**Privacy Risks of Securing Machine Learning Models against Adversarial Examples**

Liwei Song，Reza Shokri ，Prateek Mittal

（Shokri et al.在2017年提出了MIA）

**要点：**

1.结合了安全与隐私领域，而不像此前的大多研究将其分开考虑。

2.对抗防御（adversarial defense）通过保障在训练样本的小范围内模型预测不发生变化来提高鲁棒性。对抗防御天然依赖于训练集，因此更容易受到成员推理攻击。（实验表明：比起自然模型，使用了对抗防御的模型会将MIA的攻击优势增加4.5倍左右。）

3.提出两种针对对抗防御的新的MIA。

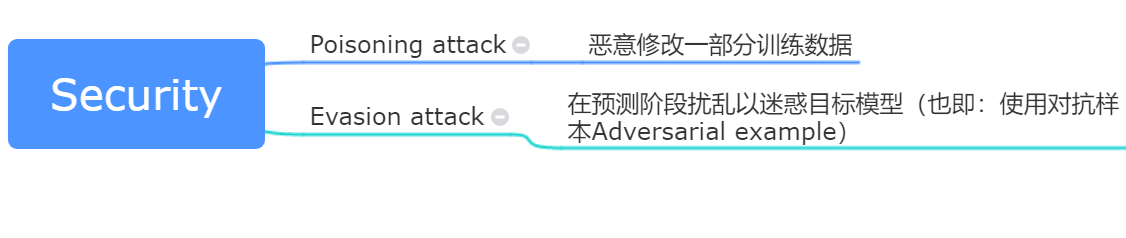
4.提出两种对策。

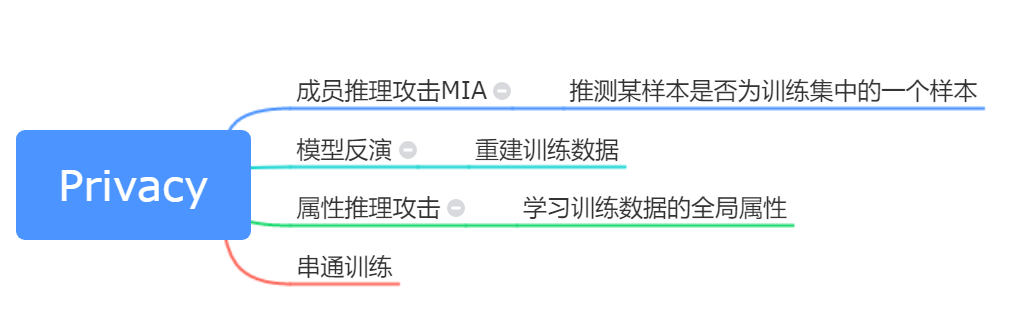
注：以下用MIA表示成员推理攻击（Membership inference attack）。

**Introduction**

1. 安全领域：敌手目标是诱发ML模型进行错误分类。

隐私领域：敌手目标是获取目标模型或训练数据的隐私信息。





本文将Evasion attack（即Adversarial example）与MIA进行了结合。

1. Contributions:

（1）提出两种针对Adversarial robust model的新的MIA，推理的准确率更高。

（2）在六种state-of-art的对抗防御方法上进行了MIA，其中3种为empirical defenses 3种为verifiable defenses。

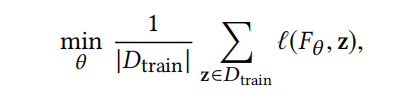
（3）探究了影响针对对抗防御模型的MIA表现的因素。

（4）提出抵御方式。

**Backgroud and related work**

这部分对Adversarial examples与MIA进行了详细阐述，这里简要总结：

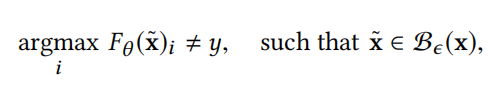
1. **Natural training algorithm：**目标是使预测与真实的label匹配，方式是最小化所有训练样本的预测损失。



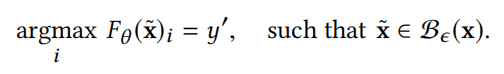
其中l用于计算预测loss，一般是交叉熵损失。

1. **Adversarial examples：**可分为Untargeted与Untargeted 两类

**Untargeted：**

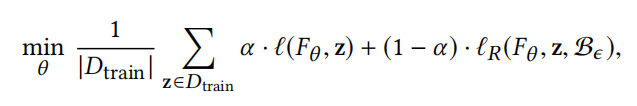


**Targeted：**



(在扰动限制范围内，分别使预测分类不等于真实分类y和使预测分类等于特定的label y')

为保障对抗鲁棒性，在算法基础上会加入一个robust loss value，即下式中的lR：



1. **对抗防御**分为两类：empirical defenses和verifiable defenses

Empirical defenses将robust loss value近似为每一步产生的对抗样本并计算他们的预测损失。本文采用的三种方法是：PGD-Based Adversarial Training，Distributional Adversarial Training，Difference-based Adversarial Training 。

Verifiable defenses相较之下可以保障鲁棒性。本文采用的三种方式是Duality-Based Verification，Abstract Interpretation-Based Verification，Interval Bound Propagation-Based Verification。

1. **MIA：**

2017年Shokri et al.使用的是影子模型进行成员推理，Yeom et.al的实验显示简单的置信度-阈值方法便可达到相似效果，本文采用的即为这种置信度方式：

设置阈值（可用影子方法找出），将预测的置信度值与阈值比较，若预测值高则表示该样本为成员。

**MIA against Robust Models**

Robust training可能会导致过拟合，引发更严重的泛化误差，且其对训练数据更为敏感（敏感指的是当有或没有一个数据参加训练时，模型表现的差异），因此和Natural models相比，Robust training可能会泄露更多成员信息。

本文使用两种新的方式进行MIA：（启发来自上一节中的3）

1. **Exploiting the Model’s Predictions on Adversarial Examples**

（以Targeted adversarial为例，Untargeted 更简单，只需要置信度-阈值策略即可，Targeted adversarial 则因为对K类有K-1个对抗样本，需要使用2017年的影子模型 ）

（1）对每一个label，用training 与test 样本的子集+上一节介绍的PGD方式生成对抗样本。

（2）计算Targeted adversarial examples的模型预测值，然后用其训练成员推理分类器（二分类器，判断成员与非成员）。

（3）使用（1）中剩下的样本来实施MIA。

（详细步骤在实验部分给出）

2. **Exploiting the Verified Worst-Case Predictions on Adversarial Examples**

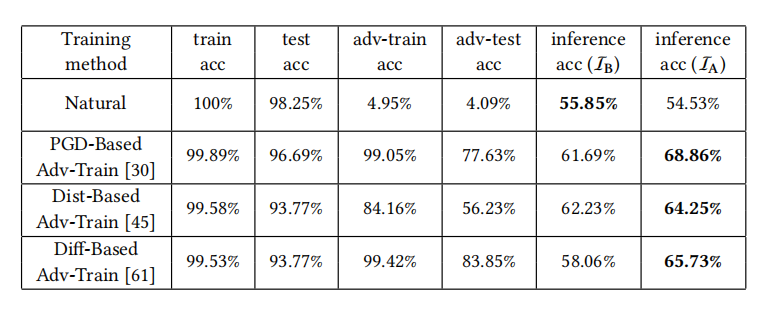
使用验证策略V获取输入的最坏情况预测，并使用最坏情况预测置信度来推测成员。本文的推理策略使用与目标模型可验证鲁棒训练过程相同的验证方法。（即上一节中3所列出的方式）

**Experiment setup**

（上述两种新方法这里仅展示第一种）

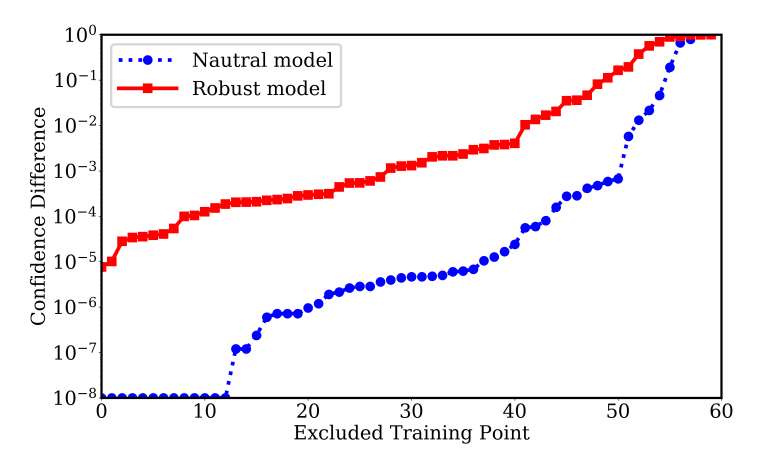
|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Yale Face | Fashion-MNIST | CIFAR10 |
| Neural network | CNN | CNN | CIFA |

1.***以在数据集Yale Face上的表现为例***，结果如图：



注意最后两列，分别指的是置信度-阈值推理攻击IB与本文的方式IA两种MIA的准确率。可以看出本文方式攻击的准确率更高，且三种防御方式是的模型更易遭受MIA。

2. ***Sensitivity Analysis***



x轴：除去的训练样本

y轴：包括和不包括这些点分别训练模型，得到的置信度的差

由图可以看出，去掉一些点后，Robust model 置信度差更大，因此对训练数据更加敏感。

3.***Inference attacks using targeted adversarial examples***

（在CIFAR10数据集上）

**Targeted adversarial examples：**

(1)对每一个输入，用本文提出的方法1计算其余九个label所对应的 Targeted adversarial examples（因为该数据集共有10个label，对每个样本除去正确的那个标签，还有九个标签）。

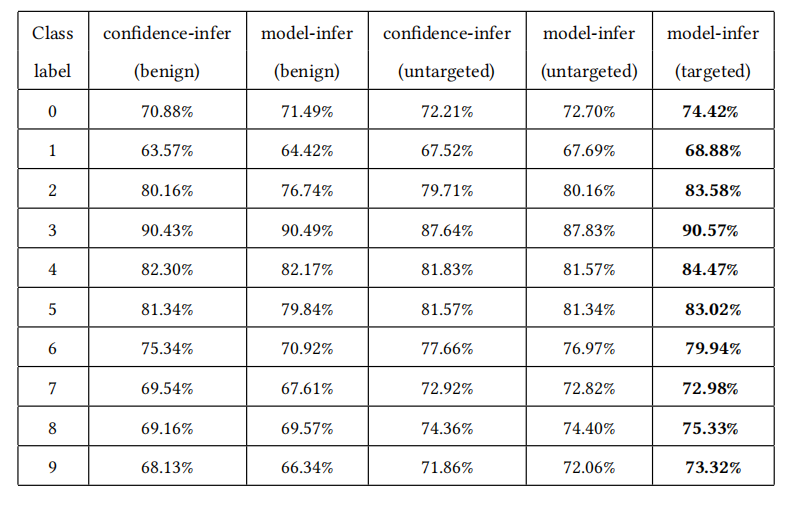
(2)对所有的Targeted adversarial examples计算其预测向量。

(3)使用17年的影子模型来执行MIA。

(4)用数据集同一类中剩下的training和test样本来测试成员推理模型。

**Untargeted adversarial examples and Benign examples:**

只需改动（3）中为置信度-阈值方式。



表格中最后一列，基于推理策略的Targeted adversarial examples的攻击准确率是最高的。（原因在于targeted包含了输入样本的标签到其余每个标签的决策边界的距离，而untargeted包含的只有最近的类的距离，因此targeted泄露的消息更多）。

**Potential countermeasures**

1.Temperature scaling

2.Regularization

**Conclusion**

本文结合安全与隐私问题，探究了针对对抗防御（即缓和对抗样本）的成员推理隐私问题，结果表明对抗防御更易受到MIA，提醒各位在考虑防御对抗样本的方案时，也要考虑MIA的影响。