

用于模型性能评估的方法、装置、设备和介质

- 5 本申请要求于 2022 年 5 月 13 日递交的，标题为“用于模型性能评估的方法、装置、设备和介质”、申请号为 202210524865.2 的中国发明专利申请的优先权。

技术领域

- 10 本公开的示例实施例总体涉及计算机领域，特别地涉及用于模型性能评估的方法、装置、设备和计算机可读存储介质。

背景技术

- 15 当前机器学习已经得到了广泛的应用，其性能通常是随着数据量的增加而提高。在一些方案中，需要集中充足的收集数据样本和标签数据用于机器学习模型的训练。然而，在很多现实场景中，存在着所谓的数据孤岛问题，即数据通常是分散隔离的，存储在不同的实体（例如，企业、用户端）上。随着数据隐私保护问题越来越受到重视，这样的集中式机器学习难以到达数据保护的目的。

- 20 当前，提出了联邦学习方案。联邦学习，指的是在保证数据隐私安全的基础上，利用各个节点的数据实现共同建模，提升机器学习模型的效果。联邦学习可以允许各个节点不离开端，以达到数据保护目的。

25 发明内容

根据本公开的示例实施例，提供了一种用于模型性能评估的方案。

在本公开的第一方面，提供了一种用于模型性能评估的方法。该方法包括：在客户端节点处，将多个数据样本分别应用到预测模型，以得到预测模型输出的多个预测得分，多个预测得分分别指示多个数

据样本属于第一类别或第二类别的预测概率；基于多个数据样本的多个真值标签和多个预测得分，确定与预测模型的预定性能指标相关的多个度量参数的值；对多个度量参数的值施加扰动，得到多个度量参数的扰动值；以及将多个度量参数的扰动值发送给服务节点。

5 在本公开的第二方面，提供了一种用于模型性能评估的方法。该方法包括：在服务节点处，从多个客户端节点分别接收与预测模型的预定性能指标相关的多个度量参数的扰动值；按度量参数聚合来自多个客户端节点的多个度量参数的扰动值，得到多个度量参数的聚合值；以及基于多个度量参数的聚合值来确定预定性能指标的值。

10 在本公开的第三方面，提供了一种用于模型性能评估的装置。该装置包括：预测模块，被配置为将多个数据样本分别应用到预测模型，以得到预测模型输出的多个预测得分，多个预测得分分别指示多个数据样本属于第一类别或第二类别的预测概率；度量确定模块，被配置为基于多个数据样本的多个真值标签和多个预测得分，确定与预测模型
15 型的预定性能指标相关的多个度量参数的值；扰动模块，被配置为对多个度量参数的值施加扰动，得到多个度量参数的扰动值；以及发送模块，被配置为将多个度量参数的扰动值发送给服务节点。

 在本公开的第四方面，提供了一种用于模型性能评估的装置。该装置包括：接收模块，被配置为从多个客户端节点分别接收与预测模
20 型的预定性能指标相关的多个度量参数的扰动值；聚合模块，被配置为按度量参数聚合来自多个客户端节点的多个度量参数的扰动值，得到多个度量参数的聚合值；以及性能确定模块，被配置为基于多个度量参数的聚合值来确定预定性能指标的值。

 在本公开的第五方面，提供了一种电子设备。该设备包括至少一个
25 处理单元；以及至少一个存储器，至少一个存储器被耦合到至少一个处理单元并且存储用于由至少一个处理单元执行的指令。指令在由至少一个处理单元执行时使设备执行第一方面的方法。

 在本公开的第六方面，提供了一种电子设备。该设备包括至少一个处理单元；以及至少一个存储器，至少一个存储器被耦合到至少一

个处理单元并且存储用于由至少一个处理单元执行的指令。指令在由至少一个处理单元执行时使设备执行第二方面的方法。

在本公开的第七方面，提供了一种计算机可读存储介质。介质上存储有计算机程序，计算机程序被处理器执行以实现第一方面的方法。

5 在本公开的第八方面，提供了一种计算机可读存储介质。介质上存储有计算机程序，计算机程序被处理器执行以实现第二方面的方法。

应当理解，本发明内容部分中所描述的内容并非旨在限定本公开的实施例的关键特征或重要特征，也不用于限制本公开的范围。本公开的其它特征将通过以下的描述而变得容易理解。

10

附图说明

结合附图并参考以下详细说明，本公开各实施例的上述和其他特征、优点及方面将变得更加明显。在附图中，相同或相似的附图标记表示相同或相似的元素，其中：

15 图 1 示出了本公开的实施例能够在其中应用的示例环境的示意图；

图 2 示出了根据本公开的一些实施例的用于模型性能评估的信令流的流程图；

图 3A 示出了根据本公开的一些实施例的确定度量参数的值的过程的流程图；

20 图 3B 示出根据本公开的一些实施例的确定度量参数的值的信令流的流程；

图 4 示出根据本公开的一些实施例的在服务节点处用于模型性能评估的过程的流程图；

25 图 5 示出根据本公开的一些实施例的在服务节点处用于模型性能评估的过程的流程图；

图 6 示出了根据本公开的一些实施例的在客户端节点处用于模型性能评估的装置的框图；

图 7 示出了根据本公开的一些实施例的在服务节点处用于模型性能评估的装置的框图；以及

图 8 示出了能够实施本公开的一个或多个实施例的计算设备/系统的框图。

具体实施方式

5 下面将参照附图更详细地描述本公开的实施例。虽然附图中示出了本公开的某些实施例，然而应当理解的是，本公开可以通过各种形式来实现，而且不应该被解释为限于这里阐述的实施例，相反，提供这些实施例是为了更加透彻和完整地理解本公开。应当理解的是，本公开的附图及实施例仅用于示例性作用，并非用于限制本公开的保护范围。

在本公开的实施例的描述中，术语“包括”及其类似用语应当理解为开放性包含，即“包括但不限于”。术语“基于”应当理解为“至少部分地基于”。术语“一个实施例”或“该实施例”应当理解为“至少一个实施例”。术语“一些实施例”应当理解为“至少一些实施例”。

15 下文还可能包括其他明确的和隐含的定义。

可以理解的是，本技术方案所涉及的数据（包括但不限于数据本身、数据的获取或使用）应当遵循相应法律法规及相关规定的要求。

可以理解的是，在使用本公开各实施例公开的技术方案之前，均应当根据相关法律法规通过适当的方式对本公开所涉及个人信息

20 类型、使用范围、使用场景等告知用户并获得用户的授权。

例如，在响应于接收到用户的主动请求时，向用户发送提示信息，以明确地提示用户，其请求执行的操作将需要获取和使用到用户的个人信息。从而，使得用户可以根据提示信息来自主地选择是否向执行本公开技术方案的操作的电子设备、应用程序、服务器或存储介质等

25 软件或硬件提供个人信息。

作为一种可选的但非限制性的实现方式，响应于接收到用户的主动请求，向用户发送提示信息的方式，例如可以是弹窗的方式，弹窗中可以以文字的方式呈现提示信息。此外，弹窗中还可以承载供用户选择“同意”或“不同意”向电子设备提供个人信息的选择控件。

可以理解的是，上述通知和获取用户授权过程仅是示意性的，不对本公开的实现方式构成限定，其他满足相关法律法规的方式也可应用于本公开的实现方式中。

如本文中所使用的，术语“模型”可以从训练数据中学习到相应的输入与输出之间的关联，从而在训练完成后可以针对给定的输入，生成对应的输出。模型的生成可以基于机器学习技术。深度学习是一种机器学习算法，通过使用多层处理单元来处理输入和提供相应输出。神经网络模型是基于深度学习的模型的一个示例。在本文中，“模型”也可以被称为“机器学习模型”、“学习模型”、“机器学习网络”或“学习网络”，这些术语在本文中可互换地使用。

“神经网络”是一种基于深度学习的机器学习网络。神经网络能够处理输入并且提供相应输出，其通常包括输入层和输出层以及在输入层与输出层之间的一个或多个隐藏层。在深度学习应用中使用的神经网络通常包括许多隐藏层，从而增加网络的深度。神经网络的各个层按顺序相连，从而前一层的输出被提供作为后一层的输入，其中输入层接收神经网络的输入，而输出层的输出作为神经网络的最终输出。神经网络的每个层包括一个或多个节点（也称为处理节点或神经元），每个节点处理来自上一层的输入。

通常，机器学习大致可以包括三个阶段，即训练阶段、测试阶段和应用阶段（也称为推理阶段）。在训练阶段，给定的模型可以使用大量的训练数据进行训练，不断迭代更新参数值，直到模型能够从训练数据中获取一致的满足预期目标的推理。通过训练，模型可以被认为能够从训练数据中学习从输入到输出之间的关联（也称为输入到输出的映射）。训练后的模型的参数值被确定。在测试阶段，将测试输入应用到训练后的模型，测试模型是否能够提供正确的输出，从而确定模型的性能。在应用阶段，模型可以被用于基于训练得到的参数值，对实际的输入进行处理，确定对应的输出。

图 1 示出了本公开的实施例能够在其中实现的示例环境 100 的示意图。环境 100 涉及联邦学习环境，其中包括 N 个客户端节点 110-

1...110-k、...110-N (其中 N 为大于 1 的整数, k=1、2、...N)
以及服务节点 120。客户端节点 110-1...110-k、...110-N 可以分
别维护各自的本地数据集 112-1...112-k、...112-N。为便于讨论,
客户端节点 110-1...110-k、...110-N 可以被统称为或单独称为客
5 户端节点 110, 本地数据集 112-1...112-k、...112-N 可以被统称
为或单独称为本地数据集 112。

在一些实施例中, 客户端节点 110 和/或服务节点 120 可以被实现
在终端设备或服务器处。终端设备可以是任意类型的移动终端、固定
终端或便携式终端, 包括移动手机、台式计算机、膝上型计算机、笔
10 记本计算机、上网本计算机、平板计算机、媒体计算机、多媒体平板、
个人通信系统 (PCS) 设备、个人导航设备、个人数字助理 (PDA)、
音频/视频播放器、数码相机/摄像机、定位设备、电视接收器、无线电
广播接收器、电子书设备、游戏设备或者前述各项的任意组合, 包括
这些设备的配件和外设或者其任意组合。在一些实施例中, 终端设备
15 也能够支持任意类型的针对用户的接口 (诸如 “可佩戴” 电路等)。
服务器是能够提供计算能力的各种类型的计算系统/服务器, 包括但不
限于大型机、边缘计算节点、云环境中的计算设备, 等等。

在联邦学习中, 客户端节点指的是提供应用训练、验证或评估预
测模型的部分数据的节点。客户端节点也可称为客户端、终端节点、
20 终端设备、用户设备等。在联邦学习中, 服务节点指的是聚合客户端
节点处的结果的节点。

在图 1 的示例中, 假设 N 个客户端节点 110 节点共同参与对预测
模型 125 的训练, 并将训练中的中间结果汇集到服务节点 120, 以由
服务节点 120 更新预测模型 125 的参数集。这些客户端节点 110 的本
25 地数据的全集构成预测模型 125 的完整训练数据集。因此, 根据联邦
学习的机制, 服务节点 120 可以确定全局的预测模型 125。

针对预测模型 125, 客户端节点 110 处的本地数据集 112 可以包
括数据样本和真值标签。图 1 具体示出了客户端节点 110-k 处的本地
数据集 112-k, 其包括数据样本集和真值标签集。数据样本集包括多

个 (M 个) 数据样本 102-1、102- i 、... 102-M (统称为或单独称为数据样本 102), 并且真值标签集包括对应的多个 (M 个) 真值标签 (ground-truth label) 105-1、105- i 、... 105-M (统称为或单独称为真值标签 105)。其中 M 为大于 1 的整数, $i=1, 2, \dots, M$ 。每个数据
5 样本 102 可以被标注有对应的真值标签 105。数据样本 102 可以对应于预测模型 125 的输入, 真值标签 105 指示数据样本 102 的真实输出。真值标签是有监督机器学习中的重要部分。

在本公开的实施例中, 预测模型 125 可以基于各种机器学习或深度学习的模型架构来构建, 并且可以被配置为实现各种预测任务, 诸如各种分类任务、推荐任务等等。相应地, 预测模型 125 也可以被称为推荐模型、分类模型, 等等。
10

数据样本 102 可以包括与预测模型 125 的具体任务相关的输入信息, 真值标签 105 与任务的期望输出有关。作为一个示例, 在二分类任务中, 预测模型 125 可以被配置为预测输入的数据样本属于第一类
15 别或是第二类别, 真值标签用于标注该数据样本实际属于第一类别或是第二类别。很多实际应用均可以被归类为这样的二分类任务, 例如在推荐任务中对推荐项目的转化 (例如, 点击、购买、注册或其他需求行为) 与否, 等等。

应当理解, 图 1 仅示出了示例的联邦学习环境。根据联邦学习算法和实际应用需要, 环境还可以不同。例如, 虽然被示出为单独的节点, 在某些应用中, 服务节点 120 除了作为中央节点外, 还可以作为
20 客户端节点, 以提供部分数据用于模型训练、模型性能评估等。本公开的实施例在此方面不受限制。

在预测模型 125 的训练阶段, 已有一些机制保护各个客户端节点
25 110 的本地数据不泄露。例如, 在模型训练过程中, 客户端节点 110 不必透漏本地的数据样本或标签数据, 而是向服务节点 120 发送根据本地训练数据计算的梯度数据, 以供服务节点 120 更新预测模型 125 的参数集。

在一些情况下, 还希望评估训练出的预测模型的性能。模型性能

的评估也需要数据，包括模型输入需要的数据样本以及数据样本对应的标签数据。预测模型的性能可以通过一个或多个性能指标来衡量。不同性能指标可以从不同角度，衡量预测模型针对数据样本集给出的预测输出与真值标签集所指示的真实输出之间的差异。通常，如果预测模型给出的预测输出与真实输出之间的差异较小，那意味着预测模型的性能较好。可以看出，通常需要基于数据样本的真值标签集来确定预测模型的性能指标。

随着数据监管体系不断加强，对数据隐私保护的要求也越来越高。对数据样本的真值标签也需要保护，避免被泄露。因此，如何既能够确定预测模型的性能指标，又保护客户端节点本地的标签数据不被泄露，是一项具有挑战性的任务。当前还没有非常有效的方案能够解决该问题。

根据本公开的实施例，提供了一种模型性能评估方案，其能够保护客户端节点本地的标签数据。具体地，在客户端节点处，在计算出与预测模型的性能指标相关的多个度量参数的值后，对所确定的度量参数的值施加扰动，得到多个度量参数的扰动值。客户端节点将度量参数的扰动值发送给服务节点。由于不需要直接发送度量参数的真实值，通过扰动值，观察者难以从扰动值推导出数据样本的真值标签。这样，可以有效避免数据泄露。

在服务节点处，服务节点从多个客户端节点接收到它们各自确定出的多个度量参数的扰动值。服务节点按度量参数聚合来自多个客户端节点的多个度量参数的扰动值。在对多个不同来源的扰动值的聚合进行聚合后，扰动被抵消。因此，基于多个度量参数的聚合值，服务节点能够准确确定模型的性能指标的值。

根据本公开的实施例，各个客户端节点无需暴露本地的真值标签集或基于真值标签确定的参数值，并且还能允许服务节点计算出模型的性能指标值。以此方式，在实现模型性能评估的同时，达到了对客户端节点本地标签数据的隐私保护目的。

以下将继续参考附图描述本公开的一些示例实施例。

图 2 示出了根据本公开的一些实施例的用于模型性能评估的信令流 200 的示意框图。为便于讨论，参考图 1 的环境 100 进行讨论。信令流 200 涉及客户端节点 110 和服务节点 120。

在本公开的实施例中，假设要评估预测模型 125 的性能。在一些实施例中，待评估的预测模型 125 可以是基于联邦学习的训练过程确定的全局预测模型，例如客户端节点 110 和服务节点 120 参与了预测模型 125 的训练过程。在一些实施例中，预测模型 125 也可以是以任何其他方式获得的模型，并且客户端节点 110 和服务节点 120 可以未参与了预测模型 125 的训练过程。本公开的范围在此方面不受限制。

10 在一些实施例中，如信令流 200 所示，服务节点 120 将预测模型 125 发送 205 给 N 个客户端节点 110。在接收 210 到预测模型 125 后，各个客户端节点 110 可以基于预测模型 125 来执行后续评估过程。在一些实施例中，也可以以任何其他适当的方式将要评估的预测模型 125 提供给客户端节点 110。

15 在本公开的实施例中，将单个客户端节点 110 的角度来描述客户端测的操作。多个客户端节点 110 可以类似操作。

在信令流 200 中，客户端节点 110 将多个数据样本分别应用 215 模型 125，以得到预测模型 125 输出的多个预测得分。假设客户端节点 110-k 的数据样本集是 X_k ，预测模型 125 被表示为 $f()$ ，那么针对数据样本集的预测得分集合可以被表示为 $s^k = f(X_k)$ 。

在本公开的实施例中，特别关注于在实现二分类任务的预测模型的性能指标。每个预测得分可以指示对应的数据样本 102 属于第一类别或第二类别的预测概率。这两个类别可以根据实际任务需要配置。

预测模型 125 输出的预测得分的取值范围可以任意设置。例如，预测得分可以是在某个连续取值区间中的取值（例如，0 到 1 之间的取值），或者可以是多个离散取值中的一个取值（例如，可以是 0、1、2、3、4、5 等离散取值之一）。在一些示例中，越高的预测得分可以指示数据样本 102 属于第一类别的预测概率越大，属于第二类别的预测概率越小。当然，相反设置也是可以的，例如越高的预测得分

可以指示数据样本 102 属于第二类别的预测概率越大，属于第一类别的预测概率越小。

客户端节点 110 基于所述多个数据样本 102 的多个真值标签（也可称为真实值标签）和模型输出的多个预测得分，确定 220 与预测模型 125 的预定性能指标相关的多个度量参数的值。

真值标签 105 用于标注对应的数据样本 102 属于第一类别或是第二类别。在下文中，为了方便讨论，将由属于第一类别的数据样本有时称为正样本、正例或正类样本，将属于第二类别的数据样本有时称为负样本、负例或负类样本。在一些实施例中，每个真值标签 105 可以具有两个取值之一，分别用于指示第一类别或第二类别。在下文的一些实施例中，为了方便讨论，可以将第一类别对应的真值标签 105 的取值设置为“1”，其指示数据样本属于第一类别，是正样本。此外，可以将第二类别对应的真值标签 105 的取值设置为“0”，其指示数据样本属于第二类别，是负样本。

在本公开的实施例中，个体客户端节点 110 根据本地数据集（数据样本和真值标签）确定与模型的性能指标相关的度量信息。通过将多个客户端节点 110 的度量信息汇总到服务节点 120，可以相当于在多个客户端节点的完整数据集基础上评估出预测模型 125 的性能。

度量信息指的是在计算模型的性能指标时需要关心的信息，通常可以由多个度量参数指示。这些度量参数的值是需要数据样本经过模型后输出的结果（即预测得分），以及数据样本对应的真值标签的基础上进行统计来得出。由客户端节点提供的度量信息的类型可以取决于要计算的具体性能指标。

在下文中，为便于理解，首先介绍用于实现二分类任务的预测模型 125 的一些示例性能指标。

预测模型 125 针对某个数据样本输出的预测得分，通常会用于与某个得分阈值相比较，并根据比较结果确定该数据样本被预测为属于第一类别或是第二类别。用于实现二分类任务的预测模型 125 的预测可能会出现四种结果。

具体地，对于某个数据样本 102，假设真值标签 105 指示其属于第一类别（正样本），预测模型 125 也预测出其为正样本，那认为该数据样本是真正样本（True Positive，TP）。如果真值标签 105 指示其属于第一类别（正样本），预测模型 125 预测出其为负样本，那认为该数据样本是假负样本（False Negative，FN）。如果真值标签 105 指示其属于第二类别（负样本），但预测模型 125 也预测出其为负样本，那认为该数据样本是真负本（True Negative，TN）。如果真值标签 105 指示其属于第二类别（负样本），但预测模型 125 预测出其为正样本，那认为该数据样本是假正本（False Positive，FP）。这四种结果可以由以下表 1 的混淆矩阵指示。

表 1

		真值标签	
		Positive (P)	Negative (N)
预测结果	P'	真正（TP）	假正（FP）
	N'	假负（FN）	真负（FN）

在衡量预测模型 125 的性能时，期望能够在多个客户端节点 110 的数据样本全集的预测结果以及真值标签全集基础上计算性能指标。在一些实施例中，预测模型 125 的性能指标还可以包括假正样本比率（FPR）和/或假正样本比率（FPR）。FPR 可以被定义为：在实际为负例的数据样本中，被模型错误地判断为阳性的比率，表示为 $FPR=FP/(FP+TN)$ ，其中 FP、TN 表示在数据样本全集中统计出的 FP、TN 的数目。TPR: 实际为阳性的样本中，被正确地判断为阳性的比率，表示为 $TPR=TP/(TP+FN)$ 。

在一些实施例中，预测模型 125 的性能指标可以包括受试者工作特征曲线（ROC）的曲线下面积（AUC）。

ROC 曲线是根据不同的分类方式（设置不同得分阈值），以假正样本比率（FPR）为 X 轴，真正样本比率（TPR）为 Y 轴，在坐标轴上绘制出的曲线。根据每个可能的得分阈值，可以计算出多个（FPR,

TPR) 对的坐标点, 将这些点连成线, 就成为特定模型的 ROC 曲线。

从定义上理解, AUC 指的是 ROC 曲线下方面积。在计算 AUC 时, 一种可能的方式是, 可以根据 AUC 的定义, 可以通过用近似算法计算 ROC 曲线下的面积来计算 AUC。

- 5 在一些实施例, 还可以从概率视角来确定 AUC。AUC 可以被认为是: 随机选择一个正样本和一个负样本, 预测模型给正样本的预测得分高于负样本的预测得分的概率。也就是说, 在数据样本集中, 将正、负样本两两相遇形成正负样本对, 其中正样本的预测得分大于负样本的预测得分的占比。如果模型能够给更多正样本输出高于负样本的预测得分, 可以认为 AUC 更高, 模型的性能更好。AUC 的取值范围在 0.5 和 1 之间。AUC 越接近 1, 说明模型的性能越好。

在上述 AUC 计算中, 均需要基于数据样本的标签数据和预测结果来确定一些度量参数的值。

- 除 AUC 之外, 预测模型 125 的性能指标还可以包括精确率 (Precision), 其被表示为 $\text{Precision} = \text{TP} / (\text{TP} + \text{FP})$ 。精确度表示, 被预测为正样本的数据样本 102 子集中, 由标签标注为正样本的概率。预测模型 125 的性能指标还可以包括召回率 (Recall), 其被表示为 $\text{Recall} = \text{TP} / (\text{TP} + \text{FN})$, 即正样本被预测的概率。预测模型 125 的性能指标还可以包括 P-R 曲线, 其以召回率为横轴, 精确度为纵轴。P-R 曲线越靠近右上角, 说明模型的性能越好。曲线下面积称作 AP 分数 (Average Precision Score, 平均精确率分数)。

在下文中, 将主要以 AUC 的确定作为示例来进行讨论。AUC 可以具有不同计算方式, 在不同计算方式下需要的度量参数也不同。

- 图 3A 示出了根据本公开的一些实施例的确定度量参数的值的过
25 程 300 的流程图。过程 300 用于确定在 AUC 的一种计算方式中需要的度量参数的值。过程 300 可以在客户端节点 110 处实现。

在框 310, 客户端节点 110 确定多个真值标签 105 中第一类标签的数目 (称为“第一数目”), 这里的第一类标签指示对应的数据样本 102 属于第一类别, 例如指示数据样本 102 是正样本。在框 320,

客户端节点 110，客户端节点 110 还可以确定多个真值标签 105 中的第二类标签的数目（称为“第二数目”），这里的第二类标签指示对应的数据样本 102 属于第二类别，例如指示数据样本 102 是负样本。

在客户端节点 110-k 处，对第一数目和第二数目的确定可以被表示如下：

$$\text{localP}_k = \sum_{i=1}^{|X_k|} y_i^k \quad (1)$$

$$\text{localN}_k = |X_k| - \text{localP}_k \quad (2)$$

其中 $|X_k|$ 表示客户端节点 110-k 的数据样本 102 的总数目； y_i^k 表示第 i 个数据样本 102 对应的真值标签 105 的值； localP_k 表示在客户端节点 110-k 处第一类标签（指示正样本的标签）的数目， localN_k 表示在客户端节点 110-k 处真值标签 105 中第二类标签（指示正样本的标签）的数目。

在上式（1）和（2）中，假设对于正样本， y_i^k 的取值为 1，对于负样本， y_i^k 的取值为 0，这样通过对 y_i^k 的加和，可以统计出由真值标签 105 指示的正样本的数目。除正样本之外的其他样本是由真值标签 105 指示的负样本的数目。在其他示例中，如果真值标签 105 用其他取值来指示正样本和负样本，还可以通过其他方式来统计 localP_k 和 localN_k ，本文对此不做限制。 localP_k 和 localN_k 可以被确定为客户端节点 110-k 处的度量信息中的两个度量参数的值。

在一些实施例中，在框 330，客户端节点 110 客户端节点 110 还可以基于多个预测得分在所有客户端节点的预测得分中的排序结果，确定第一类标签对应的数据样本 102（即，正样本）的预测得分在预测得分集合中超过的预测得分的数目（称为第三数目）。三数目可以作为性能指标的另一个度量参数的值。这个数目可以指示在数据样本 102 的总集合中，正样本的排序高于其余样本的排序的样本对个数（在升序排序的情况下）。

对于个体客户端节点 110，为了获得它的预测得分在所有客户端节点的预测得分中的排序结果，客户端节点 110 要与服务节点 120 进行信令交互。图 3B 示出根据本公开的一些实施例的确定度量参数的值的信令流 350 的流程图。

5 在信令流 350 中，客户端节点 110 将预测模型 125 输出的预测得分发送 352 给服务节点 120 进行排序。

在一些实施例中，在向服务节点 120 发送预测得分之前，客户端节点 110 可以随机调整多个预测得分的顺序，并按调整后的顺序将多个预测得分发送给服务节点。通过随机调整顺序，可以避免在一些特殊情况下，在客户端节点处将多个数据样本 102 顺序输入预测模型 125 后，所输出的预测得分具有一定的顺序，例如从大到小或从小到大的顺序，这可能会导致一定的信息泄露。随机顺序调整可以进一步加强数据隐私保护。

15 在接收 354 预测得分后，服务节点 120 对来自多个客户端节点 110 的预测得分集合进行排序 356，从而得到来自每个客户端节点 110 的预测得分在预测得分集合中的排序结果。

在一些实施例中，服务节点 120 可以对预测得分集合升序排序，并对每个预测得分 s_i^k （客户端节点 110-k 的第 i 个数据样本 102 的预测得分）分配排序值 r_i^k 。在一些实施例中，排序值 r_i^k 可以指示预测得分 s_i^k 在预测得分集合中超过的其他预测得分的数目。例如，按升序排序，最低的预测得分被分配排序值 0，指示其未超过（大于）任何其他预测得分；下一个预测得分被分配排序值 1，指示其大于集合中的 1 个预测得分，以此类推。这样的排序值的分配有利于后续的计算。

25 对于接收到预测得分的客户端节点 110，服务节点 120 将它的多个预测得分在整体的预测得分集合中的排序结果发送 358 给对应的客户端节点 110。在单个客户端节点 110 处，基于从服务节点 120 接收 360 的排序结果，客户端节点 110 可以确定 362 第一类标签对应的数据样本 102（即，正样本）的预测得分在预测得分集合中超过的预测

得分的第三数目。在一些实施例中，在客户端节点 110-k 处，第三数目可以通过以下来确定：

$$\text{localSum}_k = \sum_{i=1}^{|X_k|} y_i^k r_i^k \quad (3)$$

其中 localSum_k 表示第三数目， y_i^k 表示第 i 个数据样本 102 对应的真值标签 105 的值， r_i^k 表示第 i 个数据样本 102 对应的预测得分的排序值。如前所述，排序值 r_i^k 可以被设置为指示预测得分 s_i^k 在预测得分集合中超过的其他预测得分的数目。在上式(3)中，也假设对于正样本， y_i^k 的取值为 1，对于负样本， y_i^k 的取值为 0。这样，通过对 $y_i^k r_i^k$ 的加和，可以确定出正样本的预测得分排序超过其余样本的预测得分排序的样本数目（也是这样的预测得分的数目）。 localSum_k 可以被确定为客户端节点 110-k 处的度量信息中的另一个度量参数的值（差错值）。

localP_k 、 localN_k 和 localSum_k 均是在 AUC 的示例计算方式中需要确定的度量参数。假设 N 个客户端节点 110 处第一类标签的总数目是 $\bar{P}$ 、第一类标签的总数目是 $\bar{N}$ ，并且第一类标签对应的数据样本 102（即，正样本）的预测得分在预测得分集合中超过的预测得分的总数目是 globalSum ，那么可以通过以下方式计算模型的 AUC 的值：

$$\text{AUC} = \frac{\text{globalSum} - \frac{\bar{P}(\bar{P}-1)}{2}}{\bar{P}\bar{N}} \quad (4)$$

在一些实施例中，还可以通过其他方式来计算 AUC，并且 AUC 的计算需要其他的度量参数。在一些实施例中，客户端节点 110 可以确定在本地的多个真值标签 105 中，由真值标签 105 指示的正样本的数目和由真值标签 105 指示的负样本的数目。此外，客户端节点 110 可以确定基于预测得分集合，确定在全部数据样本 102 中，正样本的预测得分大于负样本的预测得分的数目。基于这三个数目

可以是计算 AUC 的值需要的度量参数的值。各个客户端节点 110 可以确定在各自的数据集合上所统计出的这些度量参数的值。

假设 N 个客户端节点 110 处的数据样本 102 的总数是 L，并且其中由真值标签 105 指示的正样本的数目为 m 个，负样本为 n 个。此外，每个数据样本 102 对应的预测得分为 s_i , $i \in [1, L]$ 。通过遍历正样本和负样本的两两组合，可以形成 $m*n$ 个样本对 P_i , $i \in [1, m*n]$ ，那么 AUC 可以被确定如下：

$$\text{AUC}_{\text{corr}} = \frac{\sum I(P_i)}{mn}, \text{ 其中} \quad (5)$$

$$I(P_i) = \begin{cases} 1, & P_i \text{中正样本的预测得分大于负样本的预测得分} \\ 0, & \text{其他情况} \end{cases}$$

以上讨论了对于 AUC 的一些示例计算方式。如果适应的话，还可以基于其他方式从概率统计的角度来确定 AUC。

在一些实施例中，除 AUC 之外，还可以评估预测模型 125 的其他性能指标，只要这样的性能指标是能够从多个预测得分和多个真值标签 105 中确定出。相应的，客户端节点 110 可以根据性能指标的类型和计算方式，在本地的预测得分和真值标签中确定性能指标相关的度量参数的值。本公开的实施例在此方面不受限制。

在本公开的实施例中，为了在确定预测模型 125 的性能指标的同时实现对真值标签的隐私保护，不是直接将客户端节点 110 确定的多个度量参数的值发送给服务节点 120。相反，在信令流 200 中，在客户端节点 110 处确定出多个度量参数的值后，客户端节点 110 对所述多个度量参数的值施加 225 扰动，得到多个度量参数的扰动值。客户端节点 110 将多个度量参数的扰动值发送 230 给服务节点 120。

在本公开的实施例中，通过施加扰动，可以避免暴露基于真值标签统计出的度量参数的真实值。下文将详细讨论客户端节点如何施加扰动。在本文中，“扰动”有时也称为噪音、干扰等。

在一些实施例中，期望所施加的扰动能够满足数据的差分隐私的

保护。为更好地理解本公开的实施例，下文将首先简单介绍差分隐私和随机响应机制。

假设 ϵ, δ 是大于等于0的实数，即 $\epsilon, \delta \in \mathbb{R}_{\geq 0}$ ，并且 $\mathcal{M}$ 是一个随机机制（随机算法）。所谓随机机制，指的是对于特定输入，该机制的输出不是固定值，而是服从某一分布。对于随机机制 $\mathcal{M}$ ，如果满足以下情况则可以认为随机机制 $\mathcal{M}$ 具有 (ϵ, δ) -差分隐私：对于任意两个相邻训练数据集 D, D' ，并且对于 $\mathcal{M}$ 的可能的输出的任意子集 S ，存在：

$$\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{M}(D') \in S] + \delta \quad (6)$$

10

此外，如果 $\delta = 0$ ，还可以认为随机机制 $\mathcal{M}$ 具有 ϵ -差分隐私（ ϵ -DP）。在差分隐私机制中，对于具有 (ϵ, δ) -差分隐私或 ϵ -差分隐私的随机机制 $\mathcal{M}$ ，期望其分别作用于两个相邻数据集后得到的两个输出的分布难以区分。这样的话，观察者可以通过观察输出结果，很难察觉到算法的输入数据集中的微小变化，从而达到保护隐私的目的。如果随机机制 $\mathcal{M}$ 作用于任何相邻数据集，都得到特定输出 S 的概率差不多，那么将会认为该算法难以达到差分隐私的效果。

在本公开的实施例中，关注于对数据样本的标签的差分隐私，且标签指示二分类结果。因此，遵循差分隐私的设置，可以定义标签差分隐私。具体地，假设 ϵ, δ 是大于等于0的实数，即 $\epsilon, \delta \in \mathbb{R}_{\geq 0}$ ，并且 $\mathcal{M}$ 是一个随机机制（随机算法）。如果满足以下情况则可以认为随机机制 $\mathcal{M}$ 具有 (ϵ, δ) -标签差分隐私（label differential privacy）：对于任意两个相邻训练数据集 D, D' ，它们的差异仅在于单个数据样本的标签不同，并且对于 $\mathcal{M}$ 的可能的输出的任意子集 S ，存在：

25

$$\Pr[\mathcal{M}(D) \in S] \leq e^\epsilon \cdot \Pr[\mathcal{M}(D') \in S] + \delta \quad (7)$$

此外，如果 $\delta = 0$ ，还可以认为随机机制 $\mathcal{M}$ 具有 ϵ -标签差分隐私（ ϵ -

标签 DP)。也就是说,期望在改变数据样本的标签后,从随机机制 $\mathcal{M}$ 的输出结果的分布仍较小,使得观察者难以察觉到标签的改变。

随机机制可以服从一定的概率分布。在一些实施例,可以基于高斯 (Gaussian) 分布或拉普拉斯 (Laplace) 分布来施加扰动。在一些实施例,客户端节点 110 通过确定要扰动的度量参数值的敏感度值,并且基于灵敏度值确定概率分布,用于向各个度量参数的值施加随机扰动。下面首先介绍灵敏度,在介绍如果在灵敏度基础上确定概率分布。

假设 d 是正整数, D 是数据集的合集, 并且 $f: D \rightarrow R^d$ 是函数, 用于从 D 变化到 R^d 。函数的灵敏度可以表示为 Δf , 其可以被定义为 $\Delta f = \max ||f(D_1) - f(D_2)||_1$, 其表示在 D 中的数据集 D_1 和 D_2 的所有配对中的最大值, D_1 和 D_2 相差最多一个数据元素, 并且 $||\cdot||_1$ 表示 l_1 范数。根据上述定义可知, 灵敏度指的是在变化最多一个数据元素的情况下, 函数输出的最大差异是多少。

在不同类型的概率分布中, 均可以引入灵敏度值来定义具体的概率分布方式。例如, 对于高斯分布机制, 对于任意的 $(\epsilon, \delta) \in (0, 1)$, 高斯分布具有的标准差可以被定义为 $std = \Delta \sqrt{2 \log (1.25/\delta)}/\epsilon$, 其中 Δ 表示灵敏度值。这样的高斯分布具有 (ϵ, δ) 的差分隐私 (即, (ϵ, δ) - DP)。

又例如, 对于拉普拉斯分布机制 (其以 0 为中心) 并且宽度 (scale) 为 b , 其概率密度函数被表示为: $Lap(x|b) = \frac{1}{2b} \exp(-\frac{|x|}{b})$ 。如果随机噪声 (随机扰动) 是由 $Lap(\Delta/\epsilon)$ 的拉普拉斯分布确定的, 那么可以认为这样的概率分布能够提供 $(\epsilon, 0)$ 的差分隐私。

基于上述讨论, 客户端节点 110 在施加扰动时, 可以确定与不同度量参数的扰动分别相关的灵敏度值, 并且基于灵敏度值和差分隐私机制, 确定对应的概率分布。客户端节点 110 可以按照概率分布来向对应的度量参数施加扰动值。

在一些实施例, 对于在客户端节点 110-k 处确定的多个真值标

签 105 中第一类标签的数目 localP_k 和第二类标签的数目 localN_k ，客户端节点 110-k 可以向其中任意一个数值施加扰动即可。因为另一个值可以通过将该节点处的真值标签 105 的总数减去前一数目的扰动值来确定。

5 具体地，客户端节点 110-k 确定与 localP_k 和 localN_k 的扰动相关的灵敏度值。对于标签的数目的值，可以看出，如果随机改变一个真值标签的值， localP_k 或 localN_k 最多会改变 1。因此，此处的灵敏度值可以被确定为 $\Delta = 1$ 。基于该灵敏度值，客户端节点 110-k 可以根据差分隐私机制，确定扰动所要遵循的概率分布。

10 在一些示例中，可以确定高斯分布，并基于高斯分布来施加扰动（也称为噪音或高斯噪音）。根据以上讨论的高斯分布机制，对于任意的 $(\epsilon, \delta) \in (0, 1)$ ，高斯分布的标准差可以被确定为 $\sigma = \Delta \sqrt{2 \log(1.25/\delta)}/\epsilon$ ，其中灵敏度值 $\Delta = 1$ 。这样的高斯分布机制可以满足 (ϵ, δ) 的差分隐私（即 $(\epsilon, \delta) - \text{DP}$ ）。

15 在一些示例中，可以确定拉普拉斯分布，并基于拉普拉斯分布来施加扰动（也称为噪音或拉普拉斯噪音）。如果要满足 $(\epsilon, 0)$ 的差分隐私，拉普拉斯分布的宽度可以被确定为 $b = \Delta/\epsilon$ ，即，从 $\text{Lap}(\Delta/\epsilon)$ 的分布中施加随机噪音。该分布的标准差为 $\sigma = \sqrt{2}b = \sqrt{2}\Delta/\epsilon$ 。

20 在一些实施例中，对于在客户端节点 110-k 处确定的第一类标签对应的数据样本 102（即，正样本）的预测得分在预测得分集合中超过的预测得分的第三数目 localSum_k ，客户端节点 110-k 可以确定该度量参数的扰动的灵敏度值。

25 在一些实施例中，与 localSum_k 有关的灵敏度值由服务节点 120 从全局的角度确定。这样的扰动方式可以获得全局隐私保护。从全局角度看，对于第一类标签对应的正样本的预测得分在预测得分集合中超过的预测得分的数目，其敏感度可以被确定为 $\Delta = Q - 1$ ，其中 Q 是表示 N 个客户端节点 110 处的数据样本的总数目。这里的敏感度值表示，如果改变数据样本的全集中的一个数据样本，那么 localSum_k 最大会改 $Q - 1$ 。在一些实施例中，对于某个客户端节点 110-k，其可以从服务

节点 120 接收与灵敏度值相关的信息。该信息可以是多个客户端节点的数据样本的总数目 Q ，也可以直接接收敏感度值 $Q-1$ ，或者其他能够确定灵敏度值的信息。

在一些实施例中，与 localSum_k 有关的灵敏度值由各个客户端节点 110-k 在本地确定。这样的扰动方式可以获得局部隐私保护。具体地，客户端节点 110-k 从本地的多个预测得分各自的排序结果中确定最高排序结果，并基于最高排序结果来确定灵敏度值，其可以被表示为 $\Delta_k = \max(r_i^k)$ 。也就是说，对于各个客户端节点处的本地数据集，如果改变其中一个数据样本，那么 localSum_k 最大会改变的数值与其对应的预测得分在整体预测得分集中的最高排序有关。

在客户端节点 110-k 以任意方式确定出与 localSum_k 有关的灵敏度值后，基于该灵敏度值，客户端节点 110-k 可以根据差分隐私机制，确定对 localSum_k 施加扰动所要遵循的概率分布。

在一些示例中，可以确定高斯分布，并基于高斯分布来施加扰动（也称为噪音或高斯噪音）。根据以上讨论的高斯分布机制，对于任意的 $(\epsilon, \delta) \in (0, 1)$ ，高斯分布的标准差可以被确定为 $\sigma = \Delta \sqrt{2 \log(1.25/\delta)}/\epsilon$ 。这样的高斯分布机制可以满足 (ϵ, δ) 的差分隐私（即 (ϵ, δ) -DP）。

在一些示例中，可以确定拉普拉斯分布，并基于拉普拉斯分布来施加扰动（也称为噪音或拉普拉斯噪音）。如果要满足 $(\epsilon, 0)$ 的差分隐私，拉普拉斯分布的宽度可以被确定为 $b = \Delta/\epsilon$ ，即，从 $\text{Lap}(\Delta/\epsilon)$ 的分布中施加随机噪音。该分布的标准差为 $\sigma = \sqrt{2}b = \sqrt{2}\Delta/\epsilon$ 。

以上讨论了对于一些示例度量参数施加扰动的方式。对于其他不同的度量参数，也可以以类似的方式，确定灵敏度值和对应的概率分布，用于相应地施加扰动。这里不再赘述。

通过施加随机扰动，从本地的真值标签确定出的度量参数的值不需要被暴露。客户端节点 110 可以向服务节点 120 发送度量参数的扰动值。

继续参考图 2，服务节点 120 从多个客户端节点 110 接收 235 到

它们各自提供的多个度量参数的扰动值。服务节点 120 按度量参数聚合 240 来自多个客户端节点 110 的多个度量参数的扰动值，得到多个度量参数的聚合值。服务节点 120 基于多个度量参数的聚合值来确定 245 预测模型 125 的性能指标的值。

- 5 在一些实施例中，假设从客户端节点 110 接收到的度量参数的扰动值为 $\text{localSum}'_k$, localP'_k , localN'_k , 服务节点 120 将各个客户端节点 110 的这些度量参数的值聚合（例如，加和到一起），可以分别得到：

$$\text{globalSum} = \sum_k \text{localSum}'_k \quad (8)$$

$$\bar{P} = \sum_k \text{localP}'_k \quad (9)$$

$$\bar{N} = \sum_k \text{localN}'_k \quad (10)$$

- 由于是基于扰动值确定的，这些聚合值可能不完全等同于从多个客户端节点 110 处的真值标签和预测得分统计出的值。然而，由于多个客户端节点 110 均各自应用了随机扰动，在一些实施例中，客户端节点 110 应用的随机扰动（例如，概率分布）的均值为 0。这样，通过服务节点 120 处的聚合操作，各个客户端节点 110 的随机扰动可以被互相抵消，使得这些聚合值近似于这些度量参数的真实值。

- 15 在一些实施例中，服务节点 120 可以根据上式（4）来计算预测模型 125 的 AUC 的值。在一些实施例中，对于从客户端节点 110 获得的其他度量参数的扰动值，服务节点 120 也可以类似聚合以用于计算性能指标。例如，对于上式（5）给出的 AUC 的计算方式，服务节点 120 可以从客户端节点 110 接收对应的度量参数的扰动值，并聚合后用于计算 AUC。

- 20 虽然在均值为 0 的概率分布下，聚合操作可以抵消方差，但可能与性能指标的真实值相比，所确定的值还会存在一定方差。但发明人通过反复试验和验证，发现这样的方差较小，在允许范围内。特别是随着参与的客户端节点的数目增加，方差会更小。

实际上，严格来说，即使拥有真值标签，在计算 AUC 的很多算

法中，均是通过近似的方式去逼近 AUC 的真实值，即 ROC 曲线下方面积。因此，在需要对标签数据进行隐私保护的场景中，根据本公开的各个实施例，能够在获得数据的差分隐私保护的同时，允许服务节点确定出较准确的性能指标。

- 5 下文将讨论根据本公开的一些实施例计算的性能指标（以 AUC 为例）的值与真实值之间的误差。为了方便，暂时假设在计算 AUC 时，服务节点使用的是基于真实标签统计的值，例如真实的正样本数目和负样本数目。

- 在基于全局隐私保护来施加扰动时，所计算的 AUC 的标准差是
- 10 $std(AUC) = std\left(\frac{\sum sum_i - \frac{P*(P-1)}{2}}{P*N}\right) = \frac{\sqrt{c}*\sigma}{P*N} = \frac{\sigma}{P*N/\sqrt{c}}$ ，其中 P 是正样本数目，N 是负样本数目，c 是客户端节点数目， σ 是所添加的扰动（噪声）的标准差。从上式可以看出，随着客户端数目的减少，所添加的噪声的减少，所计算的 AUC 的标准差也会减少。在全局隐私保护中应用拉普拉斯机制为例， $\sigma = \sqrt{2}(M-1)/\epsilon$ ，M 为客户端节点的总数目。因此，
- 15 AUC 的方差可以被计算为 $var(AUC) = (\frac{\sqrt{2}(M-1)/\epsilon}{P*N/\sqrt{c}})^2$ ，标准差为 $std(AUC) = \frac{\sqrt{2}(M-1)/\epsilon}{P*N/\sqrt{c}}$ 。假设每个客户端节点上只有一个数据样本，即 $c=M$ ，那么标准差为 $\frac{\sqrt{2}(M-1)/\epsilon}{P*N/\sqrt{M}}$ 。该值在一般应用中是较小值。

- 在基于局部隐私保护来施加扰动时，为方便计算，假设在一种极端情况下，每个客户端节点上只有一个数据样本，即 $c=M$ ，那么这些
- 20 客户端节点所使用的灵敏度值分别为 $[0, 1, 2, \dots, M-1]$ ，仍以拉普拉斯分布为例，所计算的 AUC 的方差如下，其在一般应用中也是较小值：

$$\begin{aligned} \text{Variance}(AUC) &= \frac{\sum_{i=1}^c \sigma_i^2}{P^2 * N^2} = \frac{\sum_{i=1}^c 2 * (i-1)^2 / \epsilon^2}{P^2 * N^2} = \frac{2 * M(M-1)(2M-1)/6/\epsilon^2}{P^2 * N^2} = \\ &= \frac{M(M-1)(2M-1)/3/\epsilon^2}{P^2 * N^2} \end{aligned}$$

- 应当理解，虽然以 AUC 为例进行说明，在一些实施例中，还可以以类似的扰动和交互方式，使服务节点 120 能够附加地或备选地计
- 25 算其他性能指标的值。

图4示出根据本公开的一些实施例的在客户端节点处用于模型性能评估的过程400的流程图。过程400可以被实现在客户端节点110处。

在框410,客户端节点110在客户端节点处,将多个数据样本分别应用到预测模型,以得到预测模型输出的多个预测得分。多个预测得分分别指示多个数据样本属于第一类别或第二类别的预测概率。在框420,客户端节点110基于多个数据样本的多个真值标签和多个预测得分,确定与预测模型的预定性能指标相关的多个度量参数的值。在框430,客户端节点110对多个度量参数的值施加扰动,得到多个度量参数的扰动值。在框440,客户端节点110将多个度量参数的扰动值发送给服务节点。

在一些实施例中,确定多个度量参数的值包括:确定多个真值标签中的第一类标签的第一数目,以作为第一度量参数的值,第一类标签指示对应的数据样本属于第一类别;以及确定多个真值标签中的第二类标签的第二数目,以作为第二度量参数的值,第二类标签指示对应的数据样本属于第二类别。

在一些实施例中,对多个度量参数的值施加扰动包括:确定与第一度量参数和第二度量参数中的一个度量参数的扰动相关的第一灵敏度值;基于第一灵敏度值和差分隐私机制,来确定第一概率分布;基于第一概率分布,对第一度量参数和第二度量参数中的一个度量参数的值施加扰动,得到一个度量参数的扰动值;以及基于多个真值标签的总数目和第一度量参数和第二度量参数中的度量参数的扰动值,确定第一度量参数和第二度量参数中的另一度量参数的扰动值。

在一些实施例中,确定多个度量参数的值包括:将多个预测得分发送给服务节点;从服务节点接收多个预测得分各自在预测得分集合中的排序结果,预测得分集合包括由多个客户端节点发送的预测得分,多个客户端节点包括客户端节点;以及基于多个预测得分各自的排序结果,确定第一类标签对应的数据样本的预测得分在预测得分集合中超过的预测得分的第三数目,以作为第三度量参数的值。

在一些实施例中，对多个度量参数的值施加扰动包括：确定与第三度量参数的扰动相关的第二灵敏度值；基于第二灵敏度值和差分隐私机制，来确定第二概率分布；以及基于第二概率分布，对第三度量参数的值施加扰动。

- 5 在一些实施例中，确定第二灵敏度值包括：从服务节点接收与第二灵敏度值相关的信息；以及基于所接收的信息来确定第二灵敏度值。

在一些实施例中，与第二灵敏度值相关的信息包括多个客户端节点的数据样本的总数目。

- 10 在一些实施例中，确定第二灵敏度值包括：从多个预测得分各自的排序结果中确定最高排序结果；以及基于最高排序结果来确定第二灵敏度值。

在一些实施例中，预定性能度量指标至少包括受试者工作特征曲线（ROC）的曲线下面积（AUC）。

- 15 图 5 示出根据本公开的一些实施例的在服务节点处用于模型性能评估的过程 500 的流程图。过程 500 可以被实现在服务节点 120 处。

- 在框 510，服务节点 120 从多个客户端节点分别接收与预测模型的预定性能指标相关的多个度量参数的扰动值。在框 520，服务节点 120 按度量参数聚合来自多个客户端节点的多个度量参数的扰动值，得到多个度量参数的聚合值。在框 530，服务节点 120 基于多个度量
20 参数的聚合值来确定预定性能指标的值。

- 在一些实施例中，对于多个客户端节点中的给定客户端节点，多个度量参数的扰动值指示以下至少一项：给定客户端节点处的多个真值标签中的第一类标签的第一数目，第一类标签指示对应的数据样本属于第一类别；多个真值标签中的第二类标签的第二数目，以作为第二
25 度量参数的值，第二类标签指示对应的数据样本属于第二类别；以及给定客户端节点处的第一类标签对应的数据样本的预测得分在预测得分集合中超过的预测得分的第三数目，预测得分由预测模型基于数据样本确定，并且预测得分集合包括由多个客户端节点发送的预测得分。

在一些实施例中，过程 500 还包括：向多个客户端节点分别发送与第二灵敏度值相关的信息。

在一些实施例中，与第二灵敏度值相关的信息包括多个客户端节点的数据样本的总数目。

- 5 在一些实施例中，预定性能度量指标至少包括受试者工作特征曲线（ROC）的曲线下面积（AUC）。

图 6 示出了根据本公开的一些实施例的在客户端节点处用于模型性能评估的装置 600 的框图。装置 600 可以被实现为或者被包括在客户端节点 110 中。装置 600 中的各个模块/组件可以由硬件、软件、固
10 件或者它们的任意组合来实现。

如图所示，装置 600 包括预测模块 610，被配置为将多个数据样本分别应用到预测模型，以得到预测模型输出的多个预测得分，多个预测得分分别指示多个数据样本属于第一类别或第二类别的预测概率。装置 600 还包括度量确定模块 620，被配置为基于多个数据样本
15 的多个真值标签和多个预测得分，确定与预测模型的预定性能指标相关的多个度量参数的值。装置 600 还包括扰动模块 630，被配置为对多个度量参数的值施加扰动，得到多个度量参数的扰动值；以及发送模块 640，被配置为将多个度量参数的扰动值发送给服务节点。

在一些实施例中，度量确定模块 620 包括：第一确定模块，被配置为确定多个真值标签中的第一类标签的第一数目，以作为第一度量参数的值，第一类标签指示对应的数据样本属于第一类别；以及第二
20 确定模块，被配置为确定多个真值标签中的第二类标签的第二数目，以作为第二度量参数的值，第二类标签指示对应的数据样本属于第二类。

25 在一些实施例中，扰动模块包括：第一灵敏度确定模块，被配置为确定与第一度量参数和第二度量参数中的一个度量参数的扰动相关的第一灵敏度值；第一分布确定模块，被配置为基于第一灵敏度值和差分隐私机制，来确定第一概率分布；第一扰动施加模块，被配置为基于第一概率分布，对第一度量参数和第二度量参数中的一个度量

参数的值施加扰动，得到一个度量参数的扰动值；以及扰动值确定模块，被配置为基于多个真值标签的总数目和第一度量参数和第二度量参数中的度量参数的扰动值，确定第一度量参数和第二度量参数中的另一度量参数的扰动值。

- 5 在一些实施例中，度量确定模块包括：得分发送模块，被配置为将多个预测得分发送给服务节点；结果接收模块，被配置为从服务节点接收多个预测得分各自在预测得分集合中的排序结果，预测得分集合包括由多个客户端节点发送的预测得分，多个客户端节点包括客户端节点；以及第三确定模块，被配置为基于多个预测得分各自的排序
10 结果，确定第一类标签对应的数据样本的预测得分在预测得分集合中超过的预测得分的第三数目，以作为第三度量参数的值。

- 在一些实施例中，扰动模块包括：第二灵敏度确定模块，被配置为确定与第三度量参数的扰动相关的第二灵敏度值；第二分布确定模块，被配置为基于第二灵敏度值和差分隐私机制，来确定第二概率分布；以及第二扰动施加模块，被配置为基于第二概率分布，对第三度
15 量参数的值施加扰动。

在一些实施例中，第二灵敏度确定模块包括：灵敏度接收模块，被配置为从服务节点接收与第二灵敏度值相关的信息；以及基于信息的确定模块，被配置为基于所接收的信息来确定第二灵敏度值。

- 20 在一些实施例中，与第二灵敏度值相关的信息包括多个客户端节点的数据样本的总数目。

在一些实施例中，第二灵敏度确定模块包括：排序确定模块，被配置为从多个预测得分各自的排序结果中确定最高排序结果；以及基于排序的确定模块，被配置为基于最高排序结果来确定第二灵敏度值。

- 25 在一些实施例中，预定性能度量指标至少包括受试者工作特征曲线（ROC）的曲线下面积（AUC）。

图7示出了根据本公开的一些实施例的在客户端节点处用于模型性能评估的装置700的框图。装置700可以被实现为或者被包括在服务节点120中。装置700中的各个模块/组件可以由硬件、软件、固件

或者它们的任意组合来实现。

如图所示，装置 700 包括接收模块 710，被配置为从多个客户端节点分别接收与预测模型的预定性能指标相关的多个度量参数的扰动值。装置 700 还包括聚合模块 720，被配置为按度量参数聚合来自
5 多个客户端节点的多个度量参数的扰动值，得到多个度量参数的聚合值。装置 700 还包括性能确定模块 730，被配置为基于多个度量参数的聚合值来确定预定性能指标的值。

在一些实施例中，对于多个客户端节点中的给定客户端节点，多个度量参数的扰动值指示以下至少一项：给定客户端节点处的多个真
10 值标签中的第一类标签的第一数目，第一类标签指示对应的数据样本属于第一类别；多个真值标签中的第二类标签的第二数目，以作为第二度量参数的值，第二类标签指示对应的数据样本属于第二类别；以及给定客户端节点处的第一类标签对应的数据样本的预测得分在预测得分集合中超过的预测得分的第三数目，预测得分由预测模型基于
15 数据样本确定，并且预测得分集合包括由多个客户端节点发送的预测得分。

在一些实施例中，装置 700 还包括：灵敏度发送模块，被配置为向多个客户端节点分别发送与第二灵敏度值相关的信息。

在一些实施例中，与第二灵敏度值相关的信息包括多个客户端节
20 点的数据样本的总数目。

在一些实施例中，预定性能度量指标至少包括受试者工作特征曲线（ROC）的曲线下面积（AUC）。

图 8 示出了能够实施本公开的一个或多个实施例的计算设备/系统 800 的框图。应当理解，图 8 所示出的计算设备/系统 800 仅仅是示
25 例性的，而不应当构成对本文所描述的实施例的功能和范围的任何限制。图 8 所示出的计算设备/系统 800 可以用于实现图 1 的客户端节点 110 或服务节点 120。

如图 8 所示，计算设备/系统 800 是通用计算设备的形式。计算设备/系统 800 的组件可以包括但不限于一个或多个处理器或处理单元

810、存储器 820、存储设备 830、一个或多个通信单元 840、一个或多个输入设备 850 以及一个或多个输出设备 860。处理单元 810 可以是实际或虚拟处理器并且能够根据存储器 820 中存储的程序来执行各种处理。在多处理器系统中，多个处理单元并行执行计算机可执行指令，以提高计算设备/系统 800 的并行处理能力。

计算设备/系统 800 通常包括多个计算机存储介质。这样的介质可以是计算设备/系统 800 可访问的任何可以获得的介质，包括但不限于易失性和非易失性介质、可拆卸和不可拆卸介质。存储器 820 可以是易失性存储器（例如寄存器、高速缓存、随机访问存储器（RAM））、非易失性存储器（例如，只读存储器（ROM）、电可擦除可编程只读存储器（EEPROM）、闪存）或它们的某种组合。存储设备 830 可以是可拆卸或不可拆卸的介质，并且可以包括机器可读介质，诸如闪存驱动、磁盘或者任何其他介质，其可以能够用于存储信息和/或数据（例如用于训练的训练数据）并且可以在计算设备/系统 800 内被访问。

计算设备/系统 800 可以进一步包括另外的可拆卸/不可拆卸、易失性/非易失性存储介质。尽管未在图 8 中示出，可以提供用于从可拆卸、非易失性磁盘（例如“软盘”）进行读取或写入的磁盘驱动和用于从可拆卸、非易失性光盘进行读取或写入的光盘驱动。在这些情况中，每个驱动可以由一个或多个数据介质接口被连接至总线（未示出）。

存储器 820 可以包括计算机程序产品 825，其具有一个或多个程序模块，这些程序模块被配置为执行本公开的各种实施例的各种方法或动作。

通信单元 840 实现通过通信介质与其他计算设备进行通信。附加地，计算设备/系统 800 的组件的功能可以以单个计算集群或多个计算机器来实现，这些计算机器能够通过通信连接进行通信。因此，计算设备/系统 800 可以使用与一个或多个其他服务器、网络个人计算机（PC）或者另一个网络节点的逻辑连接来在联网环境中进行操作。

输入设备 850 可以是一个或多个输入设备，例如鼠标、键盘、追踪球等。输出设备 860 可以是一个或多个输出设备，例如显示器、扬

声器、打印机等。计算设备/系统 800 还可以根据需要通过通信单元 840 与一个或多个外部设备（未示出）进行通信，外部设备诸如存储设备、显示设备等，与一个或多个使得用户与计算设备/系统 800 交互的设备进行通信，或者与使得计算设备/系统 800 与一个或多个其他
5 计算设备通信的任何设备（例如，网卡、调制解调器等）进行通信。这样的通信可以经由输入/输出（I/O）接口（未示出）来执行。

根据本公开的示例性实现方式，提供了一种计算机可读存储介质，其上存储有计算机可执行指令或计算机程序，其中计算机可执行指令或计算机程序被处理器执行以实现上文描述的方法。

10 根据本公开的示例性实现方式，还提供了一种计算机程序产品，计算机程序产品被有形地存储在非瞬态计算机可读介质上并且包括计算机可执行指令，而计算机可执行指令被处理器执行以实现上文描述的方法。

在本文中参照根据本公开实现的方法、装置、设备和计算机程序
15 产品的流程图和/或框图描述了本公开的各个方面。应当理解，流程图和/或框图的每个方框以及流程图和/或框图中各方框的组合，都可以由计算机可读程序指令实现。

这些计算机可读程序指令可以提供给通用计算机、专用计算机或其他可编程数据处理装置的处理单元，从而生产出一种机器，使得这
20 些指令在通过计算机或其他可编程数据处理装置的处理单元执行时，产生了实现流程图和/或框图中的一个或多个方框中规定的功能/动作的装置。也可以把这些计算机可读程序指令存储在计算机可读存储介质中，这些指令使得计算机、可编程数据处理装置和/或其他设备以特定方式工作，从而，存储有指令的计算机可读介质则包括一个制造品，
25 其包括实现流程图和/或框图中的一个或多个方框中规定的功能/动作的各方面的指令。

可以把计算机可读程序指令加载到计算机、其他可编程数据处理装置、或其他设备上，使得在计算机、其他可编程数据处理装置或其他设备上执行一系列操作步骤，以产生计算机实现的过程，从而使得

在计算机、其他可编程数据处理装置、或其他设备上执行的指令实现流程图和/或框图中的一个或多个方框中规定的功能/动作。

附图中的流程图和框图显示了根据本公开的多个实现的系统、方法和计算机程序产品的可能实现的体系架构、功能和操作。在这点上，
5 流程图或框图中的每个方框可以代表一个模块、程序段或指令的一部分，模块、程序段或指令的一部分包含一个或多个用于实现规定的逻辑功能的可执行指令。在有些作为替换的实现中，方框中所标注的功能也可以以不同于附图中所标注的顺序发生。例如，两个连续的方框实际上可以基本并行地执行，它们有时也可以按相反的顺序执行，这
10 依所涉及的功能而定。也要注意的，框图和/或流程图中的每个方框、以及框图和/或流程图中的方框的组合，可以用执行规定的功能或动作的专用的基于硬件的系统来实现，或者可以用专用硬件与计算机指令的组合来实现。

以上已经描述了本公开的各实现，上述说明是示例性的，并非穷
15 尽性的，并且也不限于所公开的各实现。在不偏离所说明的各实现的范围和精神的情况下，对于本技术领域的普通技术人员来说许多修改和变更都是显而易见的。本文中所用术语的选择，旨在最好地解释各实现的原理、实际应用或对市场中的技术的改进，或者使本技术领域的其他普通技术人员能理解本文公开的各个实现方式。

20

权利要求书

1. 一种模型性能评估的方法，包括：

在客户端节点处，将多个数据样本分别应用到预测模型，以得到
5 所述预测模型输出的多个预测得分，所述多个预测得分分别指示所述
多个数据样本属于第一类别或第二类别的预测概率；

基于所述多个数据样本的多个真值标签和所述多个预测得分，确定
与所述预测模型的预定性能指标相关的多个度量参数的值；

对所述多个度量参数的所述值施加扰动，得到所述多个度量参数
10 的扰动值；以及

将所述多个度量参数的扰动值发送给服务节点。

2. 根据权利要求 1 所述的方法，其中确定所述多个度量参数的
所述值包括：

确定所述多个真值标签中的第一类标签的第一数目，以作为第一
15 度量参数的值，所述第一类标签指示对应的数据样本属于所述第一类
别；以及

确定所述多个真值标签中的第二类标签的第二数目，以作为第二
度量参数的值，所述第二类标签指示对应的数据样本属于所述第二类
别。

20 3. 根据权利要求 2 所述的方法，其中对所述多个度量参数的所
述值施加扰动包括：

确定与所述第一度量参数和所述第二度量参数中的一个度量参
数的扰动相关的第一灵敏度值；

基于所述第一灵敏度值和差分隐私机制，来确定第一概率分布；

25 基于所述第一概率分布，对所述第一度量参数和所述第二度量参
数中的所述一个度量参数的所述值施加扰动，得到所述一个度量参数
的扰动值；以及

基于所述多个真值标签的总数目和所述第一度量参数和所述第
二度量参数中的所述度量参数的扰动值，确定所述第一度量参数和所

述第二度量参数中的另一度量参数的扰动值。

4. 根据权利要求 1 至 3 中任一项所述的方法，其中确定所述多个度量参数的所述值包括：

将所述多个预测得分发送给所述服务节点；

- 5 从所述服务节点接收所述多个预测得分各自在预测得分集合中的排序结果，所述预测得分集合包括由多个客户端节点发送的预测得分，所述多个客户端节点包括所述客户端节点；以及

基于所述多个预测得分各自的所述排序结果，确定所述第一类标签对应的数据样本的预测得分在所述预测得分集合中超过的预测得分的第三数目，以作为第三度量参数的值。

5. 根据权利要求 4 所述的方法，其中对所述多个度量参数的所述值施加扰动包括：

确定与所述第三度量参数的扰动相关的第二灵敏度值；

基于所述第二灵敏度值和差分隐私机制，来确定第二概率分布；

- 15 以及

基于所述第二概率分布，对所述第三度量参数的所述值施加扰动。

6. 根据权利要求 5 所述的方法，其中确定所述第二灵敏度值包括：

从所述服务节点接收与所述第二灵敏度值相关的信息；以及

- 20 基于所接收的信息来确定所述第二灵敏度值。

7. 根据权利要求 6 所述的方法，其中与所述第二灵敏度值相关的信息包括所述多个客户端节点的数据样本的总数目。

8. 根据权利要求 5 至 7 中任一项所述的方法，其中确定所述第二灵敏度值包括：

- 25 从所述多个预测得分各自的所述排序结果中确定最高排序结果；以及

基于所述最高排序结果来确定所述第二灵敏度值。

9. 根据权利要求 1 至 8 中任一项所述的方法，其中所述预定性能度量指标至少包括受试者工作特征曲线（ROC）的曲线下面积

(AUC)。

10. 一种模型性能评估的方法，包括：

在服务节点处，从多个客户端节点分别接收与预测模型的预定性能指标相关的多个度量参数的扰动值；

5 按度量参数聚合来自所述多个客户端节点的所述多个度量参数的所述扰动值，得到所述多个度量参数的聚合值；以及

基于所述多个度量参数的所述聚合值来确定所述预定性能指标的值。

11. 根据权利要求 10 所述的方法，其中对于所述多个客户端节点中的给定客户端节点，所述多个度量参数的扰动值指示以下至少一项：

所述给定客户端节点处的多个真值标签中的第一类标签的第一数目，所述第一类标签指示对应的数据样本属于所述第一类别；

15 所述多个真值标签中的第二类标签的第二数目，以作为第二度量参数的值，所述第二类标签指示对应的数据样本属于所述第二类别；以及

所述给定客户端节点处的所述第一类标签对应的数据样本的预测得分在预测得分集合中超过的预测得分的第三数目，所述预测得分由所述预测模型基于数据样本确定，并且所述预测得分集合包括由所述多个客户端节点发送的预测得分。

12. 根据权利要求 11 所述的方法，还包括：

向所述多个客户端节点分别发送与所述第二灵敏度值相关的信息。

13. 根据权利要求 12 所述的方法，其中与所述第二灵敏度值相关的信息包括所述多个客户端节点的数据样本的总数目。

14. 根据权利要求 10 至 13 中任一项所述的方法，其中所述预定性能度量指标至少包括受试者工作特征曲线 (ROC) 的曲线下面积 (AUC)。

15. 一种用于模型性能评估的装置，包括：

预测模块，被配置为将多个数据样本分别应用到预测模型，以得到所述预测模型输出的多个预测得分，所述多个预测得分分别指示所述多个数据样本属于第一类别或第二类别的预测概率；

度量确定模块，被配置为基于所述多个数据样本的多个真值标签
5 和所述多个预测得分，确定与所述预测模型的预定性能指标相关的多个度量参数的值；

扰动模块，被配置为对所述多个度量参数的所述值施加扰动，得到所述多个度量参数的扰动值；以及

发送模块，被配置为将所述多个度量参数的扰动值发送给服务节
10 点。

16. 一种用于模型性能评估的装置，包括：

接收模块，被配置为从多个客户端节点分别接收与预测模型的预定性能指标相关的多个度量参数的扰动值；

聚合模块，被配置为按度量参数聚合来自所述多个客户端节点的
15 所述多个度量参数的所述扰动值，得到所述多个度量参数的聚合值；
以及

性能确定模块，被配置为基于所述多个度量参数的所述聚合值来确定所述预定性能指标的值。

17. 一种电子设备，包括：

20 至少一个处理单元；以及

至少一个存储器，所述至少一个存储器被耦合到所述至少一个处理单元并且存储用于由所述至少一个处理单元执行的指令，所述指令在由所述至少一个处理单元执行时使所述设备执行根据权利要求 1 至 9 中任一项所述的方法。

25 18. 一种电子设备，包括：

至少一个处理单元；以及

至少一个存储器，所述至少一个存储器被耦合到所述至少一个处理单元并且存储用于由所述至少一个处理单元执行的指令，所述指令在由所述至少一个处理单元执行时使所述设备执行根据权利要求 10

至 14 中任一项所述的方法。

19. 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行以实现根据权利要求 1 至 9 中任一项所述的方法。

- 5 20. 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行以实现根据权利要求 10 至 14 中任一项所述的方法。