深入浅出对抗性机器学习（AML）

笔者来自于南京大学系统安全与软件安全实验室，南京大学软件新技术重点实验室，南京大学计算机科学计算机科学与技术系。南京大学计算机科学与技术学科是国家一级重点学科，15 年连续三次被评为“优秀类”国家重点实验室、计算机类第一名。

针对 adversarial machine learning 这个问题而言，AI 出身的小伙伴们可能认为，这样的工作应该只能看做模型的鲁棒性或泛化能力不够强，但是从安全角度考虑，其实所谓的“安全”概念，是从模型的设计者角度出发，考虑到模型的行为超出意料之外，让模型设计者手足无措，因此我们认为是可能存在“潜在威胁”，因而将这类行为归类为安全问题。特此解释。

那么针对模型的攻击问题，我们主要分为两大类，就是从训练阶段和推理（inference）阶段来进行讨论。

训练阶段的攻击（Training in Adversarial Settings）

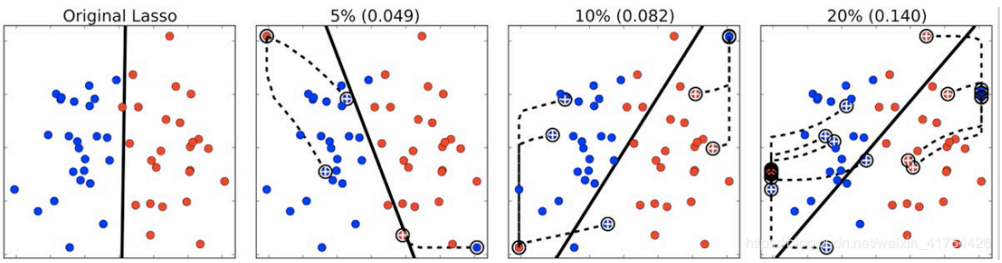
训练阶段的恶意攻击，主要的目的就是针对模型的参数进行微小的扰动，从让而模型的性能和预期产生偏差。这样的行为主要是通过数据投毒来完成的。

举个简单的例子，当我们在进行食品加工的时候，在里面少量的加入一些“有毒”的成分，那这个食品就会和当时预想的做出来不一样，从美味变成了毒物。当然，数据投毒没办法做到“一颗老鼠屎坏了一锅粥”，但是能通过尽量少的“老鼠屎”坏了尽量多的“粥”就是它的目的了。

不过在此之前，有个前提，在 PAC 理论中，有一个已经论证的结论：对于任意的学习算法而言，其置信度 β，必须满足 β≤Σ/1+Σ，其中 Σ 表示了学习准确率。那么也就是说，当我哦需要达到 90%的学习准确率（Σ=0.1），那么我被扰动的数据量必须少于 10%（0.1/1+0.1）。

（M. Kearns and M. Li, “Learning in the presence of malicious errors,” SIAM Journal on Computing, vol. 22, no. 4, pp. 807–837, 1993.）

1、标签操纵（label manipulation）

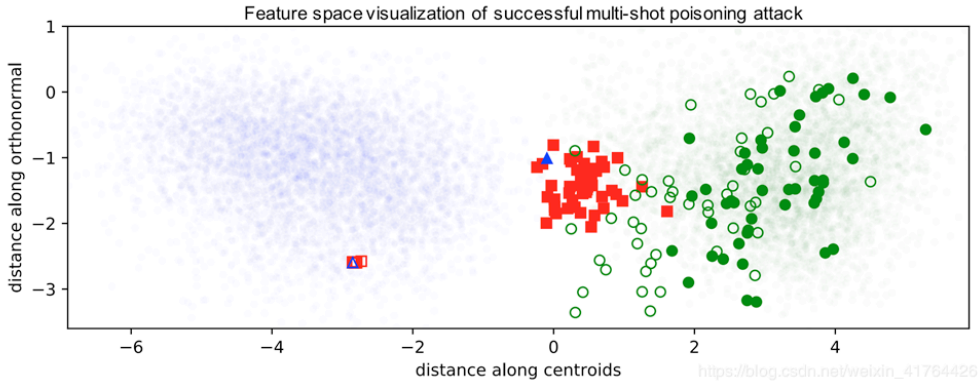


这个方法很直观，就是直接通过对于训练数据的标签进行替换，让数据样本和标签不对应，最后训练的结果一定是不如预期的。有前人在 SVM 的场景下，随机替换了约 40%的数据，对其算法进行了破坏，最后的效果也是如期的很好。其实这只是在二分类问题中起到了比较好的结果，但是在多分类的情况下并没有很好的解释，或者是实证性的研究。（这里可以有一个比较有趣的思考，如果二分类的分类替换了 40%的数据会导致模型的预测结果很不好，那么多分类的 SVM 需要替换多少数据样本的标签呢，是需要替换更少的标签，还是更多？是随机替换还是有目标性的都替换成一种？）

后来的研究则是对这个标签操纵的过程更加优化，是否能通过更少的标签替换，来实现更强烈的模型扰动，从而产生更有说服力的攻击模型。

（B. Biggio, B. Nelson, and P. Laskov, “Support vector machines under adversarial label noise.” in ACML, 2011, pp. 97–112.）

2、输入操纵（input manipulation）



在此攻击场景下，攻击者需要获知模型的算法类型，并且能结束到训练集。

比较直接的攻击方式，则是通过在线的方式获得训练数据的输入权，那么最终的结果就是直接通过恶意数据来扰动在线训练过程，自然最后的结果就是脱离预期，从而导致恶意者的操纵成功。

而当我们无法接触到在线模型的时候，我们只能通过线下的方式操纵训练数据，那么则需要构造尽量少且恶意程度尽量高的恶意样本，那么这就可以使用梯度上升的方法去达到局部分类错误最大的结果，从而完成样本构造，然后再输入到模型中进行训练。

那么，当我们无法直接接触到在线训练模型，或离线时，我们也无法解除到训练数据，我们该怎么进行输入的操纵呢？从之前的流程介绍中我们也提到了，在物理世界获取数据的时候，这阶段并没有受到很好的保护。因此这阶段的数据，我们可以通过恶意的攻击物理世界中的数据，例如交通信号灯，或者是自动驾驶摄

像头正在拍摄的图像等。通过其在数据转换之前，就进行数据的污染，或是数据表示的污染。

（M. Kloft and P. Laskov, “Online anomaly detection under adversarial impact,” in International Conference on Artificial Intelligence and Statistics, 2010, pp. 405–412.）

推理阶段的攻击（Inference in Adversarial Settings）

当训练完成一个模型之后，这个模型就可以看做一个 BOX，那么这个盒子中，对于我们如果是透明的话，我们就把它当成是“白盒”模型，如果这个盒子中，我们什么都看不了，我们就把它当成“黑盒”模型。（我们在这个部分不讨论灰盒模型 -。-）

那么针对白盒和黑盒的进攻手段自然是不同的，但是最终的目的都是希望能对模型的最终结果产生破坏，与预期脱离。其影响力以及攻击的构造粒度也是有所不同的。

1、白盒攻击（White-Box Adversarial）

当然这种所谓的“白盒攻击”，需要提供一个很“假”的前提——就是我们需要知道里面所有的模型参数，这个在现实生活中是非常不现实的。除非是，当模型被打包压缩到智能手机上之后，然后恶意者通过逆向工程来进行原有模型的复原，才有可能。当然这种情况出现的情况非常低了，因此我们需要有这种前提假设。

$$\underset{\gamma}{argmin}\, h(x+r) = l \qquad s.t. \qquad x + r \in D$$

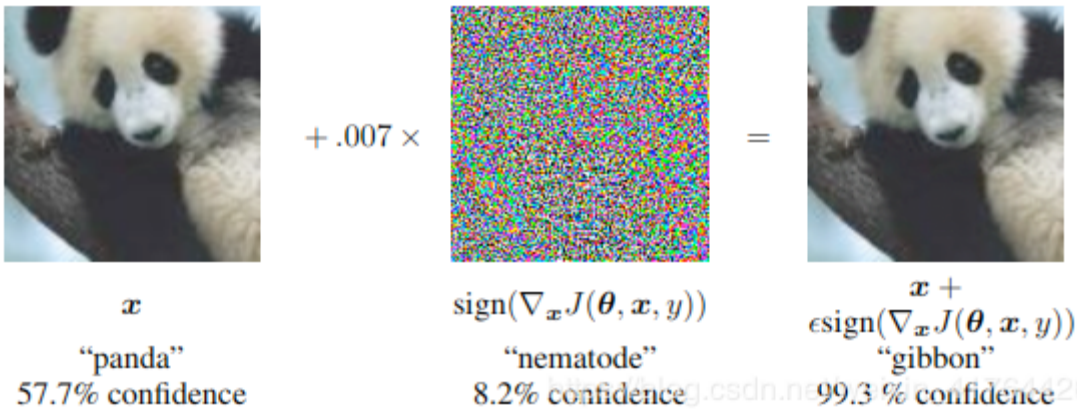
（1）

看到如上公式，其中 x 是数据样本的特征，l 是数据样本通过函数 h（x）预测的结果，l 就是预测的结果。我们的数据样本通过模型的预测结果可能是 k，但是我们希望通过尽量小的扰动 r，最后通过模型预测的结果是 l（然而 x 的分类目标并不是 l）。目标很明确。

（ B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. ˇSrndi´c, P. Laskov, G. Giacinto, and F. Roli, “Evasion attacks against machine learning at test time,” in Machine Learning and Knowledge Discovery in Databases. Springer, 2013, pp. 387–402.）

当然这样的方法对于非凸的模型，例如深度神经网络也有类似的工作，同样也是能通过较小的扰动，来达成模型的误分类目的。

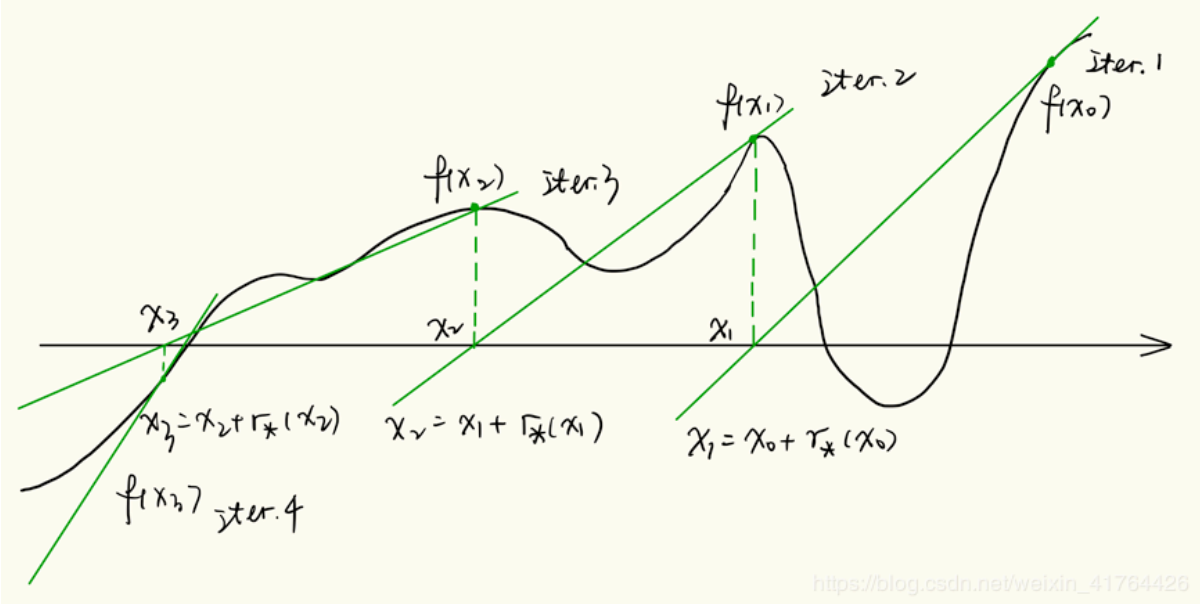
（ D. C. Liu and J. Nocedal, “On the limited memory bfgs method for large scale optimization,” Mathematical programming, vol. 45, no. 1-3, pp. 503–528, 1989. ）



当然可以看到，如式（1）中所示，我们可以很明显的看到，其实如何快速求解这个扰动“r”是个问题，因此之后就有工作专门针对这个问题进行了探索，给出了如下方法，FGSM：

（2）

如式（2）所示，通过梯度可以快速求到，通过最小的扰动获得的最后的攻击目的。



the process of FGSM

那么后续的工作无外乎就是从两个方向进行优化，一方面就是尽量少的对样本扰动，从而能达成攻击，另一方面对尽量少的样本特征进行操纵，通过算法的优化，从而能达到更高的错误识别率。

（N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in Proceedings of the 1st IEEE European Symposium on Security and Privacy. IEEE, 2016.）

有一个很有趣的现象，是这样描述的，其实在数据进入预处理步骤之前，在物理世界中，如果没有一个很好的表示形态，即使经过了预处理，模型也很难识别。这就给了研究者一些启发，对图片进行打印之后，再拍照让模型进行识别；亦或是把人脸的图片打印在玻璃上，然后再进行识别。这样的结果，都会有很高的误识别率。（虽然目前 CV 发展的势头很好，但是由此看来，还是有不少算法对于环境和背景的敏感程度很高）

大家都认为，adversarial machine learning 应该关注在分类问题上，但是其实并不然，其实如果一个 AI 系统是以 agent 为核心，或是以 multi-agent 为核心的强化学习系统的话，那也是有可以攻击的点的，例如改变环境获取的结果？等……（只是猜想），现在有课题组可以在一些固定模式下自动进行星际争霸的游戏，如果攻击了这样的系统，应该还是很有趣的。

不仅模型的预测结果是有脆弱的地方，同时，当我们拥有模型参数的时候，也是可以进行模型训练集数据分布的预测的。虽然这个并不是最重要的信息，但是也是一部分关于模型的隐私。

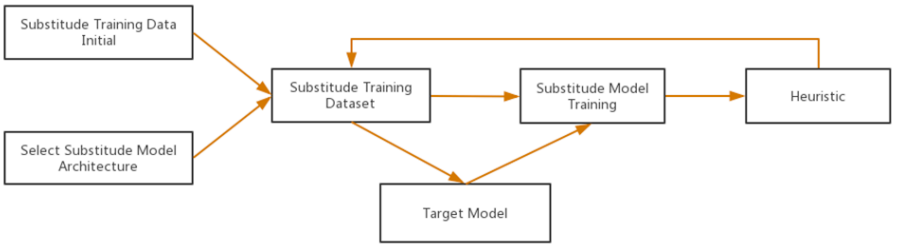
（G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici, "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers," International Journal of Security and Networks, vol. 10, no. 3, pp. 137–150, 2015.）

2、黑盒攻击（Black-Box Adversarial）

当模型处于黑盒的时候，更加符合现实的场景，但是这比白盒的模型缺少了更多的模型信息。因此，大家就从几个角度考虑如何进行模型攻击：通过输入和输出猜测模型的内部结构；加入稍大的扰动来对模型进行攻击；构建影子模型来进行关系人攻击；抽取模型训练的敏感数据；模型逆向参数等。

其中我觉得比较有意思的是两个方法，一个是加入扰动来对模型进行攻击。这个方法最主要针对的是，找到原有模型的"blind spot"，或是说"blind area"。这些区域主要是原有模型模棱两可的区域，或是 boundry，这对二分类的问题来说可能这些区域比较小或是比较狭窄，但是如果针对的是多分类问题，就可能在高维空间中提现出更多的"blind area"。因此尽量高的命中这些盲区，是这种方法致力于的方向，同时这里也提出一个思考，这样的盲区是否是可以定向搜索的，或是说是否可以用一个模糊的算法 bound 住这些区域。

第二是建立影子模型，这个 process 很有意思，通过构建一个功能性类似的模型，来仿造一个攻击空间。这有点像军事演习的意思，我想要在战场上打出好的效果，就要模拟产战场上可能发生的情况，但是目前战场的情况我一无所知，所以我只能根据大致的情况去模拟。模型也是如此，只能对黑盒的情况进行对应的训练模拟，然后对其进行“白盒”的尝试，由于模型的迁移性还不错，或者说类似的算法都有不少的相同点，因此，影子模型的攻击成效还是不错的。



Shadow Model Establishment

至此，基本上从训练阶段到推理阶段的攻击都大致介绍了一遍，也算是给自己再复习了一遍。有所不对，或是思考不够全面的，还请大家斧正。

以上。