南開大學
Nankai University

数据安全课程期末报告

论文研读：Systematic Evaluation of Privacy Risks of Machine Learning Models

---

陈睿颖

2013544

计算机学院

2023 年 7 月 4 日

目录

1 原文链接

Systematic Evaluation of Privacy Risks of Machine Learning Models

2 动机

本篇论文旨在解决机器学习模型所面临的隐私风险问题，特别是成员推理攻击的问题。

2.1 成员推理攻击

成员推理攻击是指攻击者试图确定特定输入样本是否在模型的训练中使用过，从而揭示个人的敏感信息。具体来说，它指的是给定机器学习模型和一条数据记录，能够判断该条数据记录是否是机器学习模型中的一部分。机器学习模型通常的目标是在模型训练过程中让模型的预测误差最小，这就导致了机器学习模型在训练数据集上的表现会比其他数据上表现更好。这种特性就导致了成员推理攻击的出现。

当攻击者获得一条数据记录和对机器学习模型的访问权利时，如果他能够正确判断该条数据记录是否是模型训练数据集的一部分，则认为成员推理攻击成功。

成员推理攻击大致可分为黑盒攻击和白盒攻击。

1. 黑盒成员推理攻击

黑盒成员推理攻击是一种攻击机器学习模型的方法，攻击者在不了解模型内部结构的情况下，试图确定特定输入样本是否在模型的训练中使用过。攻击者可以通过向模型输入大量的查询样本，并观察模型的输出结果来进行攻击。具体来说，攻击者可以通过观察模型的输出结果，来推断出某个输入样本是否被用于训练模型。这种攻击方法通常需要大量的查询样本，并且需要对模型的输出结果进行分析和推断，因此攻击者需要具备一定的专业知识和技能。

在黑盒成员推理攻击中，攻击者通常不了解模型的内部结构和参数，因此无法直接访问模型的训练数据。攻击者只能通过向模型输入查询样本，并观察模型的输出结果来进行攻击。为了提高攻击的准确性，攻击者通常需要使用一些机器学习技术，如分类器和聚类算法等，来对模型的输出结果进行分析和推断。黑盒成员推理攻击是一种比较常见的攻击方法，因为攻击者可以在不了解模型内部结构的情况下进行攻击，从而更容易地窃取用户的隐私信息。

2. 白盒成员推理攻击

白盒成员推理攻击是一种攻击机器学习模型的方法，攻击者在了解模型的内部结构和参数的情况下，试图确定特定输入样本是否在模型的训练中使用过。与黑盒成员推理攻击不同，白盒成员推理攻击可以更深入地分析和利用模型的内部信息，从而更准确地进行攻击。

在白盒成员推理攻击中，攻击者可以访问模型的内部结构、权重和训练数据等敏感信息。这使得攻击者能够更好地理解模型的行为和决策过程，并通过分析模型的内部信息来推断出某个输入样本是否被用于训练模型。攻击者可以通过观察模型的内部状态、梯度信息和预测结果等来进行攻击。此外，攻击者还可以使用更复杂的攻击模型，如神经网络等，来提高攻击的准确性和效果。

白盒成员推理攻击通常需要更高的技术要求和专业知识，因为攻击者需要了解模型的内部结构和参数，并能够分析和利用这些信息来进行攻击。然而，一旦攻击者成功进行了白盒成员推理攻击，他们可以更准确地确定特定输入样本是否在模型的训练中使用过，从而进一步侵犯用户的隐私。因此，对于机器学习模型的隐私保护来说，防御白盒成员推理攻击至关重要。

2.2 背景

如今，越来越多的大型互联网公司将机器学习作为其云平台上的一项服务。模型和训练算法的具体细节面对拥有数据的用户是隐藏的，模型的类型根据最终预测的效果自适应选择，且服务提供商对模型是否会过拟合考虑较少。通常，服务商仅仅收集用户所提供的数据，并将数据上传到云端后训练模型。外界访问时仅能获得服务所提供的模型的 API。

为了解决这些隐私风险，本篇论文深入研究成员推理攻击的原理和方法，并提出有效的防御策略来保护机器学习模型的隐私，以确保机器学习模型的安全性和隐私性。通过对成员推理攻击的分析和实验验证，本篇论文旨在提高对成员推理攻击的认识，并为研究人员和从业者提供有关隐私保护的指导和建议。

3 贡献与方法

本篇论文贡献主要分为两个方面。第一部分，提出了一种基于度量的攻击，将其与现有的基于神经网络的成员推理攻击结合起来。

在第二部分中，定义了一个称为隐私风险分数的指标，用来估计每个样本成为训练集中成员的可能性，提供更加细粒度的隐私风险分析。

3.1 基于度量的攻击

该方法通过度量目标模型对输入样本的预测不确定性来判断输入样本是否被用于训练模型。具体来说，该方法使用预测熵和预测方差等度量指标来评估目标模型对输入样本的预测不确定性。如果目标模型对某个输入样本的预测不确定性较低，那么攻击者就可以推断出该样本可能被用于训练模型，从而进行成员推理攻击。

与现有的基于神经网络的成员推理攻击不同，该方法不需要训练自己的神经网络模型来进行攻击，而是直接利用目标模型的输出结果来进行攻击。这种基于度量的攻击方法具有较高的攻击准确性和鲁棒性，可以有效地对抗目标模型的防御措施。此外，该方法还可以与现有的基于神经网络的成员推理攻击结合起来，形成一种更加全面和有效的攻击策略。

当前的基于度量的攻击主要可分为依据正确性、置信度和熵三个方面。在攻击中，将这些度量与某些阈值进行比较，以推断输入样本是成员还是非成员。

3.1.1 基于预测正确性的成员推理攻击

在基于预测正确性的成员推理攻击中，能够在目标模型泛化误差较大的情况下取得成功。

当目标模型无法很好的进行泛化时，就会存在能够正确预测训练数据却无法很好预测测试数据的情况。

因此，可以简单的根据输入样本是否能够正确预测，也就是公式中的 $F(x)$ 能否等于 y，来推断某个样本为是否为成员。

$$I_{\text{corr}}(F,(\mathbf{x},y)) = \mathbb{1}\left\{\underset{i}{\text{argmax}} F(\mathbf{x})_i = y\right\}$$

3.1.2 基于预测置信度的成员推理攻击

目标模型是通过最小化训练数据的预测损失来训练的，那么训练样本的预测置信度也就接近于 1，而在测试样本的预测中，即使能够正确分类，置信度也会相比训练数据有所减少。

$$I_{\text{conf}}(F,(\mathbf{x},y)) = \mathbb{1}\{F(\mathbf{x})_y \geq \tau_y\}$$

因此可以利用这种特性进行成员推理攻击，如果样本的预测置信度大于之前设定好的阈值，就认为该样本为成员，否则为非成员。

从之前影子模型的建立中得到启发，可以为不同的类标签设定不同的置信度阈值来改进该方法。通过建立影子模型，在影子模型的训练集和测试集中得到的不同置信度，从而得到针对不同类的阈值。

3.1.3 基于预测熵的成员推理攻击

由于整个训练过程是最小化训练数据的预测损失，那么训练样本的预测输出向量应该接近一个 one-hot 向量，只有所属类为 1，其余值均为 0，预测熵应接近 0。

在测试样本中，即使能够正确分类，最后的预测向量也不会像训练集中像一个 one-hot 向量，而是会有相对较大的预测熵。

与基于置信度的成员推理攻击类似，当特定类的预测熵大于设定阈值时，被认为该样本是成员，否则为非成员。

$$I_{\text{entr}}\left(F,(\mathbf{x},y)\right) = \mathbb{1}\left\{-\sum_i F(\mathbf{x})_i \log\left(F(\mathbf{x})_i\right) \leq \hat{\tau}_y\right\}$$

但是基于预测熵的方法存在一个十分严重的缺陷：预测熵并不包含有关真实标签的任何信息，也就是说虽然当分类结果正确时，熵值为 0，但当分类结果完全错误，被分为另一类时，预测熵值也会为 0。

为了解决这个问题，就需要对原有的预测熵进行适当的改进。

首先新的预测熵需要保证能够随着正确标签 $F(x)_y$ 的预测概率单调递减；另外也应该随着任何其他错误标签 $F(x)_i$ 的预测概率单调递增。

$$\begin{aligned}\text{Mentr}(F(\mathbf{x}),y) = &-\left(1-F(\mathbf{x})_y\right)\log\left(F(\mathbf{x})_y\right)\\ &-\sum_{i\neq y} F(\mathbf{x})_i \log\left(1-F(\mathbf{x})_i\right)\end{aligned}$$

这两点要求对于区分训练数据和测试数据的作用是显而易见的。当正确分类的概率为 1 时，修正熵为 0，而当错误分类的概率为 1 时，修正熵为无穷大。

对预测熵进行合理的修正后，可采用与之前的度量的类似的方式。

为不同的类别设置不同的阈值，通过阴影训练技术进行学习。当需要判定的修正预测熵小于预设阈值，认为是成员，否则认为是非成员。

3.2 当前存在的防御模型的评估

除了使用基于度量的攻击外，本篇论文还提出了另外两种方法来评估当前存在的防御模型。

3.2.1 早停法 Early Stopping

早停法的提出是由于模型在随着训练次数的增加后，训练误差和测试误差都会逐渐减小，但由于过拟合或对训练数据的记忆的影响，目标模型变得容易受到成员推理攻击。因此，可以考虑使用早停法来评估防御方法的性能。

由于早停法可以使用更短的时间和迭代次数来达到与防御措施相似的模型预测性能，如果与防御措施具有相近的对成员推理攻击的抵抗能力，那么就可以认为防御措施是无效的，这样反而会耗费时间和资源。

3.2.2 自适应攻击 Adaptive Attacks

过去的防御方法往往没有考虑自适应攻击的存在。也就是说对手了解防御机制，并能够对防御模型实行自适应攻击，非自适应攻击的完美防御性能并不意味着该防御方法是有效的。

3.3 细粒度的隐私风险分析

先前的工作通常是对样本的隐私风险进行总体评估。但是机器学习模型的学习效果最终往往会因为训练数据的不同而产生变化。这就说明了不同的样本针对隐私风险可能存在着异质性。

由于隐私风险是由于模型的成员与非成员之间的预测行为区别产生的，这就需要对单个样本进行细粒度的隐私风险分析，从而了解哪种特征的样本会具有高隐私风险。

将隐私风险评分如下定义：

$$r(\mathbf{z}) = P\left(\mathbf{z} \in D_{\text{train}} \mid O(F, \mathbf{z})\right)$$

隐私风险被定义为一个后验概率，表示在观察到目标模型在该样本上的行为后，输入样本来自训练集的后验概率。在使用贝叶斯公式对其进行展开后，整个公式可以化简为：

$$r(\mathbf{z}) = \frac{P\left(\mathbf{z} \in D_{\text{train}}\right) \cdot P\left(O(F, \mathbf{z}) \mid \mathbf{z} \in D_{\text{train}}\right)}{P(O(F, \mathbf{z}))}$$

其中 $P(O(F, \mathrm{z})) = P\left(\mathrm{z} \in D_{\text{train}}\right) \cdot P\left(\mathrm{O}(F, \mathrm{z}) \mid \mathrm{z} \in D_{\text{train}}\right) + P\left(\mathrm{z} \in D_{\text{test}}\right) \cdot P\left(\mathrm{O}(F, \mathrm{z}) \mid \mathrm{z} \in D_{\text{test}}\right)$

在黑盒成员推理攻击中，$O(F, z)$ 就可以认为是 $F(x)$。

对于先验概率，为了保证成员推理攻击的不确定性最大，让测试集和训练集以相等的概率进行采样，就可以将两个先验概率约掉。

对于条件概率分布，可以通过影子训练技术获取到。

此外，还可以通过训练影子模型来模拟目标模型的行为、根据影子训练和影子测试数据得到影子模型的预测输出以及计算影子模型训练数据和测试数据的条件分布。但为了防止影子模型中的数据大小的局限性和以类相关的限制，这里采用之前实验结果表现较好的修正预测熵的方式进一步近似多维分布。

$$P\left(F(\mathbf{x}) \mid \mathbf{z} \in D_{\text{tr}}\right) \approx \begin{cases} P\left(\text{Mentr}(F(\mathbf{x}), y) \mid \mathbf{z} \in D_{\text{train}}, y = y_0\right), & \text{when} \quad y = y_0 \\ P\left(\text{Mentr}(F(\mathbf{x}), y) \mid \mathbf{z} \in D_{\text{train}}, y = y_1\right), & \text{when} \quad y = y_1 \\ \vdots & \\ P\left(\text{Mentr}(F(\mathbf{x}), y) \mid \mathbf{z} \in D_{\text{train}}, y = y_n\right), & \text{when} \quad y = y_n \end{cases}$$

对于测试数据也使用相似的方法，最终可以得到特定样本的隐私风险得分。

4 思考

4.1 当前工作的不足

本篇论文虽然提出了一些有效的成员推理攻击和防御方法，但仍存在一些不足之处。

首先，论文中所使用的的实验数据集可能不够广泛和多样化，可能无法完全反映真实场景下的成员推理攻击和防御情况。在论文中，作者使用了几个数据集来评估成员推理攻击和防御方法的效果，包括 Texas100、CIFAR-10、Purchase100 和 Location30 数据集。然而，这些数据集可能无法完全反映真实场景下的成员推理攻击和防御情况。例如，CIFAR-10 数据集是常用的图像分类数据集，但在实际应用中，机器学习模型可能会面临更加复杂和多样化的数据类型和应用场景。但这些数据集可能无法涵盖所有可能的隐私泄露情况和攻击方式。因此，为了更好地评估成员推理攻击和防御方法的效果，需要使用更广泛和多样化的数据集，并考虑更多的攻击方式和隐私泄露情况。

其次，本文的攻击和防御方法可能无法适用于所有类型的机器学习模型和数据集，需要进一步的研究和验证。论文中提出的成员推理攻击和防御方法主要基于神经网络模型和常见的数据集，然而，不同类型的机器学习模型和数据集可能具有不同的特点和隐私泄露风险，例如，一些机器学习模型可能具有不同的结构和参数设置，可能需要针对不同的模型进行不同的攻击和防御方法。

此外，本文的成员推理攻击和防御方法可能会对模型的性能和准确性产生一定的影响，需要在攻击和防御效果和模型性能之间进行权衡。例如，在进行成员推理攻击时，攻击者可能需要访问模型的中间表示或梯度信息，这可能会增加模型的计算和存储开销，从而降低模型的性能和效率。此外，一些防御方法可能会对模型的训练过程或模型结构进行修改，从而影响模型的准确性和泛化能力。

因此，在进行成员推理攻击和防御时，需要在攻击和防御效果和模型性能之间进行权衡。具体来说，需要考虑以下几个方面：

- 攻击和防御效果：需要评估成员推理攻击和防御方法的效果，包括攻击准确性、防御效果、误判率等指标，以确定是否能够有效地保护隐私。
- 模型性能：需要评估成员推理攻击和防御方法对模型的性能和准确性的影响，包括计算和存储开销、训练时间、模型准确性等指标，以确定是否能够满足实际应用的需求。
- 应用场景：需要考虑成员推理攻击和防御方法的应用场景和需求，以确定是否需要更强的隐私保护或更高的模型性能。

最后，本文的成员推理攻击和防御方法可能无法完全解决成员推理攻击带来的隐私问题，需要进一步的研究和探索。

4.2 未来的工作

基于上述当前研究中可能存在的问题，针对未来可能需要进行的工作，我有如下想法：

1. 使用更广泛和多样化的数据集：为了更好地评估成员推理攻击和防御方法的效果，可以使用更广泛和多样化的数据集，涵盖不同类型的数据和应用场景。这样可以更全面地了解不同数据类型和隐私泄露情况对成员推理攻击和防御的影响。
2. 考虑不同类型的机器学习模型和数据集：进一步的研究可以考虑不同类型的机器学习模型和数据集，以适应更广泛的应用场景。这包括针对不同模型结构和参数设置进行攻击和防御方法的研究，以提高方法的通用性和适应性。

3. 在攻击和防御效果与模型性能之间进行权衡：需要更深入地研究成员推理攻击和防御方法对模型性能和准确性的影响，并在攻击和防御效果与模型性能之间进行权衡。这样可以找到一个平衡点，既能保护隐私，又能保持模型的性能和效率。

4. 进一步探索隐私保护方法：成员推理攻击是一个复杂的问题，需要进一步探索和研究不同的隐私保护方法。可以考虑使用不同的隐私增强技术，如差分隐私、同态加密等，以提供更强的隐私保护。

5. 考虑实际应用需求和场景：在研究成员推理攻击和防御方法时，需要考虑实际应用的需求和场景。不同的应用可能对隐私保护和模型性能有不同的要求，因此需要根据具体应用的需求来选择合适的方法和策略。

总之，未来的工作可以着重于使用更广泛和多样化的数据集进行评估、考虑不同类型的机器学习模型和数据集、在攻击和防御效果与模型性能之间进行权衡、进一步探索隐私保护方法，并根据实际应用需求和场景进行定制化研究。这样可以不断改进和完善成员推理攻击和防御方法，提高隐私保护的效果和可靠性。

4.3 个人体会与心得

本篇论文的研究方向可以抽象地理解为通过黑盒模型获取训练数据集；在未来，如果越来越多的机器学习模型用在了程序分析上，获取原始训练数据则很有可能。考虑到程序分析所使用的数据可能较一般分类学习的数据机密性更强，这就存在某些危险发生的可能。

首先，就数据隐私的重要性而言，随着机器学习模型在程序分析中的广泛应用，特别是在处理更机密和敏感的数据时，数据隐私的保护变得尤为重要。程序分析数据可能包含用户的敏感信息、商业机密或个人隐私等。因此，对于这些数据的隐私保护需要引起足够的重视。

为了应对黑盒模型攻击的风险，有必要加强数据保护措施。这包括在数据收集、存储和处理过程中采取适当的安全措施，例如数据加密、访问控制、差分隐私等。此外，可以采用模型压缩、模型集成等方法，减少对原始训练数据的依赖，从而降低隐私泄露的风险。

从社会的角度上看，为保护个人和组织的数据隐私，建立法律和监管框架也具有重要意义。这就包括制定隐私保护法规、建立数据安全标准和认证机制等，以确保机器学习模型在程序分析领域的使用符合隐私保护的最佳实践和法律要求。

此外，提高人们对数据隐私重要性的认识和教育也是至关重要的。个人和组织需要了解隐私风险，并采取适当的措施来保护自己的数据。研究人员和开发者也应该关注隐私保护，并在设计和实现机器学习模型时考虑隐私问题。

总之，我们需要意识到未来机器学习模型在程序分析领域的应用可能带来的隐私风险。而社会各界应思考应对策略，尽可能确保机器学习在程序分析中的应用能够为社会带来益处，同时保护用户和组织的隐私权益。

5 参考文献

1. Nasr M , Shokri R , Houmansadr A . Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning[C]// IEEE Symposium on Security and Privacy (SP). IEEE, 2019.

2. Shokri R , Stronati M , Song C , et al. Membership Inference Attacks against Machine Learning Models[J]. 2017 IEEE Symposium on Security and Privacy (SP), 2017.

3. Song L , Mittal P . Systematic Evaluation of Privacy Risks of Machine Learning Models[J]. 2020.

4. Nasr M, Shokri R, Houmansadr A. Machine learning with membership privacy using adversarial regularization[C]//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. 2018: 634-646.

5. Jia J, Salem A, Backes M, et al. Memguard: Defending against black-box membership inference attacks via adversarial examples[C]//Proceedings of the 2019 ACM SIGSAC conference on computer and communications security. 2019: 259-274.