机器学习中差分隐私的数据共享及发布: 技术、应用和挑战

胡奥婷[1], 胡爱群[2,3], 胡 韵[1], 李古月[1,3], 韩金广[4]

[1] 东南大学网络空间安全学院 南京 中国 211189
[2] 东南大学信息科学与工程学院 南京 中国 210096
[3] 网络通信与安全紫金山实验室 南京 中国 211111
[4] 南京财经大学江苏省电子商务重点实验室 南京 中国 210023

摘要 近年来, 基于机器学习的数据分析和数据发布技术成为热点研究方向。与传统数据分析技术相比, 机器学习的优点是能够精准分析大数据的结构与模式。但是, 基于机器学习的数据分析技术的隐私安全问题日益突出, 机器学习模型泄漏用户训练集中的隐私信息的事件频频发生, 比如成员推断攻击泄漏机器学习中训练的存在与否, 成员属性攻击泄漏机器学习模型训练集的隐私属性信息。差分隐私作为传统数据隐私保护的常用技术, 正在试图融入机器学习以保护用户隐私安全。然而, 对隐私安全、机器学习以及机器学习攻击三种技术的交叉研究较为少见。本文做了以下几个方面的研究: 第一, 调研分析差分隐私技术的发展历程, 包括常见类型的定义、性质以及实现机制等, 并举例说明差分隐私的多个实现机制的应用场景。初次之外, 还详细讨论了最新的 Rényi 差分隐私定义和 Moment Accountant 差分隐私的累加技术。其二, 本文详细总结了机器学习领域常见隐私威胁模型定义、隐私安全攻击实例方式以及差分隐私技术对各种隐私安全攻击的抵抗效果。其三, 以机器学习较为常见的鉴别模型和生成模型为例, 阐述了差分隐私技术如何应用于保护机器学习模型的技术, 包括差分隐私的随机梯度扰动(DP-SGD)技术和差分隐私的知识转移(PATE)技术。最后, 本文讨论了面向机器学习的差分隐私机制的若干研究方向及问题。

关键词 隐私保护; 差分隐私; 机器学习; 数据共享
中图法分类号 TP309.2 DOI 号 10.19363/J.cnki.cn10-1380/tn.2022.07.01

Differentially Private Data Sharing and Publishing in Machine Learning: Techniques, Applications, and Challenges

HU Aoting[1], HU Aiqun[2,3], HU Yun[1], LI Guyue[1,3], HAN Jinguang[4]

[1] School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China
[2] School of Information Science and Engineering, Southeast University, Nanjing 210096, China
[3] Purple Mountain Laboratories for Network and Communication Security, Nanjing 211111, China
[4] Key Laboratory of Electronic Commerce of Jiangsu Province, Nanjing University of Finance and Economics, Nanjing 210023, China

Abstract Recently, Machine Leaning (ML) -based data analysis and data publishing techniques have become hot topics. Compared to traditional data analysis techniques, machine learning enjoys accurate results in analyzing data structure and pattern. However, the privacy leakage issues of machine learning have become increasingly prominent. Incidents of the output predictions of machine learning models leaking users' private information of training data happen frequently. For instance, Membership Inference Attacks (MIA) leaks the participation of machine learning training data by only observing the predictions of models. With the same information, Attribute Inference Attack (AI) leaks the private attributes of machine learning training data. Differential Privacy (DP), a de facto standard for achieving privacy, is trying to incorporate machine learning technology to protect user privacy. However, as the intersection of privacy-preserving technology, machine learning technology, and machine learning attacks, comprehensive researches on this area are relatively rare. In this paper, the following researches are carried out: first, we conduct an in-depth investigation and analysis of the development process of differential privacy, including common types of definitions, properties, and implementation mechanisms, followed by concrete examples to illustrate different scenario to implement different variations of differential privacy. Besides, the analysis also includes state-of-the-art variations, called Rényi Differential Drivacy (RDP) and Moment Accoun-

通讯作者: 胡爱群, 博士, 教授, Email: aqhu@seu.edu.cn
本课题得到国家自然科学基金青年科学基金项目(No. 61801115)资助
收稿日期: 2021-04-27; 修改日期: 2021-06-02; 定稿日期: 2022-05-11

tant (MA) privacy composition technology. Second, we discuss in detail the threat model, the common privacy-related attacks and differential privacy defenses in the field of machine learning. Third, this paper takes the more common discriminative models and generative models of machine learning as examples, and expounds how differential privacy technology is applied to the protecting machine learning models, including the Differentially Private-Stochastic Gradient Desent (DP-SGD) technology and Private Aggregation of Teacher Ensembles (PATE) technology. Finally, we identify the open problems and research directions with respect to leveraging differentially privacy techniques to protect the privacy of deployed machine learning models.

Key words privacy-preserving; differential privacy; machine learning; data sharing

1 引言

数据分析和发布技术使得数据分析者可以学习大数据的共有规律。其中, 统计信息分析[1-2]和机器学习是热门应用领域。然而, 所有的数据分析任务如不添加合适的隐私保护技术都有可能泄漏个人隐私信息。这导致如今数据拥有者由于担忧个人隐私泄漏问题不愿贡献个人数据供第三方使用。欧洲针对此类问题, 已经出台了《通用数据保护法规》(GDPR)规定第三方数据使用者有权保护个人隐私。

1.1 隐私保护背景

首先, 本文举例描述数据分析任务场景以及可能存在的隐私威胁。图 1 为 Adult 公开数据库的片段截取示例。在 Adult 数据库中, 每一行代表一条个人(隐私)信息。数据分析者想要分析数据库中所包含的模式规律。例如, 统计问题“数据库中有多少人的信息满足属性 P？”属性 P 可以是“年收入超过 50K？”或者“年龄超过 50 岁”, 或者两者的交集。机器学习二分类任务可以是“基于个人的其他信息预测该人的年收入是否超过 50K”。

为了在保护数据拥有者的个人信息的同时允许数据分析者分析数据中暗藏的模式, 传统隐私保护方式有非交互式和交互式两种。其中匿名化为非交互式保护方式。匿名化指数据收集者把能表示个人身份信息的唯一识别号(例如身份证号, 学号, 姓名等)从原始数据库中去除再发布。然而, Sweeney[3]提出 87%的美国人可以通过邮编、出生日期和性别这三个组合属性唯一识别, 这暗示仅仅去除唯一识别号不足以保护个人身份不被泄漏。随后, Narayanan 和 Shmatikov 提出链接攻击(linkage attack)[4]。该攻击通过将一个公共数据库的信息链接到私有数据库从而暴露私有数据库里的隐私属性。为了应对该攻击, k-匿名[3]、l-多样化[5]、t-近似[6]等技术相继提出。但是, 这些攻击或受到背景知识攻击影响, 或缺少严谨量化的隐私定义。这些技术假设数据集中的属性可分类为隐私属性和公共属性。隐私属性需要保护而公共属性可以公开。但根据后来研究表明[7], 隐私属性和公共属性并不存在明显的分界, 因为任何属性组合皆有可能泄漏个人的独有特征规律。这个结论尤其符合如今的大数据环境。

当非交互式数据发布难以两全个人隐私保护和数据分析任务时, 交互式问答成为研究者的新方向。然而, 直接回答关于数据库的统计问答也有可能会泄漏个人隐私, 例如差分攻击。攻击者向某医疗数据库提问“数据库中有多少人患有癌症？”和“有多少除了小明的人患有癌症？”可以直接差分出小明是否患有癌症。

在以上案例场景中, 隐私保护目标是在不违反个人隐私的条件下允许数据分析者学习群体规律。因此, 如何定义个人隐私泄漏至关重要。从信息论的角度上分析, 群体规律的学习必然会导致数据分析者得到更多的信息以猜测个人隐私。例如, 某调查结果“肺癌和吸烟有紧密关系”必然会增强攻击者猜测吸烟人群是否患有肺癌的正确概率。在图 1 中, 某

	Age	WorkClass	EducationNum	MaritalStatus	Occupation	Relationship	Race	Gender	CapitalGain	CapitalLoss	HoursPerWeek	NativeCountry	Income
0	39	State-gov	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	≤50K
1	50	Self-emp-not-inc	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	≤50K
2	38	Private	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	≤50K
3	53	Private	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	≤50K
4	28	Private	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	≤50K

图 1 Adult 数据库片段截取

Figure 1 A fragment of the Adult database

机器学习分类器获得 80%的测试集正确率。然后其预测个人年薪是否超过 50K 的正确率会从原本的50%提高到 80%(假设 income 属性平衡)。这些情况是否能称为隐私泄漏？现有的隐私定义难以回答这类定性问题, 因此需要新的隐私保护定义。

差分隐私(Differential Privacy, DP)定义了“合理的可否认性”[8], 即某条个人信息是否参与调查, 调查结果都维持“大致”相同。这等同于保证攻击者几乎无法察觉某个人的信息是否用于计算调查结果。“大致”是由隐私预算 ε 控制。该参数提供隐私和实用性的折中。在实际应用中, 差分隐私机制向调查结果中加入一定量的噪声。噪声的量由隐私预算 ε 和问题敏感度控制。敏感度度量了两个汉明距离为 1 的数据库回答同一个问题的最大差值。

如今, 差分隐私已经成为执行隐私保护的实际标准。微软[9]、苹果[10]、谷歌[11-12]、美国人口调查局[13]、哈弗大学 PSI 项目[14]等都通过利用该技术分析敏感数据。本文旨在分析差分隐私技术在机器学习领域用于隐私保护的理论与应用。通过剖析差分隐私与机器学习交叉领域技术, 提出该领域存在的问题和可能的解决方向。

1.2 相关研究介绍

近年来有以下与差分隐私相关的综述性分析。在这些综述分析中, Dwork 等人[2]首先给出隐私保护分析中存在的问题以及初步的差分隐私解决方案。Dwork 和 Roth[15]总结了到 2014 年为止差分隐私出现的理论性技术。Sarwate 和 Chaudhuri[7], Ji 等人[16], Goryczka 等人[17]和 Jain 等人[18]分别强调信号处理、机器学习、多方安全计算、大数据中存在的差分隐私问题。Zhu 等人[19]介绍了差分隐私的数据共享和分析, 与本文目标类似。然而近年来, 随着差分隐私技术及机器学习技术的迅速发展, 许多新的理论突破和实践层出不穷。因此本文将涵盖更多新发展的技术和问题。

本文旨在帮助读者迅速了解差分隐私的进化发展历程, 并熟悉差分隐私机制的在机器学习领域的应用。图 2 给出常见的隐私数据分析场景架构, 其中数据拥有者提供敏感数据集; 服务提供者, 例如机器学习服务提供商(Machine Learning as a Service, MLaaS)负责数据分析和以及用户隐私保护; 常规用户旨在获取查询结果, 同时恶意用户可能成为窃取隐私信息的攻击者。

后文结构如下: 第 2 节介绍差分隐私的定义、实现机制、常用性质定理; 第 3 节介绍机器学习领域热门的威胁模型、攻击以及与差分隐私的联系; 第 4 节介绍差分隐私机制在机器学习中两种热门模型: 鉴别模型(discriminative model)以及生成模型(generative model)中的运用; 第 5 节总结差分隐私在机器学习领域应用存在的公开问题和研究方向。

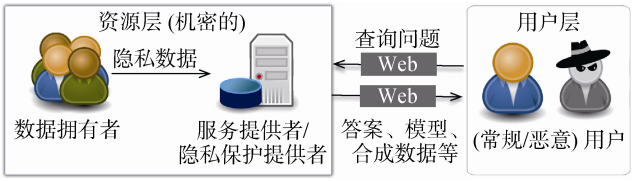


图 2 隐私保护数据分析架构

Figure 2 The framework of private data analysis

2 差分隐私预备知识

2.1 差分隐私

2.1.1 定义

直观上, 差分隐私向数据拥有者保证: 无论某一条数据是否存在于隐私数据集 D 中, 调查结果大致不变。首先, 定义相邻数据集(neighboring datasets)为: 数据集 D 和数据集 D'相差一条个人信息 $x_i \in \mathbb{N}^{|\mathcal{X}|}$, 即汉明距离$d_H(D,D')=1$。差分隐私定义为:

表 1 常用符号

Table 1 Notations

符号	含义	符号	含义
D	隐私数据库	$\mathcal{X}$	隐私数据空间
x_i	一条隐私信息	f	查询
$\mathcal{M}$	随机扰动算法	d_H	汉明距离
ϵ	差分隐私的隐私预算	δ	违反ϵ-DP 的概率
N	数据集大小	exp	指数函数

定义 1. ϵ-差分隐私[8]**.** 对任意相邻数据集 D 和数据集 D'和任意算法结果$\mathcal{S}$, 随机算法$\mathcal{M}$如果满足

$$\frac{\Pr[\mathcal{M}(D)\in\mathcal{S}]}{\Pr[\mathcal{M}(D')\in\mathcal{S}]}\leqslant\exp(\epsilon), \tag{1}$$

则$\mathcal{M}$是ϵ-差分隐私。概率空间由$\mathcal{M}$的随机性算法决定。

在定义 1 中, 隐私预算ϵ是乘性界限的自然对数。小的隐私预算ϵ指数据集 D 和数据集 D'分布距离更小, 即一条隐私数据的信息对结果的影响力十分有限。因此, 小隐私预算ϵ意味着更高的隐私保护力度。

2.1.2 机制

纯粹的差分隐私可以通过拉普拉斯(Laplace)机制、指数机制和随机应答(randomized response)实现。

1) 拉普拉斯机制

拉普拉斯机制[8]旨在向数字查询结果$f:\mathbb{N}^{|\mathcal{X}|}\rightarrow$

$\mathbb{R}^k$添加校准噪声。例如, 在图 1 中, 某计数查询或分数查询为“有多少人的年收入超过 50K？”。为了保证相邻数据集对同一个问题的答案差距足够小, 我们需要度量查询f的敏感度(sensitivity)。在拉普拉斯机制中, 用ℓ_1度量敏感度, 定义如下。

定义 2. ℓ_1敏感度. ℓ_1-敏感度定义了相邻数据集D和D'回答问题f结果的最大差

$$\Delta f = \max_{D,D'} \| f(D) - f(D') \|_1, \tag{2}$$

其中$\|\cdot\|_1$是一阶范数。

根据定义 2, 计数查询的ℓ_1-敏感度为 1。分数查询的ℓ_1-敏感度为$1/N$, N为数据集D的大小.

$$\ln \frac{\exp\left(-\frac{|f(D)|}{\frac{\Delta_1 f}{\epsilon}}\right)}{\exp\left(-\frac{|f(D)+\Delta_1 f|}{\frac{\Delta_1 f}{\epsilon}}\right)} = \frac{|f(D)+\Delta_1 f| - |f(D)|}{\frac{\Delta_1 f}{\epsilon}} \leqslant \frac{\Delta_1 f}{\Delta_1 f/\epsilon} = \epsilon,$$

根据定理 1, 拉普拉斯机制$\mathcal{M}_L$的噪声被缩放为$b = \frac{\Delta f}{\epsilon}$用以满足$\epsilon$-差分隐私。因为噪声大小与隐私损失$\epsilon$成反比, 所以隐私损失$\epsilon$越小, 所添加的的噪声幅度越大, 则查询结果越不准确。 因此, 隐私损失ϵ控制隐私与效用之间的折中。

2) 指数机制

不同于拉普拉斯机制, 指数机制[20]旨在向非数值型查询函数$f: \mathbb{N}^{|\mathcal{X}|} \to \mathcal{R}$结果$r \in \mathcal{R}$添加噪声, 特别是选择最优解。对于有些无法直接向答案中添加噪声的非数值型问题和有一些直接添加噪声会破坏数据可用性的数值型问题, 例如医疗记录、产品名称等查询, 指数机制通过定义效用函数$u: \mathbb{N}^{|D|} \times \mathcal{R} \to \mathbb{R}$将输入输出对映射为一个实数。效用函数$u$定义首选对获得更高的分数。效用函数的敏感度定义如下:

$$\Delta u = \max_{r \in \mathcal{R}} |u(D,r) - u(D',r)|. \tag{4}$$

据此, 差分隐私的指数机制$\mathcal{M}_E(D,u,\mathcal{R})$有如下定理:

定理 2. 指数机制[20]. 指数机制$\mathcal{M}_E(D,u,\epsilon)$以概率$\exp\left(\frac{\epsilon \times u(D,r)}{2\Delta u}\right)$输出答案$r$, u为效用函数, 则机制满足ϵ-差分隐私。

由于指数机制的概率密度函数为$\exp\left(\frac{\epsilon \times u(D,r)}{2\Delta u}\right)$, 则对任意回答$r \in \mathcal{R}$, 有

$$\ln \frac{\exp\left(\frac{\epsilon \times u(D,r)}{2\Delta u}\right)}{\exp\left(\frac{\epsilon \times u(D',r)}{2\Delta u}\right)} = \frac{\epsilon(u(D,r) - u(D',r))}{\Delta u} \leqslant \frac{\epsilon \Delta u}{\Delta u} = \epsilon,$$

因此满足ϵ-差分隐私。

3) 随机应答机制

拉普拉斯机制和指数机制都是在交互式问答的查询结果中添加噪声以满足差分隐私, 而随机应答应用于非交互式场景。它直接向本地个人信息中添加噪声, 然后发布净化的本地信息。该机制适用于收集众包数据(谷歌、苹果, 微软等公司运用该技术收集用户信息, 如输入习惯、表情包喜好等)。随机应答机制[21]早在差分隐私提出之前就已经广泛使用。但是后来证明其满足差分隐私要求。

由于定义 1 中应用乘性距离, 添加拉普拉斯噪声天然满足差分隐私。尺度为b、中心为 0 的拉普拉斯分布的概率密度函数为$\text{Lap}(x|b) = \frac{1}{2b}\exp(-\frac{|x|}{b})$。据此, 拉普拉斯差分隐私机制如下:

定理 1. 拉普拉斯机制[8]. 给定数值查询函数f, 隐私数据库D和拉普拉斯噪声$Y \sim \text{Lap}(\frac{\Delta f}{\epsilon})$, 如下随机扰动函数$\mathcal{M}_L$满足$\epsilon$-差分隐私。

$$\mathcal{M}_L(D,f,\epsilon) = f(D) + Y. \tag{3}$$

定理 1 的合理性可以通过计算$\mathcal{M}_L(D,f,\epsilon)$在相邻数据集上结果的最大差值对数比得出。

回顾图 1, 数据分析者想了解“Adult 数据库中多少比例的人群年收入超过 50K？”。拉普拉斯机制要求有某个可信数据中心先收集所有用户的真实答案, 计算查询结果再添加差分隐私噪声。而在随机应答模式中无需可信数据中心。数据分析者获取查询“某个人年薪是否超过 50K”的扰动答案, 再自行聚合扰动答案以逼近真实答案。具体操作如下: 每个参与回答“年薪是否超过 50K”的人, 都扔一枚硬币, 如果是“字”即诚实回答, 如果是“花”则再扔一枚硬币, 硬币是“字”回答“是”, 硬币是“花”回答“否”。假设真实答案中“是”的比例为q, 那么扰动回答中“是”的比例则为$q' = 0.5 * q + 0.25$。当发布了一批扰动答案后, 数据分析者估计真实“是”比例为$2 * q' - 0.5$。据此, 数据分析者既获得年收入大于 50K 的人数比例, 个人信息得到了保护。在该机制中, 参与问答的用户越多, 数据分析者得到的查询结果越准确。因此当数据量稀少时, 不适合该机制。

根据以上随机扰动案例, 令单个查询函数为f, 随机应答机制为$\mathcal{M}_R(D,f,\epsilon)$, 对任意$x \in D$和其随机应答$r$, 满足

$$\ln \frac{\Pr(x=0|r=0)}{\Pr(x=1|r=0)} = \frac{3/4}{1/4} = \ln 3$$

同理,

$$\ln \frac{\Pr(x=1|r=1)}{\Pr(x=0|r=1)} = \ln 3$$

因此, 随机扰动算法$\mathcal{M}_R(D,f,\epsilon)$满足3-差分隐私。更广泛的, 在二元分布中, 给定随机应答机制

$$\mathcal{M}_R(D,f,\epsilon)=\begin{cases} f(x) & \text{以概率 } q \\ 1-f(x) & \text{以概率 } 1-q \end{cases}, \quad (5)$$

则其满足ϵ-差分隐私, 且$q=\frac{e^{\epsilon}}{1+e^{\epsilon}}$。

2.1.3 性质

差分隐私定义的广泛使用离不开其满足的性质: 合成定理(composition)、对辅助信息的鲁棒性、抗后处理(post-processing)、群差分隐私(Group Differential Privacy, GDP)。这些性质可以保证差分隐私适应于模块化设计。

1) 合成定理

合成定理[15]使得差分隐私可优雅的模块化叠加。多步差分隐私算法作用于同一数据库, 其隐私消耗线性叠加。

定理 3. 合成定理[15]. 令D为隐私数据库, f为查询函数。同时输出扰动结果$\mathcal{M}_1(D,f,\epsilon_1)$, $\mathcal{M}_2(D,f,\epsilon_2)$满足$\epsilon_1+\epsilon_2$差分隐私。

根据定理 3 可知, 重复输出k次对数据库D的查询结果$\mathcal{M}(D,f,\epsilon)$满足$k\epsilon$-差分隐私。因此, 同质(homogeneous)和异质(heterogeneous)差分隐私机制的累加都是线性的。同质指每次使用相同的隐私预算, 异质指每次使用不同的隐私预算。

2) 对辅助信息的鲁棒性

差分隐私的保护效果应不受攻击者掌握的背景知识多少影响[15]。该性质可利用贝叶斯定理比较攻击者先验和后验的攻击优势差。令攻击者对数据集D分布的先验知识为$p(D)$, 扰动机制$\mathcal{M}(D,f,\epsilon)$输出结果为$r$。由定义 1 可知, $p(r|D)/p(r|D')\leqslant\exp(\epsilon)$, 在等式左边使用贝叶斯定理得

$$\frac{\frac{p(D|r)\cdot p(r)}{p(D)}}{\frac{p(D'|r)\cdot p(r)}{p(D')}}=\frac{p(D|r)/p(D'|r)}{p(D)/p(D')}\leqslant\exp(\epsilon).$$

因此, 攻击者先验和后验的攻击优势比值不超过$\exp(\epsilon)$。

3) 抗后处理

抗后处理定理指对某个差分隐私的结果做后续分析不会削弱其差分隐私保护效果。

定理 4. 抗后处理定理[15]. 令$\mathcal{M}:\mathbb{N}^{|\mathcal{X}|}\to\mathcal{R}$是$\epsilon$-差分隐私机制, $g:\mathcal{R}\to\mathcal{R}'$为任意函数。$g\circ\mathcal{M}:\mathbb{N}^{|\mathcal{X}|}\to\mathcal{R}'$依然是$\epsilon$-差分隐私。

该定理保证了, 如果算法$\mathcal{M}$是ϵ-差分隐私, 无论数据分析者如何使用差分隐私的结果, 利如使用算法g对$\mathcal{M}$的结果做进一步数据分析, 都不会降低算法$\mathcal{M}$对隐私数据的保护效果。

4) 群差分隐私

群差分隐私保护“关联的隐私数据”。当多个隐私数据之间存在确定性或者概率性的关联时, 可以将其当作一个群组。群差分隐私定义隐私损失随着群组大小的增大而线性增大。例如, 当某调查中出现一个家庭的多条隐私数据时, 他们的数据属性是高度关联的。他们可能会共享地址、邮编等。因为改变某一条记录的属性对结果产生的影响可能会比原先估计的要大。

定理 5. 群差分隐私[15]. 某差分隐私机制$\mathcal{M}(D,f,\epsilon)$在群组大小为$k$时满足$k\epsilon$-差分隐私。即, 对$d_H\left(D,D'\right)\leqslant k$以及其可能出现的扰动结果$S$

$$\Pr[\mathcal{M}(D)\in S]\leqslant\exp(k\epsilon)\Pr[\mathcal{M}(D')\in S]. \quad (6)$$

该定理用于处理数据库分析中的关联数据集。但是, 后续有研究指出当群数组关系为概率性而非确定性时, 差分隐私会引入过多的噪声[22]。

2.2 近似差分隐私

差分隐私在实际使用中, 存在几种近似定义。近似定义意在降低达成同等隐私损失ϵ而添加的噪声量(对比定义 1)。主流的近似定义有(ϵ,δ)-差分隐私[23], Concentrated 差分隐私(CDP)[24]和 zero- Concentrated 差分隐私(zCDP)[25], Moments Accountant[26], Rényi 差分隐私(RDP)[27]。其中(ϵ,δ)-差分隐私使用最为广泛。

2.2.1 (ϵ,δ)-差分隐私

(ϵ,δ)-差分隐私是历史最久且使用广泛的近似定义。该定义添加参数δ以实现$1-\delta$成功率的ϵ-差分隐私机制, 其中参数δ是一个很小的数。定义如下:

定义 3. (ϵ,δ)-差分隐私[23]. 对任意相邻数据集数据集 D 和数据集 D'和任意算法结果$\mathcal{S}$, 随机算法$\mathcal{M}$如果满足:

$$\Pr[\mathcal{M}(D)\in S]\leqslant\exp(\epsilon)\Pr[\mathcal{M}(D')\in S]+\delta, \quad (7)$$

则$\mathcal{M}$是(ϵ,δ)-差分隐私。概率空间由$\mathcal{M}$的随机性算法决定。

该定义配合高斯机制使用。高斯噪声比拉普拉斯噪声存在更为普遍, 且原始数据集中可能已经包含高斯噪声。另外, 高斯噪声的分布函数比拉普拉斯噪声更加集中。因此添加高斯噪声是更自然的选择。高斯噪声$\mathcal{N}(0,\sigma)$的概率密度函数为$f_g(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{x^2}{2\sigma^2}}$。与拉普拉斯机制类似, 高斯机制也是向查询结果中添加噪声。噪声幅度由问题敏感度与隐私预算决定。不同于拉普拉斯机制的是, 高斯机制使用ℓ_2敏感度度量两个查询结果的最大差别。

定义 4. ℓ_2敏感度[23]. ℓ_2-敏感度定义了相邻数据集D和D'回答查询f结果的最大差:

$$\Delta_2 f = \max_{D,D'} \| f(D) - f(D') \|_2, \tag{8}$$

其中$\|\cdot\|_2$是欧几里德范数。

对于查询函数f，ℓ_2敏感度通常要比ℓ_1敏感度小(特别是敏感度小于 1)。因此, 需要添加的高斯噪声幅度通常要比拉普拉斯噪声小。

定理 6. 高斯机制[23]. 给定数值查询函数f, 隐私数据库D和拉普拉斯噪声$N \sim \mathcal{N}(0,\sigma)$, 如果$c^2 > 2\ln(1.25/\delta)$, $\sigma \geqslant c\Delta_2 f/\epsilon$, 则随机函数$\mathcal{M}_G$满足$(\epsilon,\delta)$-差分隐私。

$$\mathcal{M}_G(D,f,\epsilon) = f(D) + N. \tag{9}$$

与拉普拉斯机制类似, 高斯噪声的标准差设置小于$c\Delta_2 f/\epsilon$, 其中参数c控制不满足ϵ-差分隐私机制的概率不大于δ。感兴趣的读者可以在文献[15]的附录 A 中找到推导过程。

高斯机制的优势有: 第一, 差分隐私所添加的噪声可能与数据集中原本存在的噪声同分布; 第二, 高斯噪声的和依然是高斯噪声(拉普拉斯噪声不是); 第三, 高斯分布概率密度的尾端比拉普拉斯分布降速更快, 即更集中。

其弱点为: 第一, 高斯机制为了降低同等隐私预算下的噪声添加量, 遗留δ概率的完全隐私侵犯。实际使用中, 我们一般将δ设置成远远小于数据库大小倒数的值, 即$\delta \ll 1/N$, 以保证没有用户的隐私会被侵犯。第二, 虽然基于高斯机制的(ϵ,δ)-差分隐私定义与ϵ-差分隐私同样享有对冗余信息的鲁棒性、抗后处理、群差分隐私, 但是其合成定理较为复杂。其中同质(homogeneous)(ϵ,δ)-差分隐私机制$\mathcal{M}$的合成定理[28]见定理 6。但是异质(heterogeneous)差分隐私机制的合成是$\#P$-难度[29-30]。

定理 7. 高级合成定理(advanced composition)[28]. 同质(ϵ,δ)-差分隐私机制$\mathcal{M}$重复输出k次对数据库D的查询结果满足$(\epsilon\sqrt{k\log\left(\frac{1}{\delta'}\right)} + \epsilon(e^{\epsilon}-1)k, k\delta + \delta')$差分隐私。

为了解决合成定理的难题, 后续提出了其他的近似差分隐私定义。

2.2.2 其他近似定义

在(ϵ,δ)-差分隐私之后, 研究者们提出了其他近似定义, 这里列举应用较为广泛的几种:

1) CDP 和 zCDP

Dwork 和 Rothbulum 提出 CDP[24], 其利用了亚高斯尾巴很小的特点以平均隐私损失。CDP 定义比(ϵ,δ)-差分隐私定义弱一些, 但是提供更高的可用性和更优的高斯机制合成定理。其后, Bun 和 Steinke[25]利用 Rényi 差异(Rényi divergence)捕获两分布的差异性以提出改良的 zCDP。

2) Moments Accountant

Abadi 等人[26]提出 Moments Accountant 技术, 用于跟踪深度学习中差分隐私预算的变化量。在第 3 节讲述差分隐私的随机梯度递减算法时会详述。

3) Rényi 差分隐私

2017 年, Mironov 提出 Rényi 散度是差分隐私的一种天然近似形式[27]。ϵ-差分隐私使用最大差异(maximum divergence)来度量两个相邻数据库查询差别。而 Rényi 散度使用参数α放松该最大差异条件。

表 2 中总结了几种差分隐私定义。值得注意的是, Jayaraman 和 Evans 在文献[30]中对比了 Advanced Composition(定理 7)、CDP、zCDP 和 RDP 在实际使用中的隐私预算累加效果, 总体来说, RDP 表现最优。因此在机器学习领域中常采用的差分隐私机制是高斯机制。在追踪隐私预算时, 使用 RDP+ Moments Accountant 技术, 最后将 RDP 中所定义的转化成(ϵ,δ)-差分隐私。据此, 下文简述 RDP 机制原理及其对比(ϵ,δ)-差分隐私的优势。

表 2 差分隐私常见定义

Table 2 Different definitions of differential privacy

名称	定义
ϵ-DP[8]	$\Pr[M(D)\in S] \leqslant \exp(\epsilon)\Pr[M(D')\in S]$
(ϵ,δ)-DP[23]	$\Pr[M(D)\in S] \leqslant \exp(\epsilon)\Pr[M(D')\in S] + \delta$
(ξ,ρ)-zCDP[25]	$D_\alpha(M(D)\|\|M(D')) \leqslant \xi + \rho\alpha$
(α,ϵ)-RDP[27]	$D_\alpha(M(D)\|\|M(D')) \leqslant \epsilon$

理解 RDP 机制首先要理解 Rényi 散度的定义, 以及为什么 Rényi 散度是ϵ-差分隐私的天然泛化形式。

定义 5. Rényi 散度. 对分布$P(x)$和$Q(x)$, $\alpha > 1$条件下的 Rényi 散度定义为

$$D_\alpha(P\|\|Q) \triangleq \frac{1}{\alpha-1}\log \mathbb{E}_{x\sim Q}\left(\frac{P(x)}{Q(x)}\right)^{\alpha} \tag{10}$$

根据定义 5, Rényi 散度在$\alpha \in (1,\infty)$连续。当α逼近 1, $\lim_{\alpha\to 1} D_\alpha(P\|\|Q) = E_{x\sim P}\log\frac{P(x)}{Q(x)}$, 这就是 KL 散度。如果$\alpha$逼近$\infty$, $\lim_{\alpha\to\infty} D_\alpha(P\|\|Q) = \sup_{x\in \mathrm{supp}Q}\log\frac{P(x)}{Q(x)}$, 这就是$\epsilon$-差分隐私。因此$\epsilon$-差分隐私是 Rényi 差异在$\alpha=\infty$的特殊形式。此时, 定义 1 可以重写为: 对任意相邻数据集数据集 D 和数据集 D'和任意算法结果$\mathcal{S}$, 随机算法$\mathcal{M}$如果满足$\mathrm{D}_\infty(\mathcal{M}(\mathrm{D})\|\|\mathcal{M}(\mathrm{D}')) \leqslant \epsilon$, 则为$\epsilon$-差分隐私。当$\alpha \in (1,\infty)$, Rényi 差异就是介于

KL(期望)散度和最大散度中间的度量方式。该程度由α的大小控制。综上, 得出依赖Rényi 散度的Rényi差分隐私定义。

定义 6. Rényi 差分隐私[27]. 对任意相邻数据集数据集 D 和数据集 D'和任意算法结果$\mathcal{S}$, 随机算法$\mathcal{M}$如果满足$D_\alpha(M(D)||M(D')) \leq \epsilon$, 则称为$(\alpha,\epsilon)$-RDP。

根据定义 6, (∞,ϵ)-RDP 就是ϵ-DP。Rényi 差分隐私有以下两个重要性质。

定理 8. (α,ϵ)-RDP 转换成(ϵ,δ)-DP[27]。对任意δ和α, (α,ϵ)-RDP 可以转换成$(\epsilon+\frac{\log 1/\delta}{\alpha-1},\delta)$-DP。

定理 9. RDP 的合成定理. 给定α, 异质(α,ϵ_i)-RDP 机制$\mathcal{M}_i$的累加满足$(\alpha,\sum_i \epsilon_i)$-RDP。

RDP 定义与(ϵ,δ)-DP 相比较有什么优缺点? 首先, RDP 依然享有冗余信息的鲁棒性、抗后处理、群差分隐私。更重要的是, 同ϵ-DP, 它能够优雅的累加异质差分隐私算法的隐私预算(见定理 9), 但是这在(ϵ,δ)-DP 定义中是难题(见 2.2.1 分析)。除此之外, 依据定理 8, (α,ϵ)-RDP 在累加结束后可便捷转换成(ϵ,δ)-DP 定义。之所以需要转换, 是因为(ϵ,δ)-DP 定义更容易理解且δ具有物理意义(即不满足ϵ-DP 的概率)。

RDP 的实现机制? RDP 有多种实现机制, 其中高斯机制较为常用。当查询f的ℓ_2-敏感度为 1 时, 添加$N\sim\mathcal{N}(0,\sigma)$噪声的高斯机制$\mathcal{M}_G$满足$(\alpha,\alpha/(2\sigma^2))$-RDP。该证明可以由直接计算两个高斯分布的$D_\alpha(\mathcal{N}(0,\sigma),\mathcal{N}(\mu,\sigma))$ 的Rényi 散度得出。

综上, 本节总结了差分隐私常用的定义形式(表 2)和实现机制以及其利弊(表 3), 下一节我们详述机器学习中存在的威胁模型及与差分隐私的联系。

表 3 差分隐私实现机制

Table 3 Mechanisms for Differential Privacy

实现机制	DP 定义	实现形式	优缺点
Laplace 机制	ϵ-DP	$\mathcal{M}_{\mathrm{L}}(D,f,\epsilon)=f(D)+\mathrm{Lap}(\frac{\Delta f}{\epsilon})$	优点: 实现方便, 满足严格差分隐私定义; 线性累加。缺点: 可能会添加过多噪声(比高斯机制)。
随机扰动	ϵ-DP	$\mathcal{M}_{\mathrm{R}}(D,f,\epsilon)=\begin{cases} f(x) & \text{以概率 } p \\ 1-f(x) & \text{以概率 } 1-p \end{cases}$	优点: 直接扰动原始数据集而非查询结果, 适合无可信中心场景, 利如众包数据收集分析; 实现简单。缺点: 单个隐私数据添加较大噪声量, 应用场景有限。
指数机制	ϵ-DP	$\mathcal{M}_{\mathrm{E}}(D,u,\epsilon)=r$ 以概率 $\exp(\frac{\epsilon\times u(D,r)}{2\Delta u})$	优点: 实现简单, 但需设定合适的效用函数; 可返回非数值查询结果; 适合对噪声敏感的查询问题。缺点: 应用场景有限。
高斯机制	(ϵ,δ)-DP	$\mathcal{M}_{\mathrm{G}}(D,f,\epsilon)=f(D)+\mathcal{N}(0,\sigma)$ 其中$c^2>2\ln(1.25/\delta)$, $\sigma\geqslant c\Delta_2 f/\epsilon$	优点: 高斯噪声在信号中更常见, 且其叠加依然是高斯噪声; 在同等隐私预算下比拉普拉斯机制添加噪声量小; 缺点: 异质差分隐私机制的合成是#P-难度(采用 RDP 定义可解决)。

3 机器学习中的隐私威胁模型与攻击

随着机器学习的深入发展, 深度学习已经成为寻找数据规律的重要手段。一般的, 机器学习通过建立模型、优化损失函数来拟合数据。但是, 机器学习模型如果用来拟合个人敏感数据, 例如医疗数据、人口普查信息、学校数据、银行数据等, 会对个人隐私保护提出挑战。当攻击者获取机器学习模型后, 模型输出特性可能泄漏训练数据的隐私信息。例如某个人的信息是否存在于隐私数据集中(成员猜测攻击), 或者猜测某个人的隐私属性(属性猜测攻击)。

3.1 隐私威胁模型

讨论攻击之前, 首先需要定义威胁模型。威胁模型可以用来度量攻击者能力及其抵抗方法的有效性。具体包括以下三个方面: 攻击者的目标、知识和能力。攻击者的目标根据不同攻击类型有所不同, 我们将在 3.2 节详述。攻击者的知识和能力在机器学习领域主要体现在以下两个方面: 模型知识和数据集知识。

模型知识: 白盒子M^{W}和黑盒子M^{B}。白盒子攻击者掌握目标机器学习模型的模型架构和模型参数。黑盒子攻击指的是攻击者只能接入模型 API, 即查询模型并获取返回的预测结果(可能包含预测结果的概率), 但是不知道模型参数。许多黑盒子模型假设攻击者知道目标模型的架构。因为当攻击者使用现有机器学习及服 MLaaS 时,其能够复现目标模型的架构。

数据集知识: 攻击者是否拥有额外数据集。攻击者能力由强到弱依次分为: (1)可获取部分训练集$D_{\mathrm{aux}}^{\mathrm{P}}$; (2)获取同分布数据集$D_{\mathrm{aux}}^{\mathrm{S}}$; (3)无额外数据集$D_{\mathrm{aux}}^{\mathrm{N}}$。第一种情况下, 攻击者获得部分训练集; 第二种情况下, 攻击者获取与训练集同分布但不相交数据集(例如, 对抗生成网络生成的人工合成数据集,

又叫影子数据集); 第三种情况, 攻击者没有任何额外数据集。

综合以上两类攻击者知识, 共有 6 种可能的威胁模型: $\langle M^{\mathrm{B}}, D_{\mathrm{aux}}^{\mathrm{P}}\rangle$, $\langle M^{\mathrm{B}}, D_{\mathrm{aux}}^{\mathrm{S}}\rangle$, $\langle M^{\mathrm{B}}, D_{\mathrm{aux}}^{\mathrm{N}}\rangle$, $\langle M^{\mathrm{W}}, D_{\mathrm{aux}}^{\mathrm{P}}\rangle$, $\langle M^{\mathrm{W}}, D_{\mathrm{aux}}^{\mathrm{S}}\rangle$, $\langle M^{\mathrm{W}}, D_{\mathrm{aux}}^{\mathrm{N}}\rangle$。

3.2 隐私威胁攻击

常见的在机器学习领域与隐私保护(privacy protection)相关的攻击分为以下几类: 成员猜测攻击(membership inference attack), 模型反演攻击(model inversion attacks), 属性猜测攻击(attribute inference attack), 模型窃取攻击(model stealing attack), 无意识记忆(unintended memorization)。值得注意的是, 对抗样本攻击(adversarial samples)[31]是另一类较为热门的威胁到机器学习模型安全的议题, 但是属于模型安全领域(model security), 与隐私保护无关, 因此不在本文讨论范围内。

3.2.1 成员猜测攻击

成员猜测攻击[32], 也叫追踪攻击, 攻击目标是猜测某条个人信息是否在目标模型的训练集中。成员猜测攻击导致的直接后果是当某个数据集本身具有敏感属性, 则探知成员信息存在性可直接泄漏其隐私属性。例如, 用某一癌症数据库作为训练集训练模型, 当攻击者探知A属于该训练集, 则可知A患有癌症。另外, 成员猜测攻击还可以用于检测非法数据使用[33]。成员猜测攻击的原理是目标模型对于训练集和其他数据的表现不同。例如, 目标分类器在预测阶段, 对训练集会表现出更高的置信度, 而对其他数据则表现较低的置信度。因此, 过拟合的模型比泛化的模型更容易受到成员猜测攻击影响。根据攻击者掌握的目标模型知识不同, 分为黑盒子攻击[32, 34]和白盒子[35]攻击。在白盒子攻击下, 攻击者通常利用模型梯度向量进行攻击; 在黑盒子攻击下, 攻击者通常利用对标签预测的置信度向量进行攻击。最新的文献中[36-37], 也有提出仅仅需要黑盒子模型预测的标签(无需置信度)也可以发起有效成员猜测攻击。

差分隐私技术可以有效防止成员猜测攻击。从定义上可以看出, 差分隐私限制某一条信息对查询结果的影响, 因此限制了成员猜测攻击的成功率上限。大多研究也表明[32, 38-39], 差分隐私可以有效防止成员猜测。

3.2.2 模型反演攻击

模型反演攻击, 指的是攻击者拥有白盒子模型, 意图重构部分训练集, 或者部分类表征。例如, Fredikson 等人[40]依据目标黑盒子模型和部分公共属性猜测个体基因型。随后, Fredrikson 等人[41]提出模型反演可以恢复部分训练集面部信息。但是, 该攻击只有当类成员近似时, 才能发挥作用, 例如 MNIST 数据集, 人脸识别数据集, 等。此后, Hitaj 等人[42]提出在合作深度学习(collaborative deep learning)模式下的模型反演攻击。攻击者利用多方在线学习所传递的更新梯度以训练自己的 GAN 生成模型[43], 使得该生成模型能够恢复部分人脸。近期, Zhang 等人[44]提出通过 GAN 生成模型以及公共信息反演深度学习模型以人工合成(恢复)训练集图片。综上, 模型反演攻击与下文即将提到的属性推测攻击有相似之处。但是模型反演攻击所恢复的部分敏感属性可能是和标签高度相关的。其攻击的成功性也许依赖于目标模型达到预期的泛化能力[45], 因此模型反演攻击是否需要防护在隐私保护领域的尚有争议。

3.2.3 属性猜测攻击

属性猜测攻击旨在猜测与鉴别模型任务无关的属性值。比如, 某用于预测年龄的鉴别模型, 攻击者可以从中推测出种族属性。或者, 某张 Bob 的人脸用于学习分类性别, 但是攻击者却可以用其判断 Bob 这张照片中其他人是否戴眼镜。该攻击说明, 某些模型过度学习(overlearning)以致模型信息中包含了许多与原始任务无关的信息。Melis 等人[45]提出在合作模型学习中利用在线学习中的更新信息推测隐私属性。比如, 泄漏合作学习中的参与者在每轮更新的参与情况。防止过度学习并不十分容易。现有抵抗措施主要包括学习训练集的替代表征[46] $\boldsymbol{x} \to z$, 使得其与目标最为相关的部分属性, 并尽量降低无关属性的信息量s。但是, Song 和 Shmatikov[47]的研究表示, 即使训练集已经被压缩为不泄漏隐私的替代表征z, 逆向工程仍然可以轻易的估计出逆向函数$T(z) \to \boldsymbol{x}_{\mathrm{approx}}$, 再用$\boldsymbol{x}_{\mathrm{approx}}$训练攻击模型以估计隐私信息量$s$。即, 作者认为过度学习可能是本能的, 在不过度学习的条件下满足目标任务分类, 也许不太可能。

3.2.4 模型窃取攻击

模型窃取攻击指的是, 攻击者只能提问黑盒子目标模型 API(例如 Google, Amazon, BigML 等), 推测出目标模型的参数, 以使得窃取的模型和目标模型具有相似特性。模型窃取攻击有严重后果。其一, 许多模型是私有财产并且按照查询问题次数计费, 因此偷取模型侵犯模型所有者的权利; 其二, 模型窃取攻击给许多需要白盒子知识的攻击(比如属性知识攻击, 对抗生成样本攻击等)提供便利。Tramèr 等人[48]提出用解方程攻击偷取逻辑回归模型(logistic regression), 用寻找路径攻击偷取决策树模型。尽管模型窃取攻击没有直接泄漏用户隐私, 但是其侵犯模型所

有者权益, 并提高了攻击者在其他类型攻击者的模型知识和取胜概率。

3.2.5 无意识记忆

无意识记忆由 Carlini 等人[49]在 2019 年提出。作者发现语言生成模型可能会无意间暴露训练集隐私。比如, 某个文字自动补全模型会自动补全隐私信息。例如, 输入“我的社交密码是 078-”, 生成模型此时会自动补全后半部分“-05-1120”。隐私信息的自动补全说明模型不仅学习到语言模式, 还记住了部分训练集中存在的隐私信息。更有甚者, 作者发现, 模型记忆并非源自模型过拟合训练集, 因为模型在训练初期已经发生记忆, 并非在后期。因此某些预防过拟合的正则化方法, 例如提前停止训练(early- stopping)或者 dropout 等并不能有效防止模型记忆。幸运的是, 作者发现, 只需要很小隐私预算ϵ的差分隐私技术即可以有效防止模型记忆。

3.3 差分隐私抵抗机制

差分隐私机制从定义上防止成员猜测攻击, 模型记忆, 并弱化属性猜测攻击。但是, 其对模型反演攻击和模型窃取攻击的弱化效果不明显。具体可参考 Liu 等人[38]的研究。表 4 总结了以上提到的五种攻击以及差分隐私对它们的抵抗能力。

表 4 机器学习模型的隐私威胁

Table 4 The attacks that threat the machine learning models

攻击类型	DP 机制	攻击目标	目标模型	攻击者知识	代表文献
成员知识攻击	有效	猜测某条信息是否属于训练集	鉴别模型	$\langle M^{\mathrm{B}}, D_{\mathrm{aux}}^{\mathrm{S}}\rangle$	[32, 50]
				$\langle M^{\mathrm{B}}, D_{\mathrm{aux}}^{\mathrm{P}}\rangle, \langle M^{\mathrm{B}}, D_{\mathrm{aux}}^{\mathrm{N}}\rangle$	[50]
				$\langle M^{\mathrm{W}}, D_{\mathrm{aux}}^{\mathrm{S}}\rangle, \langle M^{\mathrm{W}}, D_{\mathrm{aux}}^{\mathrm{P}}\rangle$	[35]
			生成模型	$\langle M^{\mathrm{B}}, D_{\mathrm{aux}}^{N}\rangle$	[39, 51-53]
				$\langle M^{\mathrm{W}}, D_{\mathrm{aux}}^{\mathrm{N}}\rangle$	[40]
模型反演攻击	效果不明显	重构部分训练集	鉴别模型	$\langle M^{\mathrm{W}}, D_{\mathrm{aux}}^{\mathrm{N}}\rangle$	[42-43, 45]
				$\langle M^{\mathrm{W}}, D_{\mathrm{aux}}^{\mathrm{S}}\rangle$	[44]
属性猜测攻击	有一定效果	猜测与目标分类无关的属性信息	鉴别模型	$\langle M^{\mathrm{W}}, D_{\mathrm{aux}}^{\mathrm{S}}\rangle, \langle M^{\mathrm{W}}, D_{\mathrm{aux}}^{\mathrm{P}}\rangle$	[45, 47]
模型窃取攻击	效果不明显	窃取模型参数	鉴别模型	$\langle M^{\mathrm{B}}, D_{\mathrm{aux}}^{\mathrm{S}}\rangle, \langle M^{\mathrm{B}}, D_{\mathrm{aux}}^{\mathrm{P}}\rangle$	[48, 54]
无意识记忆	有效	无意识生成与目标无关的个人信息	生成模型	-	[49]

为了能够尽量减少对机器学习可用性的影响, 不修改模型结构及损失函数, 主流差分隐私抵抗机制研究分为梯度扰动(gradient perturbation)[55, 26]和知识转移(knowledge transfer)[57-58]两种差分隐私方案。梯度扰动旨在修改训练过程中的梯度更新算法, 在每个迭代周期的随机梯度递减算法结果中添加差分隐私噪声。知识转移机制基于采样和聚合架构(Sample and Aggregate Framework, SAF), 将非隐私的学生模型采用差分隐私机制聚合出一个满足差分隐私机制的老师模型然后发布。第 4 节将详细描述目标/输出/梯度扰动和知识转移两种差分隐私技术在鉴别模型和生成模型中的运用。

4 机器学习中的差分隐私方法

鉴别模型主要指的是分类器模型, 即给予目标属性, 模型判断其属于哪个类别。鉴别模型在机器学习任务中应用广泛。生成模型, 本文主要指对抗生成模型(Generative Adversarial Nets, GAN), 用于生成与训练集近似分布的人工合成数据集。由于常见的 GAN 分为一个鉴别器(discriminator)和一个生成器(generator)。所以许多针对鉴别模型的差分隐私机制可以微调以适应 GAN 模型。下文将首先介绍鉴别模型中的差分隐私机制, 再介绍这些机制如何微调以保护 GAN 模型。

4.1 鉴别模型

4.1.1 目标扰动和输出扰动机制

机器学习领域, 在早期经验风险最小化(Empirical Risk Minimization, ERM)优化凸函数时, 研究者率先提出了两种方式: 目标扰动[59-61]和输出扰动[58-59]。其中 Chauhuri 等人[58]以逻辑回归(logistic regression)为例, 给出目标扰动和输出扰动的敏感度分析方法。但是其敏感度分析方法依赖目标函数为强凸函数。随着神经网络(neural networks)的深入发展, 损失函数不再是凸函数, 因此依赖强凸函数条件的分析敏感度的方法不再可行, 隐私保护的方法逐渐转入梯度扰动[26, 55]。梯度扰动无需损失函数为强凸性。且敏感度分析可以通过梯度裁剪实现。表 6 总结了 3 种扰动的实现机制。

机器学习任务背景: 设训练集为$\{(\boldsymbol{x}, y)\}_N$, 其中$\boldsymbol{x}$是属性, y是标签, 机器学习目标是根据属性$\boldsymbol{x}$预测

标签y。目标函数有如下基本形式:

$$\mathcal{J}(\theta)=\frac{1}{N}\sum_{i=1}^{N}\ell(\theta,\boldsymbol{x}_i,y_i)+\lambda R(\theta) \quad (11)$$

其中, ℓ为某种损失函数, 例如交叉墒或均方差函数。R为正则函数。机器学习的目标是找出最优参数使得$\theta^*=\arg\min_{\theta}\mathcal{J}(\theta)$。

目标扰动和输出扰动。分别对应表 5 中的方案#1 和方案#3。目标扰动在损失函数中加入噪声。输出扰动在输出结果中加入噪声。难点都在于敏感度分析。Chauhuri 等人[58]给出在二分类的逻辑回归任务中采用 smooth 敏感度技术等方法推导出的差分隐私加噪幅值, 使之可以应用于实际案例中。但是, 该机制只能应用于二分类模型, 且要求目标函数为强凸型, 训练集为较低维度。

表 5 差分隐私噪声添加方法

Table 5 Differential private noise addictive mechanisms

输入: 训练集$\{(\boldsymbol{x},y)\}_N$ 输出: 模型参数θ
初始化θ
#1. 目标扰动: 在损失函数中添加噪声β
$\mathcal{J}(\theta)=\frac{1}{N}\sum_{i=1}^{N}\ell(\theta,\boldsymbol{x}_i,y_i)+\lambda R(\theta)+\beta$ (12)
For i in range (*epochs*):
#2. 梯度扰动: 在梯度中添加噪声β
$\theta=\theta-\eta(\nabla J(\theta)+\beta)$ (13)
End
#3. 输出扰动: 在输出中添加噪声
返回$\theta+\beta$

4.1.2 梯度扰动机制

随机梯度下降(Stochastic Gradient Decent , SGD)是目前优化神经网络损失函数的常用方法。它在每个周期随机采样部分训练集, 计算经验梯度以估计总体梯度并更新参数。如果损失函数并非强凸(神经网络中, 一般都不是强凸), 则随机梯度下降会优化至某个局部最优点。差分隐私的随机梯度扰动(DP-SGD)旨在将符合差分隐私规范的噪声添加到每个周期的经验梯度中, 用扰动的梯度估计更新网络, 以使得每个周期更新的网络参数都满足差分隐私机制。

根据表 5 公式(12), 随机梯度扰动在周期t的基本形态如下

$$w_{t+1}=w_t-\eta_t\left(\frac{1}{b}\sum_{(\boldsymbol{x}_i,y_i)\in B_t}\frac{\partial\mathcal{J}(\boldsymbol{x}_i,y_i)}{\partial_{w_t}}\right) \quad (14)$$

其中, w_t是第t轮模型的权重, η_t是学习率, $\mathcal{J}$是损失函数, B_t为第t轮选择的训练集批次(batch)且$b=|B_t|$。

添加差分隐私扰动的形式如下:

$$w_{t+1}=w_t-\frac{\eta_t}{b}\left(\sum_{(x_i,y_i)\in B_t}\frac{\partial\mathcal{J}(x_i,y_i)}{\partial_{w_t}}+Z(t)\right) \quad (15)$$

Laplace 机制。对于所添加的噪声, Song 等人首先采用 Laplace 机制[55]。令估计梯度函数的ℓ_1敏感度小于 1, 即$\frac{\partial\mathcal{J}(x_i,y_i)}{\partial_{w_t}}\le 1$。根据定理 1, Laplace 噪声的幅值应设为$b=\Delta_1 f/\epsilon=1/\epsilon$以保证$\epsilon$-差分隐私。

高斯机制。此后, Abadi 等人[26]提出基于高斯机制的 DP-SGD, 并使用 moments accountant 以跟踪隐私预算消耗。首先, 在公式(15)之前, 将梯度函数的二阶范数$\|\frac{\partial\mathcal{J}(x_i,y_i)}{\partial_{w_t}}\|_2$裁剪至$C$。其次, 按照公式(15)进行梯度递减, 其中$Z(t)\sim\mathcal{N}(0,\sigma)$, $\sigma\ge\sqrt{2\ln(1.25/\delta)}\frac{C}{\epsilon}$。根据定理 6, 发布差分隐私的经验梯度估计对该批次数据集满足(ϵ,δ)- DP。值得注意的是, 公式(15)只用到了一个批次的敏感数据集, 因此其对整个数据集$\{(\boldsymbol{x},y)\}_N$满足$(q\epsilon,q\delta)$- DP, 其中$q=b/N$。此处分析用到了子采样隐私放大定理, 该定理在 DP-SGD 中经常用到。

定理 10. 子采样隐私放大定理[63-64]. 如果机制$\mathcal{M}$满足(ϵ,δ)-DP, 则$\mathcal{M}\circ\text{subsample}$满足$(\log(1+q(e^{\epsilon}-1)),q\delta)$-DP, 其中$q=b/N$指子采样率。

根据定理 10, 当ϵ较小时, q采样率的(ϵ,δ)-DP 机制隐私将大约放大至$(q\epsilon,q\delta)$- DP。值得注意的是, 定理 10 中的子采样定理针对(ϵ,δ)-DP 差分隐私定义。而我们前文提及, RDP 因为其固有优势, 常用于计算累加隐私。为此, Wang 等人[61]提出了针对 RDP 机制的子采样放大定理。

TensorFlow Privacy①。 RDP 可以看作是 moments accountant 技术的一个实例化, 其中比较著名的开源实现是 TensorFlow 的 Privacy 项目。我们简要介绍其隐私追踪思路: 通过添加噪声采样自$\mathcal{N}(0,\sigma^2)$, 将非隐私保护的 SGD 算法修改为高斯机制 DP-SGD, 则单轮对某个批次满足$(\alpha,\alpha C^2/2\sigma^2)$-RDP, 其中$C$为裁剪阈值。然后根据 RDP 的子采样放大定理[61]计算其放大后的隐私预算。之后, 根据 RDP 线性叠加定理, 对k轮迭代线性叠加隐私消耗。最后根据定理 8 遍历部分α参数找出最小ϵ, 并将 RDP 转换为(ϵ,δ)-DP。

① https://github.com/tensorflow/privacy

4.1.3 知识转移

知识转移方法指的是从一群非隐私保护的老师模型(teacher ensembles)中以隐私保护的模式把模型知识转移到一个新的学生模型(student model)中, 使得学生模型满足隐私保护, 并将学生模型发布给使用者。其中代表性的案例为 Private Aggregation of Teacher Ensembles (PATE)[①][56]。PATE 可以看成是 SAF 技术[62]在深度学习中的一个实例化应用。PATE 的训练过程可以分解为两部分: teacher ensembles 训练(图 3 左侧)和 student model 训练(图 3 右侧)。

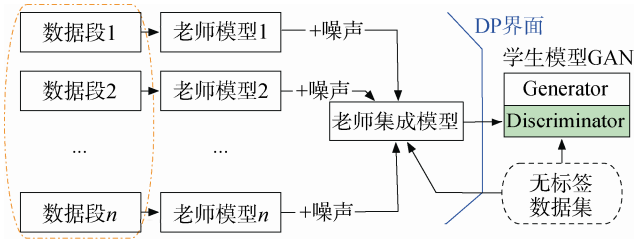


图 3 PATE 系统图

Figure 3 The framework of PATE

• **Teacher ensembles 训练。**首先, 对于隐私数据集$\{(\boldsymbol{x},y)\}_N$将其分成均等大小的n份, 对于每一份数据集$\{(\boldsymbol{x},y)\}_i$训练一个神经网络做分类任务。由此, 总共获取n个老师模型T_i。当有用户用老师模型做标签预测时, 老师模型们会集合出一个加噪的标签返回给用户。Papernot[56-57]采用 Laplace 机制和高斯机制返回扰动最大值。到此为止, 机制返回了一个差分隐私的预测标签给用户, 用户可以直接使用。但是, 因为用户每次通过 DP 接口界面获取答案时, 都会消耗隐私, 因此当隐私预算消耗殆尽, 老师模型就只能丢弃。为了解决这个问题, PATE 系统又添加了学生模型, 以便更高效的转移老师模型的知识。

• **Student model 训练。**学生模型的训练主要由公开的无标签数据集和老师模型预测的加噪标签训练。学生模型较好的选择是半监督的 GAN[63], 半监督的 GAN 的鉴别器是一个$m+1$的分类器, 除了原始m个标签类别外, 额外添件一个“fake”类别。在训练该学生 GAN 时, 除了标签是询问老师得到的加噪标签, 其他都和原始半监督的 GAN 的训练过程类似。训练结束后, 公开鉴别器(discriminator), 可以当作是满足差分隐私的分类器使用。

4.1.4 DP-SGD VS PATE

对于 DP-SGD 和 PATE 两种截然不同的隐私策略, 我们从以下三个角度对比其优劣。

• 隐私保护: 基于 SAF 技术的 PATE 架构与 DP-SGD 有略微不同的隐私假设。PATE 假设属性$\boldsymbol{x}$及其分布并非是需要保护的。其保护的是与$\boldsymbol{x}$关联的标签y的值。拿图 1 举例, PATE 保护其他属性与收入(income)之间的关联性, 但是并不保护某个人的公共属性(婚姻状态 marital staturs 等)。该隐私保护对数据集的假设要强于 DP-SGD, 且并非所有数据集都满足此要求。例如图 1 中的 Adult 数据集、医疗数据集等的个人属性也可能也是需要隐私保护的。

• 可用性: PATE 天然适合于分布式架构。PATE 无需修改现有模型架构, 但是 DP-SGD 需要修改梯度下降策略。PATE 只能用于分类任务, 而 DP-SGD 可以应用于线性回归、分类任务、生成任务等。当用分类准确度来衡量发布的差分隐私架构可用性时, 在同等隐私预算下, PATE 可能优于 DP-SGD。这是因为 PATE 从公共分布中获取了更多与分类任务无关的先验知识。且其用数据相关的隐私分析。

• 计算复杂度: 在计算复杂度这一项, DP-SGD 对比 PATE 有优势。一个典型的 PATE 模型需要 250 个老师模型才能获取隐私和有效性的较优平衡。除此之外, PATE 如果采用数据相关的隐私预算分析, 计算消耗也很大。

4.2 生成模型

生成模型有多种, 本文专指对抗生成模型 GAN。GAN 有很强的分布模仿能力, 能够生成与原始训练集分布近似的高纬度数据集。因此许多研究者用其当作天然的规避隐私保护的方法, 生成并发布合成数据集, 并用人工合成数据集替代隐私数据集发布使用。但是近年来研究发现 GAN 本身并没有严格证明的隐私保护性能, 特别的, 成员猜测攻击对 GAN 也有攻击效果[40, 42-43, 45, 68]。根据第 3 节, 差分隐私机制能够抵抗成员猜测攻击, 因此研究差分隐私的 GAN 对于隐私保护至关重要。

GAN 基本知识: GAN 的基本结构如图 4 所示,

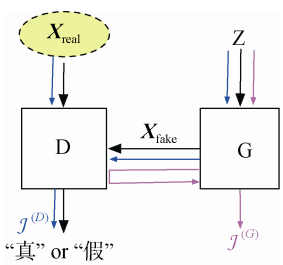


图 4 GAN 结构图

Figure 4 The structure of GAN

① https://github.com/tensorflow/privacy/tree/master/research

包括一个鉴别器网络(Discriminator)和一个生成器网络(Generator)。敏感训练集为$\boldsymbol{X}_{\text{real}}$。生成器和鉴别器相互博弈, 生成器要生成更加逼真的数据, 鉴别器提高鉴别能力以鉴别出人造数据和训练集的区别。两者的损失函数如下。

$$\mathcal{J}^{(D)}\left(\theta^{(D)},\theta^{(G)}\right)=-\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}}\log D(\boldsymbol{x})-\frac{1}{2}\mathbb{E}_{\boldsymbol{z}}\log\left(1-D\left(G(z)\right)\right) \tag{16}$$

$$\mathcal{J}^{(G)}=-\frac{1}{2}\mathbb{E}_{\boldsymbol{z}}\log D\left(G(\boldsymbol{z})\right) \tag{17}$$

鉴别器和生成器同时优化自己的损失函数, 最后达到平衡点。从公式(16)(17)以及图 4 中的损失函数流程可以看出, 只有鉴别器网络 D 的损失函数用到了敏感训练集$\boldsymbol{X}_{\text{real}}$, 生成器网络 G 在训练过程中没有直接接触敏感数据, 而是使用 D 返回的信息进行梯度更新。因此只需要保证鉴别器网络的差分隐私安全, 根据抗后处理定理(定理 4), 生成器的参数及其输出也可以自动保持差分隐私。值得注意的是, 生成器的输出为人工合成数据集, 因此差分隐私的 GAN 可以用来生成并发布满足差分隐私的合成数据集。

4.2.1 基于梯度扰动的差分隐私 GAN

DP-GAN 是 GAN 技术与 DP-SGD 技术的结合。总体思路是对 GAN 的鉴别器做差分隐私的随机梯度递减。根据抗后处理定理, 生成器参数也能保持差分隐私。Xie 等人[65]在鉴别器中的 Wasserstein 距离[66]的梯度优化上加入高斯机制以满足差分隐私, 且也用 moment accountant 技术追踪隐私。Frigerio 等人 [67]将其拓展至生成连续、时间序列、以及离散的合成数据, 并证明其差分隐私的合成数据集可以抵抗成员猜测攻击。

对比 4.1.2 节, DP-SGD 在鉴别模型和生成模型上的技术十分类似, 都是修改随机梯度递减 SGD 使其满足差分隐私。值得注意的是 GAN 训练时只需要修改鉴别器的随机梯度递减, 无需修改生成器的 SGD。在 DP-GAN 隐私追踪时, 注意追踪鉴别器的 SGD 周期即可(有些 GAN 为了增加稳定性会增加一轮生成器周期里对应的鉴别器周期数)。

4.2.2 基于知识转移的差分隐私 GAN

PATE-GAN[68]和 DP-GAN 采用的方法截然不同, 源自 PATE 系统。如图 5 所示, 作者把整个 PATE 系统都当成是 GAN 的鉴别器, 让其与额外添加的生成器博弈。训练结束后, 发布差分隐私的生成器。PATE-GAN 不再需要无标签的公共数据集来训练学生模型, 取而代之的是, 用生成器的部分难以被老师模型鉴别出真假的数据当“真数据”。PATE-GAN 的差分隐私部分与 PATE 类似, 也是用 SAF 技术聚合所有老师模型输出结果。图 5 中生成器的差分隐私也是依赖于抗后处理性质。

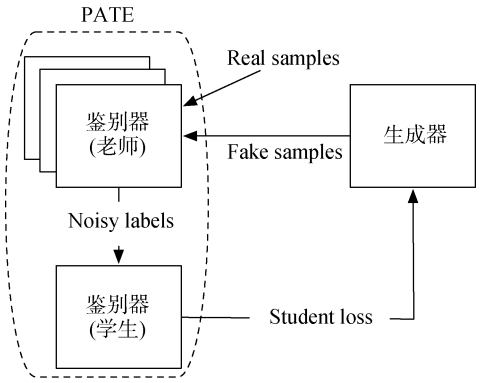


图 5 PATE-GAN 系统

Figure 5 The architecture of PATE-GAN

5 总结和展望

上文详细讨论了差分隐私技术在机器学习领域的发展历程, 包括定义、实现机制和常用性质。并且针对实际攻击, 分析并比较了差分隐私的抗攻击能力。此后, 给出了目前主流的差分隐私的鉴别模型和生成模型保护方案。本节将讨论差分隐私技术在机器学习领域的公开问题以及研究方向。

(1) 模型隐私安全和功能性安全存在折中

一直以来, 机器学习模型的隐私安全和功能性安全处于两个相对平行的研究线路。本文探讨的是模型的隐私安全, 即模型是否泄漏个人隐私。还有一类安全指模型的功能性安全, 例如对抗样本攻击、样本毒化等, 指的是存在恶意攻击者可以用肉眼难以分辨的数据模型的发生误判。差分隐私目前公认对模型的隐私安全有一定的保护效果。但是近期许多研究[69]发现模型的功能性安全可能与隐私安全有对立性, 即防止模型的功能性安全的措施可能会加重隐私安全威胁。因此差分隐私如何同模型功能性安全的抵抗措施有效结合全面防护机器学习的安全性有待研究。

(2) 差分隐私保护机制不是万能

根据本文表 4 的总结, 差分隐私可以防止成员猜测攻击和无意识记忆, 对属性猜测攻击有一定弱化效果。但是对防止模型反演、模型窃取攻击效果不明显。甚至有研究发现[38], 模型窃取攻击和成员猜测攻击的成功率是负相关的。差分隐私机制的效果和攻击原理有直接关系。如果攻击依赖于模型过拟合, 那么差分隐私有明显效果; 如果攻击不是依赖

于模型过拟合, 甚至利用模型的泛化能力, 那么差分隐私没有直接抗攻击效果。因此依赖差分隐私单一机制并不能解决机器学习隐私安全的所有攻击, 应考虑多机制结合以全面防护隐私泄漏问题。

(3) 隐私预算追踪方法有待提高

许多研究表示目前针对机器学习的差分隐私机制牺牲过多有效性以保证安全[30]。另外一些研究也在试图寻在更加严谨的差分隐私预算追踪方法[70]。例如, 目前的 DP-SGD[26]研究假设攻击者可以获取机器学习模型每一轮迭代参数(权重更新), 而不仅仅是可以获取最终训练好的模型的参数。在实际中, 该攻击条件假设太强, 但是这却是目前唯一一种已知的分析 DP-SGD 隐私累加的方式[71]。为此, Feldman 等人[70]推导出直接分析最后一轮模型隐私的方法, 但是其证明依赖损失函数是凸函数的假设, 在神经网络下还没有解决方法。另外, Nasr 等人[71]提出在不同的攻击者能力下, 应该制定不同的差分隐私下限。差分隐私一直考虑最恶劣的攻击条件来保护隐私安全。然而实际环境中很少有攻击者能达到如此强的攻击能力。因此, 针对不同攻击强度细化不同的差分隐私下限有待研究。

(4) 联邦学习模式中差分隐私存在局限性

联邦学习通常指掌握自己部分训练集的多方, 在不泄漏个人训练集的前提下, 共同训练综合模型。原理是训练的每个周期, 各方先下载综合模型, 然后用自己的训练集计算梯度更新并上传, 中心利用各方上传的梯度加权平均更新综合模型。差分隐私机制通常类似 SAF(见图 3), 用差分隐私的方式传递扰动的梯度平均。但是 2017 年 Hitaj 等人[42]研究发现, 即使是差分隐私保护的联邦学习依然不安全。当有恶意参与者存在时, 其可以窃取其他合规参与者的隐私信息。目前还没有可靠的用于联邦学习的差分隐私机制。这使得目前联邦学习的安全性只能依赖计算量以及通信量开销巨大的多方安全计算技术或者是同态加密技术。

(5) GAN 模型中差分隐私存在局限性

差分隐私技术在对抗生成模型(GAN)中的应用尚在探索阶段。比如, 较为先进的 WGAN-GP[72]尚没有差分隐私版本。因为梯度惩罚部分用到了真实训练集, 其隐私预算追踪是个难点。除此之外, 对抗生成模型与鉴别模型的网络架构以及性质也有所不同。其中, 对抗模型的过拟合程度难以衡量(差分隐私主要保护模型过拟合)[39, 51]。对抗模型的随机性可能使得非差分隐私的 GAN 可能天生含有弱差分隐私性质[72]。因此, 在 GAN 中的差分隐私机制可能需要考虑其特点进行定制。比如, 实验性衡量原始非隐私保护的 GAN 的隐私保护程度, 再补充加噪。

参考文献

[1] Dwork C, Smith A. Differential Privacy for Statistics: What we Know and what we Want to Learn[J]. *Journal of Privacy and Confidentiality*, 2010, 1(2): 135-154.

[2] Dwork C. A Firm Foundation for Private Data Analysis[J]. *Communications of the ACM*, 2011, 54(1): 86-95.

[3] Sweeney L. k-Anonymity: A Model for Protecting Privacy[J]. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2002, 10(5): 557-570.

[4] Narayanan A, Shmatikov V. Robust De-Anonymization of Large Sparse Datasets[C]. *2008 IEEE Symposium on Security and Privacy*, 2008: 111-125.

[5] Machanavajjhala A, Kifer D, Gehrke J, et al. *L* -Diversity[J]. *ACM Transactions on Knowledge Discovery from Data*, 2007, 1(1): 3.

[6] Li N H, Li T C, Venkatasubramanian S. T-Closeness: Privacy beyond K-Anonymity and L-Diversity[C]. *2007 IEEE 23rd International Conference on Data Engineering*, 2007: 106-115.

[7] Sarwate A D, Chaudhuri K. Signal Processing and Machine Learning with Differential Privacy: Algorithms and Challenges for Continuous Data[J]. *IEEE Signal Processing Magazine*, 2013, 30(5): 86-94.

[8] Dwork C, McSherry F, Nissim K, et al. Calibrating Noise to Sensitivity in Private Data Analysis[C]. *Theory of Cryptography*, 2006: 265-284.

[9] Ding B, Kulkarni J, Yekhanin S. Collecting Telemetry Data Privately. [EB/OL]. 2017: ArXiv Preprint ArXiv: 1712.01524.

[10] Differential Privacy Team, Apple. Learning with Privacy at Scale. https://machinelearning.apple.com/research/learning-with-privacy-at-scale. 2017.

[11] Erlingsson Ú, Pihur V, Korolova A. RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response[C]. *The 2014 ACM SIGSAC Conference on Computer and Communications Security*, 2014: 1054-1067.

[12] Fanti G, Pihur V, Erlingsson Ú. Building a RAPPOR with the Unknown: Privacy-Preserving Learning of Associations and Data Dictionaries[J]. *Proceedings on Privacy Enhancing Technologies*, 2016, 2016(3): 41-61.

[13] Machanavajjhala A, Kifer D, Abowd J, et al. Privacy: Theory Meets Practice on the Map[C]. *2008 IEEE 24th International Conference on Data Engineering*, 2008: 277-286.

[14] Gaboardi M, Honaker J, King G, et al. PSI ({\Psi}): a Private data Sharing Interface[EB/OL]. 2018: ArXiv Preprint ArXiv: 1609.04340.

[15] Dwork C, Roth A. The Algorithmic Foundations of Differential Privacy[J]. *Foundations and Trends® in Theoretical Computer Science*, 2013, 9(3/4): 211-407.

[16] Ji Z, Lipton ZC, Elkan C. Differential Privacy and Machine Learning: a Survey and Review[EB/OL]. 2021: ArXiv Preprint ArXiv:1412.7584

[17] Goryczka S, Xiong L. A Comprehensive Comparison of Multiparty Secure Additions with Differential Privacy[J]. *IEEE Transactions on Dependable and Secure Computing*, 2017, 14(5): 463-477.

[18] Jain P, Gyanchandani M, Khare N. Differential Privacy: Its Technological Prescriptive Using Big Data[J]. *Journal of Big Data*, 2018, 5: 15.

[19] Zhu T Q, Li G, Zhou W L, et al. Differentially Private Data Publishing and Analysis: A Survey[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2017, 29(8): 1619-1638.

[20] McSherry F, Talwar K. Mechanism Design via Differential Privacy[C]. *48th Annual IEEE Symposium on Foundations of Computer Science*, 2007: 94-103.

[21] Warner S L. Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias[J]. *Journal of the American Statistical Association*, 1965, 60(309): 63-69.

[22] Cao Y, Yoshikawa M, Xiao Y H, et al. Quantifying Differential Privacy in Continuous Data Release under Temporal Correlations[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2019, 31(7): 1281-1295.

[23] Dwork C, Kenthapandi K, McSherry F, et al. Our Data, Ourselves: Privacy Via Distributed Noise Generatio [C]. *Annual International Conference on the Theory and Applications of Cryptographic Techniques*. 2006: 486-503.

[24] Dwork C, Rothblum GN. Concentrated Differential Privacy[EB/OL]. 2016: ArXiv Preprint ArXiv: 1603.01887

[25] Bun M, Steinke T. Concentrated Differential Privacy: Simplifications, Extensions, and Lower Bounds[C]. *Theory of Cryptography*, 2016: 635-658.

[26] Abadi M, Chu A, Goodfellow I, et al. Deep Learning with Differential Privacy[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 308-318.

[27] Mironov I. Rényi Differential Privacy[C]. *2017 IEEE 30th Computer Security Foundations Symposium*, 2017: 263-275.

[28] Dwork C, Rothblum G N, Vadhan S. Boosting and Differential Privacy[C]. *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*, 2010: 51-60.

[29] Kairouz P, Oh S, Viswanath P. The Composition Theorem for Differential Privacy[C]. *IEEE Transactions on Information Theory*, 2015: 4037-4049.

[30] Jayaraman B, Evans D. Evaluating Differentially Private Machine Learning in Practice[EB/OL]. 2019: arXiv: 1902.08874[cs.LG]. https://arxiv.org/abs/1902.08874.

[31] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[EB/OL]. 2014: ArXiv Preprint ArXiv: 1312.6199.

[32] Shokri R, Stronati M, Song C Z, et al. Membership Inference Attacks Against Machine Learning Models[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 3-18.

[33] Maini P, Yaghini M, Papernot N. Dataset Inference: Ownership Resolution in Machine Learning[EB/OL]. 2021: arXiv preprint arXiv:2104.10706.

[34] Yeom S, Giacomelli I, Fredrikson M, et al. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting[C]. *2018 IEEE 31st Computer Security Foundations Symposium*, 2018: 268-282.

[35] 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]. *2019 IEEE Symposium on Security and Privacy*, 2019: 739-753.

[36] Choo C A C, Tramer F, Carlini N, et al. Label-only Membership Inference Attacks[EB/OL]. 2020: ArXiv Preprint ArXiv: 2007.14321.

[37] Li Z, Zhang Y. Label-Leaks: Membership Inference Attack with Label[EB/OL]. 2020: ArXiv Preprint ArXiv: 2007.15528.

[38] Liu Y G, Wen R, He X L, et al. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models[EB/OL]. 2021: arXiv: 2102.02551[cs.CR]. https://arxiv.org/abs/2102.02551

[39] Chen D F, Yu N, Zhang Y, et al. GAN-Leaks: A Taxonomy of Membership Inference Attacks Against Generative Models[C]. *The 2020 ACM SIGSAC Conference on Computer and Communications Security*, 2020: 343-362.

[40] Fredrikson M, Lantz E, Jha S, et al. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing[J]. *Proceedings of the USENIX Security Symposium UNIX Security Symposium*, 2014, 2014: 17-32.

[41] Fredrikson M, Jha S, Ristenpart T. Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures[C]. *The 22nd ACM SIGSAC Conference on Computer and Communications Security*, 2015: 1322-1333.

[42] Hitaj B, Ateniese G, Perez-Cruz F. Deep Models under the GAN: Information Leakage from Collaborative Deep Learning[C]. *The 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017: 603-618.

[43] Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative Adversarial Networks[J]. *Communications of the ACM*, 2020, 63(11): 139-144.

[44] Zhang Y H, Jia R X, Pei H Z, et al. The Secret Revealer: Generative Model-Inversion Attacks Against Deep Neural Networks[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 250-258.

[45] Melis L, Song C Z, de Cristofaro E, et al. Exploiting Unintended Feature Leakage in Collaborative Learning[C]. *2019 IEEE Symposium on Security and Privacy*, 2019: 691-706.

[46] Mireshghallah F, Taram M, Jalali A, et al. Not all Features are Equal: Discovering Essential Features for Preserving Prediction Privacy[C]. *The Web Conference 2021*, 2021: 669-680.

[47] Song C, Shmatikov V. Overlearning reveals sensitive attributes[EB/OL]. 2019: arXiv preprint arXiv:1905.11742.

[48] Tramèr F, Zhang F, Juels A, et al. Stealing Machine Learning Models via Prediction APIs[C]. *The 25th USENIX Conference on Security Symposium*, 2016: 601-618.

[49] Carlini N, Liu C, Erlingsson Ú, et al. The secret sharer: Evaluating and testing unintended memorization in neural networks[C]. *28th USENIX Security Symposium*. 2019: 267-284.

[50] Salem A, Zhang Y, Humbert M, et al. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models[EB/OL]. 2018: arXiv preprint arXiv: 1806.01246.

[51] Hayes J, Melis L, Danezis G, et al. LOGAN: Membership Inference Attacks Against Generative Models[J]. *Proceedings on Privacy Enhancing Technologies*, 2019, 2019(1): 133-152.

[52] Hilprecht B, Härterich M, Bernau D. Monte Carlo and Reconstruction Membership Inference Attacks Against Generative Models[J]. *Proceedings on Privacy Enhancing Technologies*, 2019, 2019(4): 232-249.

[53] Park N, Mohammadi M, Gorde K, et al. Data Synthesis Based on Generative Adversarial Networks[J]. *Proceedings of the VLDB Endowment*, 2018, 11(10): 1071-1083.

[54] Orekondy T, Schiele B, Fritz M. Knockoff Nets: Stealing Functionality of Black-Box Models[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019: 4949-4958.

[55] Song S, Chaudhuri K, Sarwate A D. Stochastic Gradient Descent with Differentially Private Updates[C]. *2013 IEEE Global Conference on Signal and Information Processing*, 2013: 245-248.

[56] Papernot N, Abadi M, Erlingsson Ú, et al. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data[EB/OL]. 2017: arXiv preprint arXiv: 1610.05755

[57] Papernot N, Song S, Mironov I, et al. Scalable Private Learning with PATE[EB/OL]. 2018: arXiv preprint arXiv: 1802.08908

[58] Chaudhuri K, Monteleoni C, Sarwate A D. Differentially Private Empirical Risk Minimization[J]. *Journal of Machine Learning Research*: *JMLR*, 2011, 12: 1069-1109.

[59] Hamm J, Cao Y, Belkin M. Learning privately from multiparty data[C]. *International Conference on Machine Learning*, 2016: 555-563.

[60] Kasiviswanathan S P, Lee H K, Nissim K, et al. What can we Learn Privately? [J]. *SIAM Journal on Computing*, 2011, 40(3): 793-826.

[61] Wang Y X, Balle B, Kasiviswanathan S. Subsampled Rényi Differential Privacy and Analytical Moments Accountant[J]. *Journal of Privacy and Confidentiality*, 2020, 10(2): 1226-1235.

[62] Nissim K, Raskhodnikova S, Smith A. Smooth Sensitivity and Sampling in Private Data Analysis[C]. *The thirty-ninth annual ACM symposium on Theory of computing*, 2007: 75-84.

[63] Salimans T, Goodfellow I, Zaremba W, et al. Improved techniques for training gans[EB/OL]. 2016: arXiv preprint arXiv:1606.03498.

[64] Long Y, Bindschaedler V, Wang L, et al. Understanding membership inferences on well-generalized learning models[EB/OL]. 2018: arXiv preprint arXiv:1802.04889.

[65] Xie L, Lin K, Wang S, et al. Differentially private generative adversarial network[EB/OL]. 2018: arXiv preprint arXiv:1802.06739.

[66] Chen N G, Li C M. Hyperspectral Image Classification Approach Based on Wasserstein Generative Adversarial Networks[C]. *2020 International Conference on Machine Learning and Cybernetics*, 2020: 53-63.

[67] Frigerio L, Oliveira A S, Gomez L, et al. Differentially Private Generative Adversarial Networks for Time Series, Continuous, and Discrete Open Data[C]. *ICT Systems Security and Privacy Protection*, 2019: 151-164.

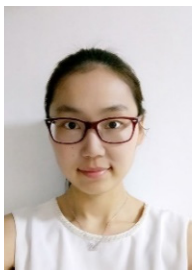
[68] Jordon J, Yoon J, Van Der Schaar M. PATE-GAN: Generating synthetic data with differential privacy guarantees[C]. *International Conference on Learning Representations*, 2018.

[69] Song L W, Shokri R, Mittal P. Privacy Risks of Securing Machine Learning Models Against Adversarial Examples[C]. *The 2019 ACM SIGSAC Conference on Computer and Communications Security*, 2019: 241-257.

[70] Feldman V, Mironov I, Talwar K, et al. Privacy Amplification by Iteration[C]. *2018 IEEE 59th Annual Symposium on Foundations of Computer Science*, 2018: 521-532.

[71] Nasr M, Songi S, Thakurta A, et al. Adversary Instantiation: Lower Bounds for Differentially Private Machine Learning[C]. *2021 IEEE Symposium on Security and Privacy*, 2021: 866-882.

[72] Lin Z, Sekar V, Fanti G. On the Privacy Properties of GAN-generated Samples[C]. *International Conference on Artificial Intelligence and Statistics*, 2021: 1522-1530.



胡奥婷 于 2018 年在东南大学网络安全专业获得硕士学位。现在东南大学网络空间安全专业攻读博士学位。 研究领域为差分隐私, 机器学习隐私保护。研究兴趣包括: 云存储安全, 大数据安全。Email: aotinghu@seu.edu.cn

胡爱群 于 1992 年在东南大学信号于信息处理专业获得博士学位。现为东南大学信息科学与工程学院移动通信国家重点实验室教授/博导。研究领域包括物理层安全技术、内生安全及隐私保护。Email: aqhu@seu.edu.cn

胡韵 于 2015 在西安电子科技大学计算机应用专业获得硕士学位, 现在东南大学信息与通信工程专业攻读博士学位, 研究领域为网络安全、信息安全, 研究兴趣包括, 数据安全追踪、差分隐私技术。Email: yun_hu@seu.edu.cn

李古月 于 2017 年在东南大学信息与通信工程获得博士学位。现任东南大学网络空间安全学院任副教授。研究领域为物理层安全, 移动通信安全。Email: guyuelee@seu.edu.cn

韩金广 于 2013 年在澳大利亚卧龙岗大学计算机科学与软件工程学院获得博士学位。现任南京财经大学教授, 研究领域包括密码学, 访问控制, 区块链, 隐私保护系统。Email: jghan22@gmail.com