要评估这个威胁模型的局限性和弱点。

5.1 设计一个安全敏感的学习者

以下三本书中描述了如何设计一个框架。

* Frank R. Hampel, Elvezio M. Ronchetti, Peter J.Rousseeuw, and Werner A. Stahel. Robust Statistics:The Approach Based on Influence Functions.Probability and Mathematical Statistics. John Wileyand Sons, 1986.

* Peter J. Huber. Robust Statistics. John Wiley and Sons, 1981.

* Ricardo A. Maronna, Douglas R. Martin, and Victor J. Yohai. Robust Statistics: Theory and Methods. John Wiley and Sons, New York, 2006.

在经典统计学中，数据样本是一个共同的模型分布产生的，而安全领域的数据却不是这样。由于存在攻击者的干扰样本，我们需要扩充已知和未知的样本。因此样本参数的估计、检验、线性模型和其他经典变量都存在鲁棒变量。问题9：传统机器学习中的稳健性测量使用的影响函数IF（influence function）和崩溃点BP（breakdown point）是否可以在安全学习中使用？是否存在更加合适、细粒度的稳健性测量方法？问题10：当攻击者对于样本数据的污染是有限的情况下（第二节问题可以量化计算的情况下），是否可以使用鲁棒性措施设计一个安全的安全学习者？传统的工具可以使用，新的工具还在研究。。。。

5.2 正交专家（异常值分解（singular value Decomposition）的一种方法）

安全敏感框架最后的组成是基于3.2中所描述基于游戏理论的，与攻击者进行在线博弈的系统。这一框架的相关技术已经被开发出来了。开发者会利用专家建议对学习者进行加权从而进行预测。问题11：我们如何设计在线学习的安全防御框架？这种设计可以自动完成么？我们可以使用异常值分解（svd）的方法进行设计。理论上可以自动完成。