

**本科毕业设计说明书**

**基于差分隐私的回归分析器的研究与编程实现**

RESEARCH AND PROGRAMMING OF REGRESSION ANALYSIS

WITH DIFFERENT PRIVACY

|  |  |
| --- | --- |
| 学院（部）： | 计算机科学与工程学院 |
| 专业班级： | 信息安全14-2班 |
| 学生姓名： | 余方超 |
| 指导教师： | 方贤进教授 |

2018 年 5 月 31 日

基于差分隐私的回归分析器的研究与编程实现

摘要

伴随着互联网技术的不断发展，人类社会进入了数据爆炸式增长的时代。物联网传输系统，电子商务交易平台，社交网络平台等等，他们无时无刻不在产生大量的数据。而数据挖掘技术的不断改进和成熟，使得越来越多的人们认识到了这些海量的数据中所存在的巨大的价值，数据成为社会发展的驱动力。但是这些海量的数据同时包含了关于个人的隐私数据，如果应用不当，会造成大规模的隐私泄露。为了更加安全合理的开发这些数据资源，隐私保护技术成为人们研究的热点。而传统的隐私保护技术始终无法完全解决隐私保护的现实要求，差分隐私技术应运而生。

本文将主要介绍差分隐私技术在回归分析领域的运用。回归分析技术广泛运用在统计学和机器学习领域，而回归模型的发布机制中也存在着隐私泄露的风险。文中将深入讨论基于差分隐私的函数机制和直接添加噪声的机制，并将其运用在回归分析模型的构建当中，降低隐私泄露风险。同时还将总结和分析这两种实现机制的性能，衡量在不同隐私预算情况下两种机制的优缺点，实现隐私性和实用性间的平衡。

**关键词：隐私泄露，隐私保护，差分隐私，回归分析，函数机制**

RESEARCH AND PROGRAMMING OF REGRESSION ANALYSIS

WITH DIFFERENT PRIVACY

**ABSTRACT**

With the continuous development of Internet technology, human society has entered an era of explosive growth of data. Internet of Things transmission systems, e-commerce trading platforms, social networking platforms, and so on, they generate a lot of data at all times. The continuous improvement and maturity of data mining technology make more and more people realize the enormous value of these massive data. Data has become the driving force of social development. However, these massive data also contain personal privacy data. If it is not properly applied, it will cause a large-scale disclosure of privacy. In order to develop these data resources more safely and reasonably, privacy protection technology has become a hot topic of research. The traditional privacy protection technology has not been able to completely solve the practical requirements of privacy protection, differential privacy technology came into being.

This article will mainly introduce the application of differential privacy technology in the field of regression analysis. Regression analysis techniques are widely used in the field of statistics and machine learning, and there is also the risk of privacy leakage in the publishing mechanism of regression models. In this article, we will discuss in depth the functional mechanism based on differential privacy and the mechanism of directly adding noise, and apply it to the construction of regression analysis model to reduce the risk of privacy leakage. At the same time, it will also summarize and analyze the performance of these two implementation mechanisms, measure the advantages and disadvantages of the two mechanisms under different privacy budgets, and achieve a balance between privacy and practicality.

**Keywords: Privacy Leakage, Privacy Protection, Differential Privacy Regression Analysis, Function Mechanism**

目录

[摘要 I](#_Toc514769665)

[**ABSTRACT** II](#_Toc514769666)

[1 绪论 1](#_Toc514769667)

[1.1 引言 1](#_Toc514769668)

[1.2 研究背景 1](#_Toc514769669)

[1.3 传统的隐私保护技术 2](#_Toc514769670)

[1.3.1 基于限制发布的技术 2](#_Toc514769671)

[1.3.2 基于数据加密的技术 3](#_Toc514769672)

[1.3.3 基于数据失真的技术 4](#_Toc514769673)

[1.4 差分隐私技术 4](#_Toc514769674)

[1.5 本文的主要工作 5](#_Toc514769675)

[2 回归分析 7](#_Toc514769676)

[2.1 线性回归模型 8](#_Toc514769677)

[2.2梯度下降 9](#_Toc514769678)

[2.3批量梯度下降算法分析 9](#_Toc514769679)

[3 差分隐私 11](#_Toc514769680)

[3.1 背景 11](#_Toc514769681)

[3.2 差分隐私相关概念 11](#_Toc514769682)

[3.2.1 差分隐私定义 11](#_Toc514769683)

[3.2.2 敏感度 12](#_Toc514769684)

[3.3 噪声机制 12](#_Toc514769685)

[3.3.1 Laplace分布 12](#_Toc514769686)

[3.3.2 Laplace机制 13](#_Toc514769687)

[3.3.3 生成Laplace噪声 13](#_Toc514769688)

[4 基于差分隐私的线性回归 15](#_Toc514769689)

[4.1 直接添加噪声 16](#_Toc514769690)

[4.2 函数机制 16](#_Toc514769691)

[4.2.1 算法分析 16](#_Toc514769692)

[4.2.2 正则化 18](#_Toc514769693)

[5 代码实现与结果分析 19](#_Toc514769694)

[5.1 准备工作 19](#_Toc514769695)

[5.1.1 开发环境 19](#_Toc514769696)

[5.1.2 实验数据集 19](#_Toc514769697)

[5.1.3 性能度量标准 19](#_Toc514769698)

[5.2 直接添加噪声机制 20](#_Toc514769699)

[5.2.1 程序流程 20](#_Toc514769700)

[5.2.2 核心代码分析 20](#_Toc514769701)

[5.3 函数机制 22](#_Toc514769702)

[5.4 结果分析 23](#_Toc514769703)

[5.4.1 算法误差 23](#_Toc514769704)

[5.4.2 算法性能 25](#_Toc514769705)

[5.5 函数机制的隐私性分析 25](#_Toc514769706)

[总结 29](#_Toc514769707)

[参考文献 30](#_Toc514769708)

[致谢 31](#_Toc514769709)

1 绪论

* 1. 引言

隐私是一种与公共利益、群体利益无关，当事人不愿他人知道或他人不便知道的个人信息，当事人不愿他人干涉或他人不便干涉的个人私事，以及当事人不愿他人侵入或他人不便侵入的个人领域。隐私包含的含义很广泛，对于个人来说，最重要的隐私是个人的身份信息，利用该信息可以直接或间接追溯到某人；而对于集体而言，隐私是指代表一个集体行为的敏感信息。

随着网络技术的不断发展，隐私泄露问题频发，隐私保护问题越来越受到关注。在学术界，学者们曾提出了各种各样的隐私保护模型和算法，但是这些模型和算法大多都是建立在攻击者没有掌握最大背景知识这一前提下，在实际运用中会受到很大的限制。直到差分隐私技术的提出，为人们找到了解决隐私保护问题的新思路。

* 1. 研究背景

随着互联网技术的不断发展，网络虚拟社会日趋成熟，人们通过社交软件、购物平台以及各种各样的互联网应用方便了我们生活的同时，也在不经意间遗留下大量的个人信息，这些信息之间往往存在着累积性和关联性，大量的信息碎片拼凑起来很有可能暴露个人的隐私。而随着大数据时代的到来，数据挖掘技术广泛的应用在社会生活的方方面面，数据资源成为核心竞争力。企业或者某些机构出于各种不同的目的大规模的收集用户信息，造成了大量的隐私泄露事件。2006年，知名的网络影片提供商Netflix为了提高自己推荐系统的准确度，举办了一个机器学习的比赛。比赛中，Netflix公布了一个数据集，里面包含很多真实用户的视频播放和购买记录，为了保护用户的隐私，该公司删除了用户账号等可以唯一标识用户的信息。Netflix认为这样就能保证用户的隐私不被泄露。然而，事实上在这个数据集发布一个月后，两个学者利用链接攻击技术，挖掘出了这个数据集中某些记录对应的用户的真实身份信息。他们从网上获得了一些包含用户信息的电影浏览记录，然后把这些记录和Netflix数据集里的浏览记录进行匹配，就可以反向推理出在Netflix数据集中的个体对应的是哪些人，即攻击者通过其他方式获得了关于该数据集的背景知识，于是这些人在Netflix中的购买记录等隐私就泄露了。最近Facebook也深陷隐私泄露的丑闻当中，剑桥分析公司通过收购Facebook中一款性格测试的APP，从而掌握了近30万用户的信息，该公司又通过挖掘这些用户朋友圈的相关数据，最终导致5000万Facebook用户数据泄露。而这家公司的核心业务之一就是作为总统竞选的智囊团，他们通过获得的用户数据，分析用户政治倾向，为合适用户推送特朗普竞选总统的政策和广告，在一定程度上帮助特朗普赢得了大选。由此可以看出，大数据时代的隐私泄露问题不仅关系到个人的利益，而且很有可能对国家社会造成危害。

* 1. 传统的隐私保护技术

隐私保护技术与传统的隐秘技术有所不同，它的核心思想是保护隐私数据与个体之间的关系，换句话说就是隐私数据可以被任何人得到，但是却不能把它对应到某个特定的人身上。例如，某研究机构要开发一套疾病诊断系统，它需要医院提供大量的医疗数据，这些医疗数据涉及了大量的病患隐私信息。隐私保护技术需要做的就是对这些医疗数据进行适当的处理，使得研究机构能很好的利用这些数据的同时，无法根据这些信息还原出任何一个病患的身份信息。基于这一思想，很多隐私保护的机制相继被提出，这些机制大致可分为三类：基于限制发布的技术、基于数据加密的技术、基于数据失真的技术。

1.3.1 基于限制发布的技术

基于限制发布的技术的核心思想是选择性的发布原始数据、发布精度较低的敏感数据。当前这类技术的研究方向主要集中在“数据匿名化”，保证对敏感数据及隐私的披露风险在可容忍的范围内，即在隐私性和数据可用性间进行折中，在保证敏感数据对应个体的隐私性的基础上，选择性的发布可能泄露隐私的敏感数据的信息。基于限制发布的技术主要包括k-匿名、l-diversity、t-closeness，其中k-匿名技术是最早被广泛认同的隐私保护模型。

k-匿名算法的思想是：要求发布的数据中每一个个体的记录都要与其他至少k-1个个体的记录不可区分，也就是公开数据中的任意准标识信息，相同的记录出现至少k次。例如表1.1为用户的信息统计表，表格中的一条记录对应唯一一个用户。

表1.1 原始数据

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 姓名 | 性别 | 年龄 | 邮编 | 购买偏好 |
| 李明 | 男 | 24 | 100083 | 电子产品 |
| 赵磊 | 男 | 23 | 100084 | 家用电器 |
| 小红 | 女 | 26 | 100102 | 护肤品 |
| 小紫 | 女 | 27 | 100104 | 厨具 |
| 李雷 | 男 | 36 | 102208 | 图书 |
| 刘磊 | 男 | 36 | 102201 | 图书 |

将表格中公开的属性分为三类：

(1)标识符：姓名、电话、地址等能唯一标识某一个体的属性，，这些数据将在发布前删除。

(2)准标识符：通过结合多个属性能唯一标识一个个体，如邮编、姓名、性别等等。

(3)敏感属性：包含隐私数据的属性，如收入、购买偏好等等。这些数据是直接公开的。

举个例子，现在对上面表格中的数据进行2-匿名保护得到表1.2。如果攻击者想确定李明的购买偏好，通过查询他的年龄、邮编、性别，攻击者会发现数据中至少有两个人有着相同的年龄、性别和邮编。这样攻击者就无法区分这两条记录中到底哪一个是李明了，从而保证了数据的隐私性。

表1.2 2-匿名

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 姓名 | 性别 | 年龄 | 邮编 | 购买偏好 |
| \* | 男 | （20，30） | 10008\* | 电子产品 |
| \* | 男 | （20，30） | 10008\* | 家用电器 |
| \* | 女 | （20，30） | 10010\* | 护肤品 |
| \* | 女 | （20，30） | 10010\* | 厨具 |
| \* | 男 | （30，40） | 10220\* | 图书 |
| \* | 男 | （30，40） | 10220\* | 图书 |

但是上述理论是基于攻击者没有掌握任何用户的背景知识的基础上，如果攻击者现在已经获取了赵磊的所有信息以及李明的邮编信息，则攻击者能很快确认表中的第一条记录即为李明的用户信息，从而确定他的购买偏好。基于k-匿名存在的种种缺陷，人们又提出l-diversity、t-closeness等算法，但这些算法依然不能完全解决背景知识攻击的问题，新的攻击方法相继出现。从根本上说，由于对攻击者所掌握的背景知识和攻击方式给出了过多的假设，而这些假设在实际生活中往往不完全成立，这是传统的隐私保护模型最大的缺陷。

1.3.2 基于数据加密的技术

建立在密码体制的基础上的隐私保护方法中实现了数据接受方对原始数据的不可见性，同时加密算法不会对数据造成损失，保证了数据的机密性、可用性以及数据的隐私性。同态加密、安全多方计算是使用比较广泛的加密技术。

同态加密是对密文进行特定的代数运算得到仍然是加密的结果的一种加密形式，将加密结果进行解密所得到的结果与对明文进行相同运算的结果一样。也就是说，同态加密技术可以令人们在加密的数据中进行数据的查询等操作，并且可以得出正确的结果，而在整个处理过程中不用对数据进行解密。

安全多方计算利用机密机制形成交互计算的协议，参与计算的各实体均以私有数据参与计算，计算结束后，各实体只能得到正确的结果，而不能得到他人的隐私数据。

不论是同态加密技术还是安全多方计算他们的优点在于保证了数据的隐私性和准确性，缺点在于他们都是基于公钥密码体制，算法复杂度高，计算效率低，且消耗资源较多。

1.3.3 基于数据失真的技术

基于数据失真的技术通过对数据添加噪声，使敏感数据产生一定范围内的失真，但同时保持某些数据或数据属性统计方面的性质，其中随机化技术使比较常见的技术。

随机化方法是一种为集中式数据进行隐私保护数据挖掘的重要方法。该方法的基本思想是通过对数据添加噪声，使得原始数据集的分布概率能够保留下来，而每条记录信息很难恢复，以此达到隐私保护的目的。

算法简单是随机化方法最大的优点，但是在另一方面也造成了随机化方法的一些固有缺陷。针对不同的应用，随机化方法都需要设计特定的算法进行处理，这大大限制了随机化方法的实用性。

* 1. 差分隐私技术

为了彻底解决传统隐私保护技术存在的问题，2006年C.Dwork等人提出了一个新的隐私保护模型，即差分隐私。差分隐私的核心思想也是属于数据失真技术的范畴。通俗的说，差分隐私的基本思路是：如果我现在需要公布一个数据集，给研究机构做数据分析，但我希望数据集中每一个个体的信息都不会被泄露出去，那么我可以通过添加噪声的方法对这个数据集进行扰动。当然添加的扰动必须在可控的范围内，否则数据就丧失了可用性。然而如果添加的扰动太小，也不能起到保护隐私的作用。C.Dwork在他的论文中提出了一个数学上的描述，来测量一个扰动机制究竟能够带来多大程度上的保密性。简单来说就是，你给我一个数据集A，如果我有一个扰动机制，可以让我先对A做一个扰动得到，再从原数据集A里随意拿掉一行记录得到B，对这个数据集B做扰动得到，如果得到的和几乎是一模一样的（对同一随机算法这两个数据集的输出概率分布几乎相同），那么我就认为这个扰动机制可以保护隐私。因为在这个扰动机制下，A里面任何单独一行数据存在或不存在都几乎不影响结果。

更简单的说，就是一个个体在或者不在这个数据集中不会对查询结果产生影响，即攻击者通过对该数据集的任何查询或者背景知识都无法准确推断出我是否在这个数据集中。这种隐私保护定义足够强大。因为既然你在不在这个数据集中都不会影响最终的查询结果，那么我们可以认为你就不在这个数据集中，而如果你都不在这个数据集中，你的数据自然不会泄露。

差分隐私技术是建立在严格的数学推理证明的基础之上，它的提出引起了国内外学术界的广泛关注。差分隐私技术的研究目前主要分为两个方面，一方面是对于差分隐私理论的不断完善，另一方面是如何将差分隐私技术运用到实际，真正解决实际生活中的隐私泄露问题。

理论性研究主要讨论的是两类问题。一类是关于差分隐私的数据发布机制，另一类是关于数据挖掘机制。基于差分隐私的数据发布机制分为交互式发布和直接发布。交互式发布有点像数据库的中间件，就是针对用户每一次的查询，数据库算出结果之后，该结果进行差分隐私算法处理之后再呈现给用户。用户无法拿到原始的数据，而只能通过对数据库的查询获得加噪音处理后的结果。直接发布就是把整个数据集一次性的进行差分隐私处理后完整公开，用户可以拿到这份数据，但是单一的每条记录可能都是修改过的，只是在统计学意义上，依然保留原始数据的特征。依然可以进行数据挖掘、分类等。而对数据挖掘机制的研究主要是对传统的机器学习算法进行修改使其符合差分隐私的定义。

应用型研究主要讨论的问题比较广泛。差分隐私应用于推荐系统，由于推荐系统需要收集大量用户数据进行推荐算法的训练，如何既能保证收集到的用户数据的隐私性，又能使数据训练出来的推荐算法的性能更好；差分隐私解决基于位置服务（LBS）的隐私保护问题，现在很多手机有定位功能，位置信息属于个人隐私信息，而有很多APP是基于位置的服务的（外卖、地图等），如何保护个人位置隐私数据的同时，商家又能够提供相应的服务，传统基于位置的隐私保护多用的是近似算法或一些模糊算法，能否和如何将差分隐私引入该领域去解决这一问题，也是目前研究的一个热点；差分隐私解决收集用户信息的问题，在移动互联网时代，各种各样的APP以及浏览器蓬勃发展，但是这些事物的出现都在加剧隐私泄露的风险，APP会收集用户的数据，浏览器也会记录用户的访问记录，差分隐私技术也许会帮助人们解决这一问题。

* 1. 本文的主要工作

本文首先介绍隐私保护相关的概念和意义从而引出差分隐私技术，然后分别介绍了回归分析技术的相关概念和实现算法以及差分隐私技术相关的概念和实现机制，最后将差分隐私应用到回归模型的构建当中，并对结果进行分析。

主要章节安排：

第一章主要介绍差分隐私保护技术的研究意义和背景

第二章主要介绍回归分析技术的相关概念和算法

第三章主要介绍差分隐私技术的相关概念和算法

第四章主要介绍基于差分隐私的回归分析技术实现的机制

第五章主要内容为通过编程实现基于差分隐私的回归分析技术的两种机制并进行比较。

第六章主要内容是对本文进行总结

2 回归分析

回归分析方法是在众多的相关变量中，根据实际问题考察其中一个或多个变量（因变量）与其余变量（自变量）的依赖关系，建立变量之间的统计模型，从而对变量（自变量）未来的趋势进行预测。

例如，现在有表2.1中的房价数据，那么该如何利用这些数据来对未来的房价进行预测呢。

表2.1 房价数据

|  |  |
| --- | --- |
| 面积（） | 房价（万元） |
| 487 | 100 |
| 667 | 150 |
| 1037 | 299 |
| 1459 | 352 |
| … | … |

现在根据这些数据，在坐标轴上将它们表示出来,横坐标表示面积，纵坐标表示房价，如图2-1所示。

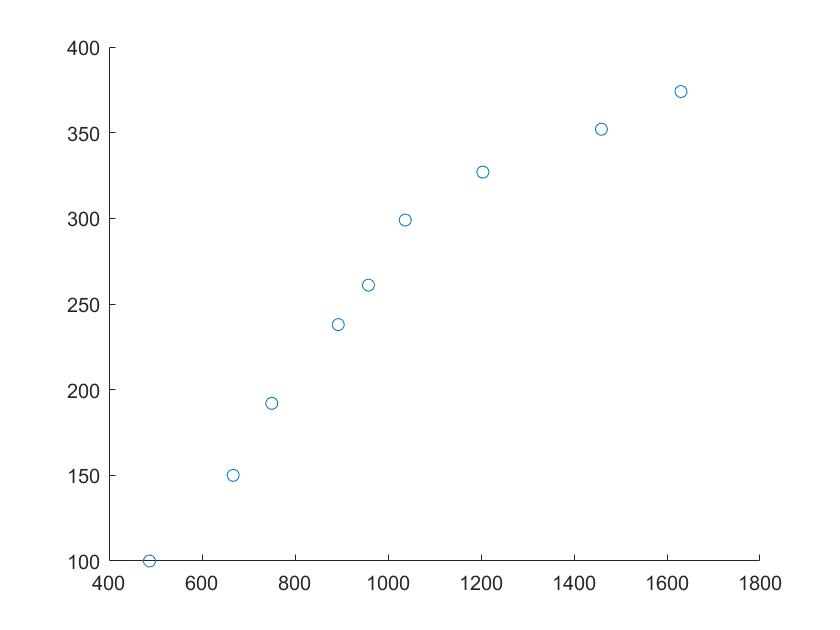
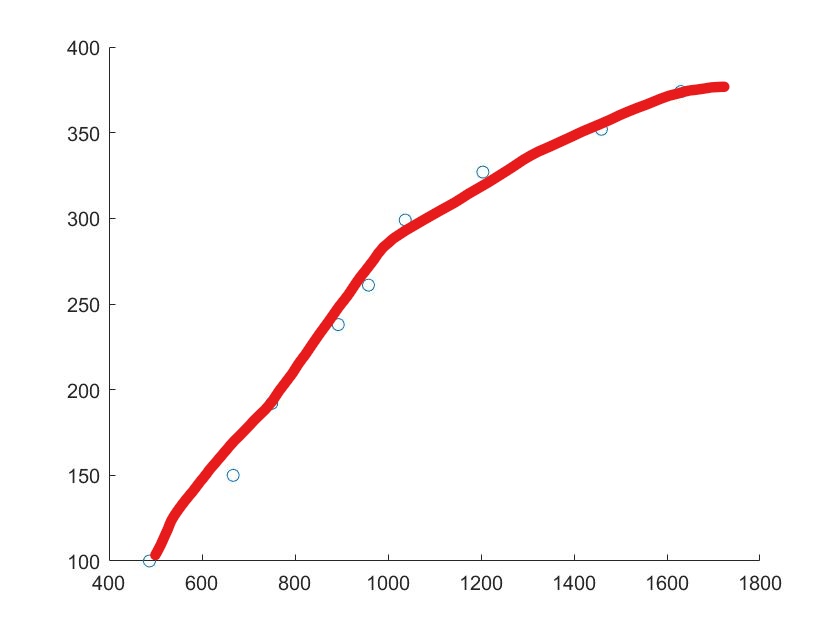


图2-1 房价数据

如果现在要预测面积为750的房子的房价是多少，该如何实现呢？利用这些已知的数据，可以大致的描述面积与房价之间的关系。如图2-2所示，将坐标轴上的点用一条平滑的曲线连接起来，即回归曲线，通过建立面积与房价间回归模型，然后对模型进行不断的优化，减少误差，即可对特定面积的房价进行预测。这就是常见的回归分析。

图2-2 回归模型

回归分析应用十分广泛。在统计学和经济学领域，人们通过分析大量数据，建立回归模型来预测股市走势、预测企业收益等等；在机器学习领域，人们通过学习所得的回归模型，实现分类和预测功能。

回归分析按照变量的多少，分为一元回归和多元回归分析；按照自变量和因变量之间的关系类型，可分为线性回归分析和非线性回归分析。本文主要讨论多变量情况下的线性回归模型。

2.1 线性回归模型

给定由个属性描述的示例其中均是在第个属性上的取值，线性模型试图获得一个通过属性的线性组合来进行预测的函数，即

一般用向量形式写成

（注：增加属性）

其中。获得之后，模型就得以确定。

以某地区的房价问题来构建线性回归模型，假设，某地区的房价由面积和所处地段两个属性所决定。

给定数据集,。现在希望通过线性回归分析获得一个模型去预测某一特定面积和所处地段的房屋的价格，该如何获得这一模型呢，也就是说我们该如何来确定呢？要想确定，关键在于如何衡量和之间误差（建模误差）。要使获得的模型越精确，就要求建模误差越小，即。在回归分析中，我们的目标就是便是选择出使得建模误差的平方和能够最小的模型参数，即使得代价函数

最小。

对于解决的最小值问题的方法有很多，例如最小二乘法和梯度下降法，本文将采用梯度下降法解决这一问题。

2.2梯度下降

梯度下降法（Gradient descent）是一个一阶最优化算法，通常也称为最速下降法。梯度下降思想是：建模之前随机选择一个参数组合，例如：，计算代价函数，然后沿着梯度下降的方向寻找下一个让代价函数值下降最快的参数模型。重复上述步骤，直到获得一个局部最小值。选择不同的初始参数组合，可能会得到不同的局部最小值。

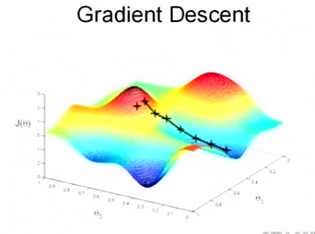
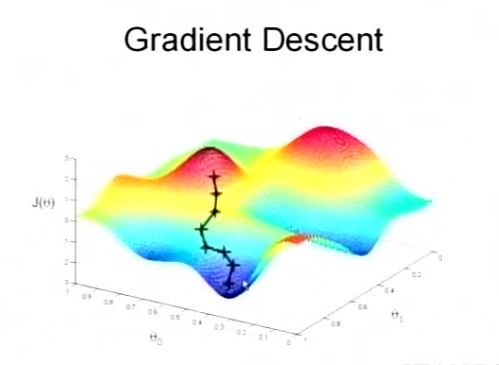


图2-3 梯度下降

由于现实生活中存在着各种各样的数据集，对于不同的数据集我们会采用不同的梯度下降法。当分析的数据集很大时，我们一般采用随机梯度下降，可以提高数据运算的效率；当分析的数据集规模较小时，我们采用批量梯度下降，可以提高拟合模型的精度。本文将采用批量梯度下降法进行回归分析。

2.3批量梯度下降算法分析

在进行批量梯度算法分析前，有表2.2定义：

表2.2 数据定义

|  |  |
| --- | --- |
| 基本定义 | |
| m | 数据集中的样本数量 |
| n | 属性的维数 |
| x | 自变量 |
| y | 因变量 |
|  | 学习率，代价函数每次下降的幅度 |

批量梯度下降的核心算法如下所示：

进行批量梯度下降的关键在于对代价函数进行求导，即

经推导可得

在批量梯度算法中，通过不断的迭代，直到

从而得到代价函数的局部最优解，同时可以不断调节起始点的位置和的值，使获得的局部最优解接近全局最优解。

3 差分隐私

3.1 背景

虽然隐私保护技术已有数十年的历史，但是由于传统隐私保护机制存在的各种问题，隐私保护技术一直没有得到很好的应用。直到差分隐私技术的出现，这一问题才得到了很好的解决。而一些传统的隐私保护的思想也在差分隐私中得到体现。

差分隐私保证攻击者能获取的个人数据几乎和他们从没有这个人记录的数据集中能获取的相差无几。差分隐私的定义足够强大，因为不论自己在不在数据集中，该数据集的分析者都将得出关于该个人的相同的结论。由于其敏感的个人信息与系统的输出几乎完全不相关，因此用户可以确信处理其数据的组织不会侵犯他们的隐私。分析者几乎"无法获得个人信息"意味着他们被限制在关于任何个人的看法的微小变化中。

3.2 差分隐私相关概念

3.2.1 差分隐私定义

定义1-差分隐私）：给定临近数据集D1和D2,即D1、D2至多相差一条数据，若数据集在隐私算法K中的任意输出都满足，则算法K满足-差分隐私，

其中D1、D2作为算法K的输入，而K(D1)、K(D2)分别为D1、D2的输出；参数为隐私保护预算，由算法设计者指定，在实际应用中一般取很小的值，例如0.01，0。1或者1等，的值越小表示差分隐私保护级别越高；表示结果为S的概率，同时也是隐私被泄露的风险。

例如，如表3.1所示，D一组医疗数据集，其中的每一条记录表示个体是否患有近视，0表示视力正常，1表示患有近视。

表3.1 医疗数据集

|  |  |
| --- | --- |
| 姓名 诊断结果 | |
| 李明 | 0 |
| 赵雷 | 1 |
| 小红 | 1 |
| 刘磊 | 0 |
| 王阳 | 1 |

数据集为用户提供查询服务，那么如何实现在保证患者隐私不被泄露的前提下，提供较为准确的查询结果。假设，用户输入参数为i，查询函数f(i)=count(i)的作用是获取数据集中前i行满足“诊断结果”为1的记录数量，并将结果反馈给用户。如果攻击者想推测王阳是否患有近视，并且攻击者已经知道王阳的记录在在数据集中的位置（此数据集中王阳的位置为i=5），它可以通过count(5)-count(4)推断出正确结果。

但如果f是一个满足差分隐私保护的查询函数，假设f(i)=count(i)+noise，其中noise服从某种随机分布的噪声。那么对于只相差一条记录的数据集来说，查询函数f的输出结果可能相同，那么攻击者就无法通过这种方式获得想要的结果，这是差分隐私技术应用的最简单的方式。

3.2.2 敏感度

从上面例子我们可以看出，对于差分隐私技术而言，最关键的环节在于如何添加噪声。如果添加的噪声过小，攻击者可以通过多次查询，分析噪声分布规律，从而达到攻击目的；而如果添加的噪声过大，使得查询的结果严重偏离实际，失去了可用性。这里提出一个敏感度的概念，它指的是删除数据集中任一条记录对查询结果造成的最大改变，它是衡量噪声大小的关键 。

定义2（敏感度）：设有函数f：D，输入为一数据集，输出为d维实数向量，对于任意临近数据集D1、D2，

称为函数f的敏感度。其中，是f(D1)和f(D2)之间的一阶范数距离。

函数的敏感度是函数自身的特性，独立于数据集，不同的函数有不同的敏感度，例如在上面例子中提到的计数函数count(i)，它的敏感度。

3.3 噪声机制

为了设计满足差分隐私的算法，对不同的需求将采用不同的机制。一般对于运算结果为数值型数据的算法，采用Laplace机制；对于运算结果为非数值型结果的算法，采用指数机制。本文将主要讨论Laplace机制。

3.3.1 Laplace分布

在概率论和统计学中，拉普拉斯是一种连续概率分布，也叫做双指数分布。如果随机变量的概率密度函数满足：

称该随机变量满足Laplace分布。如图3-1所示是Laplace分布的概率密度函数曲线。

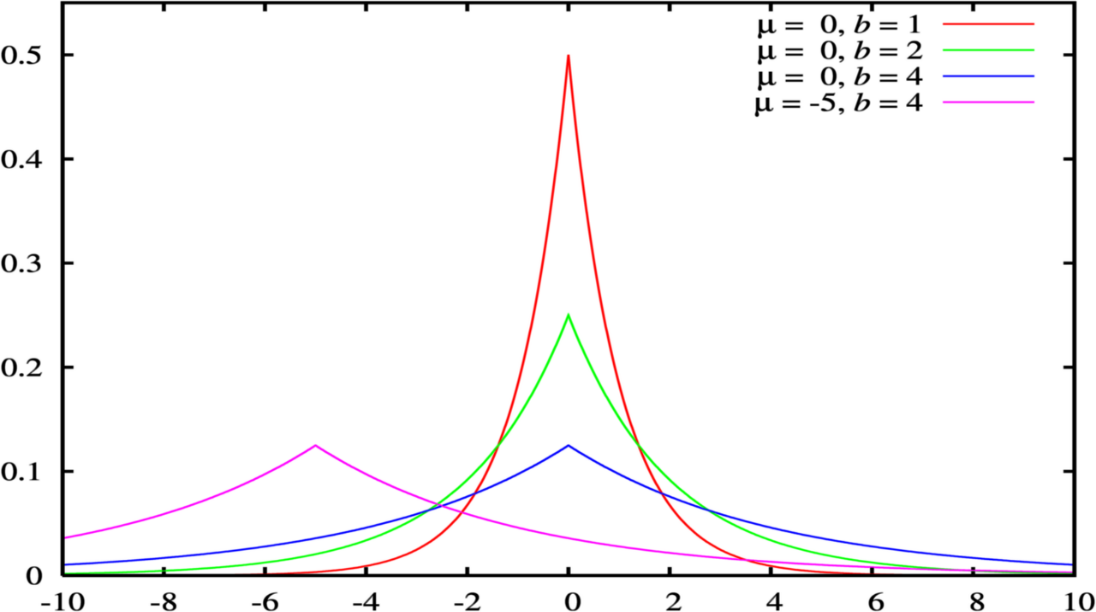


图3-1 Laplace概率密度函数

由概率密度函数推导出拉普拉斯分布的概率累计函数为：

期望和方差对于概率分布来说是两个最基本的重要属性，对于拉普拉斯分布来说，它的期望和方差分别为和。

3.3.2 Laplace机制

Laplace机制通过向查询结果添加服从Laplace分布的随机噪声来实现差分隐私。令=0，尺度参数为的Laplace分布为Lap(b),即概率密度函数为

定义3 (Laplace机制)：给定数据集D，设有函数：，其敏感度为，那么随机算法满足-差分隐私保护，其中YLap(/）为随机噪声，服从尺度参数为/的laplace分布

由定义可知，对于Laplace机制来说，噪声的大小由敏感度和隐私预算共同决定。而敏感度是由查询函数自身的特性所确定，一旦查询函数确定下来，敏感度也就确定下来，所以此时添加噪声的大小由隐私预算所决定。隐私预算越大，所添加的噪声越小，隐私预算越小，所添加的噪声越大。

3.3.3 生成Laplace噪声

现在已经确定了随机算法K(D)以及噪声的生成机制，那已知噪声满足Laplace分布，那该如何来生成Laplace噪声呢。

生成Laplace噪声之前，先生成满足Uni(-0.5,0.5)的均匀分布的随机值。假设Uni(-0.5,0.5)，，则满足Uni(0,1)的均匀分布，即。由于Laplace概率密度函数的值域区间为[0,1],满足Laplace分布的所要求的值域区间。然后通过求解概率密度函数的反函数即可求得满足Laplace分布的噪声：

x即为我们所需的噪声值。

4 基于差分隐私的线性回归

本文在开头介绍了回归分析的各种应用场景，通过建立回归模型可以很好的解决分类和预测问题。假如，现在有一家数据分析机构，它通过收集大量的医疗统计数据，建立了一个能够预测癌症患病风险的回归模型，该机构获得的数据和回归模型如下表所示。其中收入水平指数表示个体的收入水平，收入水平越高，指数越小；生活习惯指数用来描述个体的生活习惯是否良好，生活习惯指数越高，生活习惯越不健康；环境质量指数用来描述环境质量的好坏，环境质量指数越高，环境质量越好；家族病史指数用来统计家族患病数据，家族病史指数越大，患病风险越大。通过这些数据来预测患病风险，患病风险数据越大，患病的几率就增加。一般认为患病风险超过0.5，即认为患有癌症。

表4.1 统计数据

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 收入水平指数 | 生活习惯指数 | 环境质量指数 | 家族病史指数 | 患病风险 |
| 0.56 | 0.41 | 0.45 | 0.32 | 0.431 |
| 0.37 | 0.59 | 0.50 | 0.62 | 0.547 |
| 0.48 | 0.33 | 0.61 | 0.40 | 0.447 |
| 0.63 | 0.47 | 0.47 | 0.54 | 0.520 |

通过线性回归分析，得到向量（0.1，0.2，0.3，0.2，0.3）。该机构将未经任何处理的回归模型发布。

现在有一个攻击者，他拥有与该数据分析机构类似的医疗统计数据，但其中关于个体患病风险这一敏感数据并不完整。攻击者想获取这些医疗统计数据对应个体的患病风险数据，来推测他们是否患有癌症，以实现一些不法的目的，他通过这家数据结构所提供的回归模型就可以实现这一目的。

假设攻击者手中有这样一条医疗数据（0.61，0.60，0.48，0.63），攻击者希望获得这条医疗数据所对应个体是否患有癌症，他只需要将这些数据代入该机构所发布的回归模型，即

0.1+0.2\*0.61+0.3\*0.60+0.2\*0.48+0.3\*0.63=0.687

由计算结果可知该个体患有癌症，该患者的隐私信息遭到泄露。

通过上面这个例子可以看出，发布一个未经任何处理的回归模型存在隐私泄露的风险。那能否将差分隐私的思想应用到回归分析的过程当中，以实现隐私保护？答案是肯定的，可以通过在回归分析的过程中注入一定的噪声，在保证回归模型准确性的同时实现隐私保护。下面本文将讨论如何来实现这一机制。

4.1 直接添加噪声

在前面分析差分隐私定义的时候，举了一个计数函数的例子，在例子中为了防止攻击者获得数据集的隐私信息，在查询结果后面直接添加噪声。同样，为了实现对线性回归分析的隐私保护，同样可以在线性回归模型的后面直接添加Laplace噪声：

K(D)满足ε-差分隐私。

这种机制对于处理一般的简单问题可以起到作用，但是对于复杂问题来说，这种机制存在明显的缺陷。我们知道我们所添加的噪声量的大小在一定程度上取决于敏感度，敏感度越大，产生的噪声越大，那我们所得出的模型的准确性也就越低。对于一些复杂的线性回归问题，可能会存在敏感度过大的问题，所以此时如果采用这种机制，尽管保证了隐私性，但同时使得出的结果失去了可用性。

4.2 函数机制

直接添加噪声的机制无法满足人们的要求，应该试图找出一种机制，实现计算结果的可用性与数据隐私性之间的平衡。函数机制的出现很好的解决了这一问题。

函数机制[6]其实是Laplace机制的一种延伸，它不是简单将噪声添加到回归模型的后面，而是通过扰动代价函数得到加噪后的代价函数，再通过计算出，此时再将得到的回归模型发布。通过这种机制得到的参数模型满足ε-差分隐私，攻击者无法通过回归模型获得敏感信息。

4.2.1 算法分析

在进行算法分析之前，我们有表4.1参数定义

表4.2 参数定义

|  |  |
| --- | --- |
| 定义 | |
| D | 数据集 |
|  | 数据集中的第i条记录 |
| d | 自变量的维度 |
|  | 回归模型的参数向量 |
|  | 第i条记录回归值与实际值之差 |
|  | ,即代价函数 |
|  |  |
|  | 对添加噪声之后的代价函数 |
|  |  |
|  | 向量中一项或多项的乘积，如 |
|  | 表示多项式集合，j指示具体的项中的系数 |

根据这些定义，我们将对之前定义的一些公式进行改写：

那对于线性回归分析来说

我们将拆分成了一个多项式，函数机制的核心思想就是对含有项的系数添加噪声，获得一个新的目标函数。在对新的目标函数进行优化得到最终结果。具体算法如表4.2所示

表4.3 函数机制

|  |  |
| --- | --- |
| 数据集为D，目标函数为，隐私预算为 | |
| 1： | Set |
| 2： | For each 0 do |
| 3: | For each do |
| 4: | Set |
| 5: | End for |
| 6: | End for |
| 7: | Let = |
| 8: | Compute |
| 9: | Return |

在进行该算法之前，要首先对数据进行预处理，要求处理后的数据集满足

，

经过计算，对于该线性回归模型来说，其敏感度。

4.2.2 正则化

对于上述算法，可以简单的将代价函数看作一个一元二次方程，即,算法目的是求使得代价函数为最小值的的值。在一元二次函数中要保证函数取得最小值，首要的前提是保证二次项系数大于0，即a>0。那么在该算法中同样要考虑这一点，代价函数必须要存在最小值。为了推导过程更简洁，将代价函数表示为矩阵的形式：

的系数矩阵添加噪声后即得到，这里得就相当于一元二次方程的二次项系数。为了使加噪后的代价函数能够取得最小值，必须保证为对称的正定矩阵。因为为对称矩阵，所以在对添加噪声时，可以对称的添加噪声，使得获得的也为对称矩阵。那如何保证为正定矩阵呢？这里要用到正则化的思想，令：

I是一个的单位矩阵，是正则化参数，由用户指定，值的大小决定了拟合的效果。

通过这种方式，在后面迭代的过程中，会逐渐接近正定矩阵，但是这还不能完全保证为正定矩阵。为了达到目的，必须进行下一步操作。将进行特征分解，得到,为特征值的对角矩阵，为的特征向量矩阵，且为单位矩阵。所以，我们可以得到：

删除对角矩阵中特征值为负数所对应的行和列，假设中含有k个负数的特征值，则最终变成了的对角矩阵，记为。同时删除Q中所对应的特征向量，使得Q变成了的矩阵，记为。则代价函数表示为：

此时得到的必定是一个对称正定矩阵。

5 代码实现与结果分析

这一节的主要工作是通过程序来构建基于差分隐私的线性回归模型，并对该模型进行评估。同时，这里将采用上述所提到的两种基于差分隐私的线性回归模型的机制进行回归模型的构建，通过对比来分析两种机制的优缺点。

5.1 准备工作

准备工作主要分为三个部分，搭建开发环境、实验数据集预处理以及误差衡量标准。

5.1.1 开发环境

构建模型所采用的开发语言为MATLAB。MATLAB是一种高级技术计算语言和交互式环境,广泛运用在用于算法开发、数据可视化以及数值分析等领域。使用MATLAB可以很方便的实现矩阵运算和绘制图像，在机器学习领域运用广泛。搭建开发环境时，直接在win10系统中下载安装即可使用。

5.1.2 实验数据集

本文采用的数据集为美国人口普查数据，数据集包含188846条数据。数据集包含13个属性，即年龄、性别、教育程度、是否残疾、婚姻状况、每周工作时间、住宅所有权、出生地、居住当地年数、子女个数、家庭成员、汽车数量和每年收入。由于婚姻状况超过两个值，将婚姻状况转变为单身/已婚两个属性，转变后数据集变为14维。我们将每年收入这一属性作为回归分析的预测值，其余属性用来构建模型预测每年收入。

对于这些原始的数据集，各个属性的取值的尺度参差不一，这样的数据会使得在进行梯度下降算法时难以收敛，所以本文将采用特征缩放的方法对原始数据集进行处理，将原始数据的每个属性的取值都限制在[-1,1]区间内。

在实验过程中，将该数据集分为两个部分：训练数据集和测试数据集。训练数据集中数据的记录条数大约占数据集总数的80%，用于构建回归模型；测试数据集中数据的记录条数大约占数据集总数的20%，用于对数据集进行评估。

5.1.3 性能度量标准

本文针对差分隐私的性能衡度量的标准由两个方面组成：算法误差、算法性能。

算法误差主要衡量的是生成的回归模型在测试数据集上与真实值之间的方差；算法的性能主要衡量的是算法运行所需要的时间。

5.2 直接添加噪声机制

这一节，将对直接添加噪声的差分隐私机制的具体实现进行叙述，包括程序执行流程和核心代码分析

5.2.1 程序流程

该机制的具体实现流程如图5-1所示

数据集

数据预处理

测试数据集

梯度下降优化

生成属性参数

计算敏感度

生成噪声

预测值添加噪声

计算误差

循环取平均值

结束

开始

训练数据集

图5-1 程序流程图

5.2.2 核心代码分析

（1）数据预处理：前面提到在使用该数据集时需要对数据集进行相应的处理，使得在进行梯度下降的过程中更容易收敛。这里采用的处理方式为特征缩放。

其中表示属性值，表示数据集中该属性的最小值,表示数据集中该属性的最大值与最小值的差值。核心代码如下所示（矩阵表示）：

RawData = load('Data2\_Linear.dat');

%以矩阵形式导入数据

[DataRow, DataCol] = size (RawData);

%获取矩阵维度

RawData\_min = min(RawData,[],1);

RawData\_max = max(RawData,[],1);

%获得矩阵中每一维数据中的最大值和最小值

Data=(RawDataones(DataRow,1)\*RawData\_min)./(ones(DataRow,1)\*(RawData\_max-RawData\_min));

Data = (Data-0.5).\*2;

特征缩放的结果使得所有的数据的取值分布在[-1,1]区间内

（2）梯度下降优化：在这里直接调用了MATLAB内置的fminunc函数来求解最优值，他的核心思想与梯度下降类似，并在此基础上进行了优化，核心代码如下所示：

options = optimset('LargeScale','off');

%LargeScale表示进行大规模搜索，寻找最优解

options.MaxFunEvals = 100000;

%设置迭代次数

[opti\_theta,~,~,~]= fminunc(@object, theta, options);

%对代价函数进行迭代，寻找最优解

(3) 生成拉普拉斯噪声：在生成噪声之前，必须先计算敏感度，数据的敏感度计算由Calculate\_sense函数来实现。然后我们生成拉普拉斯噪声，生成噪声由函数Create\_noise函数实现，核心思想是逆累计分布函数，核心代码如下所示：

b=sense/privacy;

a=rand(1,1000000)-0.5;

s=a(randperm(row));

%生成随机化均匀分布

noise=mu-b\*sign(s).\*log(1-2\*abs(s));

运行这段代码生成的噪声如图5-2所示：

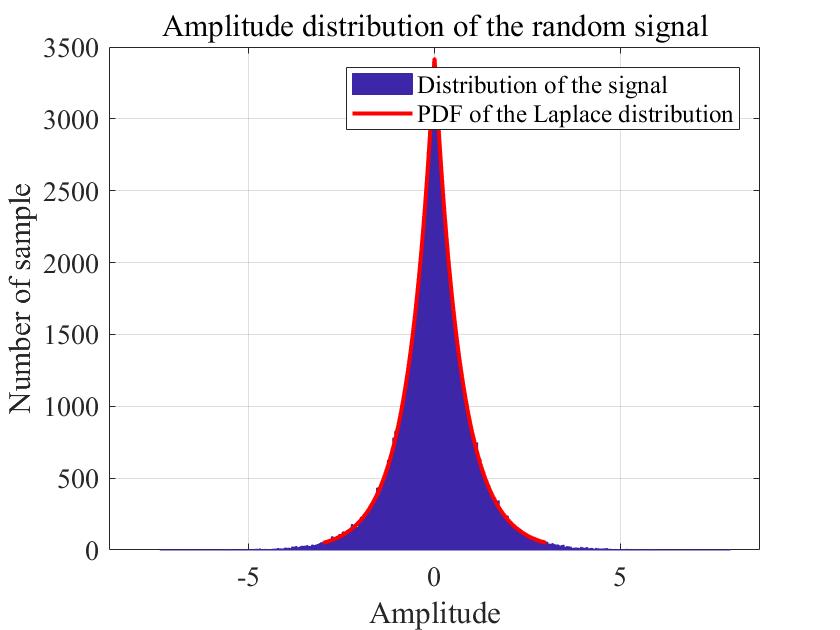


图5-2 Laplace噪声

5.3 函数机制

采用函数机制具体的程序流程与直接添加噪声机制相似,在实现函数机制的过程中需要注意的是添加噪声和正则化的问题。

（1）添加噪声：函数机制中，需要对代价函数的系数添加噪声，添加噪声的具体代码如下所示：

Coe2 = R0+laprnd(d,d,0,Sensitivity\*(1/ep));

Coe2 = 0.5\*(Coe2'+Coe2);

Coe2 = Coe2+5\*sqrt(2)\*Sensitivity\*(1/ep)\*eye(d);

Coe1 = R1+laprnd(d,1,0,Sensitivity\*(1/ep));

(2) 正则化：正则化的目的就是使得加噪后的代价函数能够取得最小值，核心代码如下所示：

[vec,val]=eig(Coe2);

del = diag(val)<1e-8;

val(del,:) = [];

val(:,del) = [];

vec(:,del) = [];

5.4 结果分析

本文在分析过程中选择6组隐私预算数据进行测试，分别是0.1、0.2、0.4、0.8、1.6、3.2。

5.4.1 算法误差

首先来讨论一下没有经过差分隐私算法生成的回归模型，经过计算所得到的参数模型如下表所示：

表5.1 未使用差分隐私算法

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| -0.3778 | 0.0037 | 0.3152 | -0.1163 | 0.0373 | 0.1756 | -0.0417 |
|  |  |  |  |  |  |  |
| 0.0286 | 0.0151 | 0.0575 | 0.1693 | 0.0968 | 0.0152 | -0.0297 |

直接添加噪声机制得到的回归模型为：

表5.2 直接添加噪声回归模型

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | -0.3819 | -0.3817 | -0.3857 | -0.3804 | -0.3775 | -0.3821 |
|  | 0.0039 | 0.0036 | 0.0036 | 0.0041 | 0.0038 | 0.0040 |
|  | 0.3167 | 0.3141 | 0.3161 | 0.3142 | 0.3133 | 0.3135 |
|  | -0.1195 | -0.1185 | -0.1219 | -0.1165 | -0.1178 | -0.1180 |
|  | 0.0383 | 0.0375 | 0.0383 | 0.0368 | 0.0391 | 0.0384 |
|  | 0.1745 | 0.1772 | 0.1757 | 0.1766 | 0.1742 | 0.1779 |
|  | -0.0426 | -0.0415 | -0.0421 | -0.0415 | -0.0414 | -0.0419 |
|  | 0.0264 | 0.0269 | 0.0269 | 0.0265 | 0.0264 | 0.0266 |
|  | 0.0141 | 0.0145 | 0.0145 | 0.0153 | 0.0150 | 0.0146 |
|  | 0.0561 | 0.0549 | 0.0528 | 0.0559 | 0.0594 | 0.0540 |
|  | 0.1686 | 0.1697 | 0.1686 | 0.1702 | 0.1693 | 0.1705 |
|  | 0.0924 | 0.0967 | 0.0961 | 0.0952 | 0.0956 | 0.0962 |
|  | 0.0146 | 0.0142 | 0.0147 | 0.0138 | 0.0134 | 0.0145 |
|  | -0.0272 | -0.0290 | -0.0287 | -0.0294 | -0.0283 | -0.0298 |

函数机制得到的回归模型为：

表5.3 函数机制回归模型

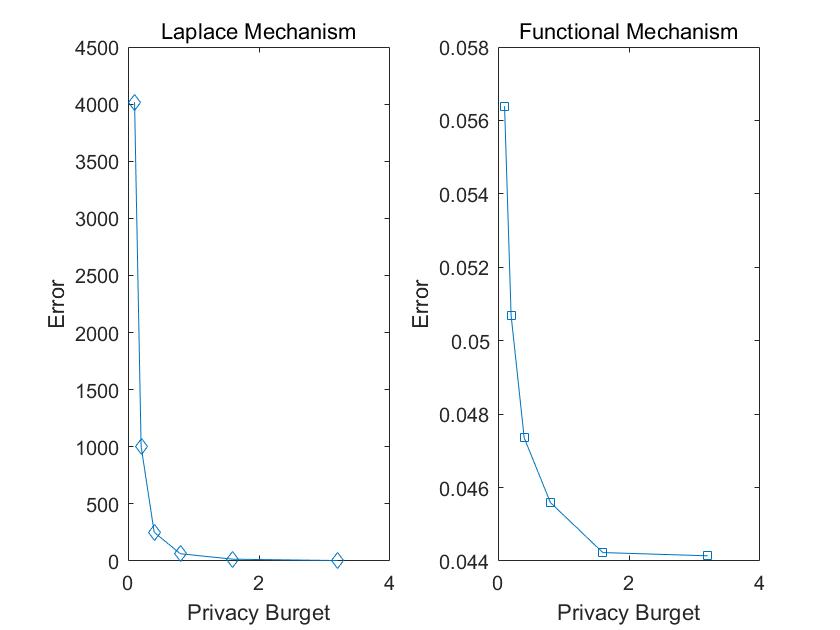
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
|  | 0.0366 | -0.0732 | -0.2775 | -0.1320 | -0.2224 | -0.2629 |
|  | 0.0214 | 0.0042 | 0.0165 | 0.0084 | 0.0094 | 0.0044 |
|  | 0.2543 | 0.3007 | 0.0449 | 0.2468 | 0.2924 | 0.2864 |
|  | -0.0402 | -0.0078 | 0.2881 | 0.0939 | -0.0480 | -0.1083 |
|  | 0.1119 | 0.0376 | 0.0549 | 0.0121 | 0.0349 | 0.0447 |
|  | 0.0889 | 0.1828 | 0.1534 | 0.1102 | 0.1313 | 0.1888 |
|  | -0.0411 | -0.0196 | -0.0328 | -0.0447 | -0.0436 | -0.0393 |
|  | 0.0313 | 0.1165 | 0.0321 | 0.0563 | 0.0414 | 0.0173 |
|  | -0.0304 | 0.0205 | 0.0140 | 0.0201 | 0.0180 | 0.0102 |
|  | 0.1527 | 0.1056 | 0.1056 | 0.1002 | 0.0601 | 0.1026 |
|  | 0.1626 | 0.1536 | 0.1882 | 0.1816 | 0.1706 | 0.1762 |
|  | -0.0197 | 0.1301 | 0.1430 | 0.1224 | 0.0832 | 0.0895 |
|  | -0.2334 | -0.1577 | 0.0231 | -0.0297 | -0.0017 | 0.0050 |
|  | 0.0484 | -0.1352 | 0.0234 | 0.0309 | 0.0065 | -0.022 |

三种模型所产生的误差分别为：

表5.4 误差值

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| 直接加噪（1.0e+03\*） | 3.999 | 1.012 | 0.2553 | 0.0635 | 0.0160 | 0．0040 |
| 函数机制 | 0．0569 | 0．0493 | 0.0481 | 0.0452 | 0.0445 | 0.0438 |
| 未使用差分隐私算法 | 0.0109 | | | | | |

根据该表格绘制的图像如图5-3所示，通过这些表格中的数据以及图像可以看出，这两种基于差分隐私的回归分析算法的实现机制，函数机制构建的模型显然比直接添加噪声机制构建的模型要好得多。直接添加噪声的机制将噪声直接添加在预测值的末尾，由于该模型的敏感度较大，所以造成添加的噪声量很大，使得添加噪声后的数据与原始数据的差值变得很大，从而引起了巨大的误差；而对于函数机制而言，噪声是添加在代价函数的系数上，在后续的各种计算过程中会逐渐降低由于系数添加的噪声量所带来的差值，从而使得获得的回归模型计算出的预测值更加接近真实值。同时，在图中我们也可以看出，不论是函数机制还是直接添加噪声的机制，误差值随着隐私预算的增大而减小。但是虽然误差值减小了，这并不意味着模型拟合的效果更好，因为误差值减小的代价是增大了隐私预算，而隐私预算增大使得拟合后的模型的可用性降低。

图5-3 误差分析对比

5.4.2 算法性能

算法性能是衡量算法训练和测试回归模型所需要的时间

表5.5 两种机制的算法性能

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| 直接加噪 | 5.0301 | 4.6825 | 5.000 | 5.0613 | 4.8430 | 4.5083 |
| 函数机制 | 0．1015 | 0.1027 | 0.1063 | 0.1102 | 0.1072 | 0.1107 |
| 未使用差分隐私算法 | 4.4074 | | | | | |

由上表可以看出，函数机制训练和测试回归模型所需要的时间远远低于直接加噪机制所需要的时间。函数机制的效率主要是由于在进行正则化的过程中间接对代价函数进行了优化，例如它会消除特征值为负数所对应的行和列，从而降低了矩阵的维度，这些因素导致了函数机制在代价函数求解的过程中更容易趋于最优解，求解所消耗的时间大大降低。而对于直接加噪机制，由于未对代价函数进行处理，算法需要消耗大量的时间去求解最优解，从而消耗大量时间。

5.5 函数机制的隐私性分析

通过这些分析可以得出结论，基于差分隐私的回归模型实现机制中，不论是算法误差还是算法性能上面，函数机制都要明显优于直接添加噪声的机制，而且函数机制在保证高效率的同时也保证了较低的误差。但是函数机制的最终目的是要实现数据的隐私性，那么函数机制的隐私性体现在哪些地方呢？

在实验数据集中我们随机取出数据集总数的20%作为测试数据集，测试数据集一共包含37515条数据。对于这37515条数据，将通过直方图的形式，描绘出敏感数据的分布情况。其中横坐标表示数值分布区间，纵坐标表示数量。

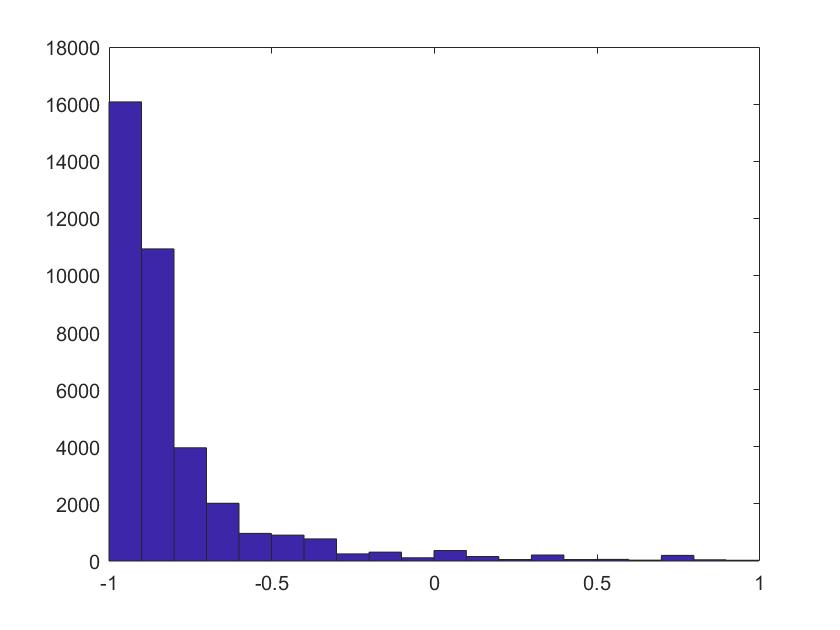
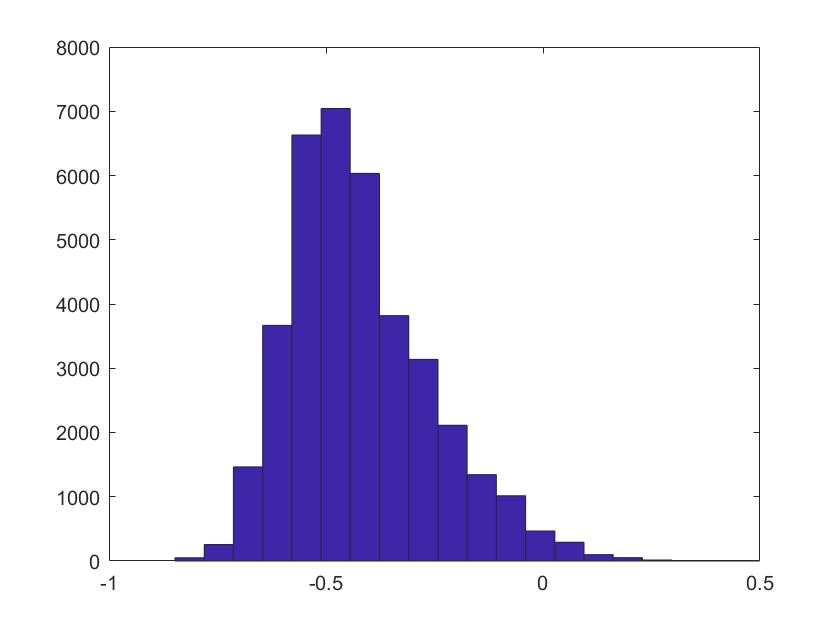
图5-4表示的是原始数据集中的敏感数据的直方分布，由图像我们可以看出，原始数据中的敏感数据主要分布在[-1,-0.5]这个区间范围内。

图5-4 原始数据分布

图5-5表示的是测试数据集通过上述未经隐私保护处理的回归模型计算所获得敏感数据的直方分布，该数据集的敏感数据主要分布在[-0.7,-0.3]范围内。

图5-5 回归模型数据分布

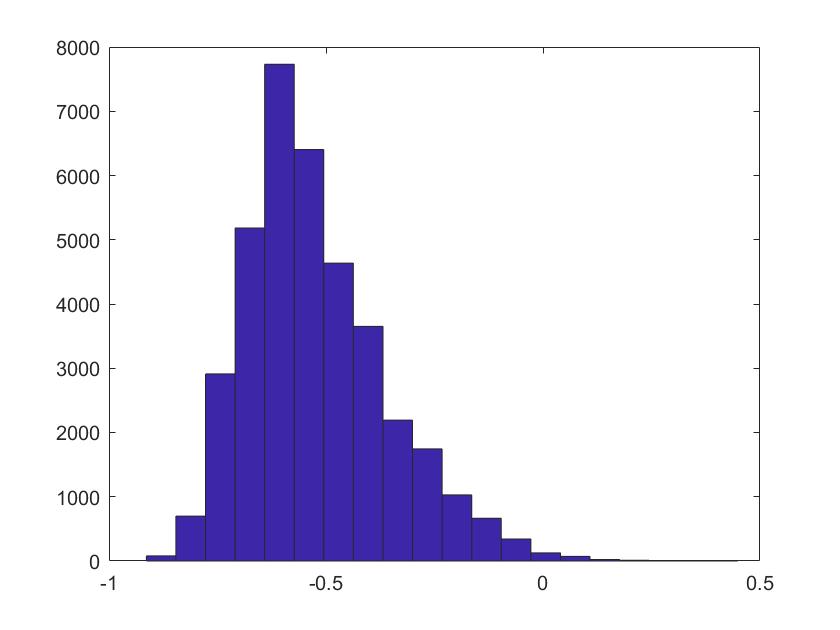
图5-6表示的是测试数据集通过上述函数机制处理的回归模型计算所获得敏感数据的直方分布，该数据集的敏感数据主要分布在[-0.7,-0.3]范围内。

图5-6 函数机制数据分布

通过这三个数据分布图可以看出，函数机制的数据分布与原始数据分布存在较大的差距，而这种差距恰恰说明了函数机制能够提供隐私保护，因为如果函数机制的数据分布无限接近于原始数据的分布，则对于每一个敏感数据个体来说，他们的预测值肯定也会接近于真实值，而这种情况下的模型存在很大的数据泄露的风险；而当我们比较函数机制的数据分布和回归模型的数据分布时我们发现，这两者之间的数据分布十分的接近，这就说明经过差分隐私处理的函数机制模型依然会保留回归模型中的统计特性，这就体现出了函数机制的可用性。

综上所述，基于差分隐私的函数机制回归模型具有误差小、性能好而同时又能兼具隐私性和实用性。

总结

随着隐私泄露危机的不断加剧，人们越发的重视个人的隐私问题，而差分隐私技术是建立在没有量化攻击者背景知识的基础上，相对于传统的隐私保护技术，差分隐私技术具备更好的应用价值。差分隐私技术与回归分析技术结合只是差分隐私技术应用领域的一部分，未来差分隐私技术将应用于机器学习、数据挖掘与发布等各个领域。

目前差分隐私技术的研究大多处于理论研究的范围，本文的主要贡献在于全面的介绍了差分隐私技术和回归分析技术的相关概念和算法，同时将差分隐私技术与回归分析技术相结合，并且实现了两种基于差分隐私的回归模型。虽然取得了一些成果，但依然存在很多需要完善的地方，比如：构建的回归模型相对来说比较简单，可以通过增加回归模型的维度，使得拟合的结果更加精确；构建直接添加噪声的模型时，由于代价函数求最小值迭代次数过多导致算法性能下降在后期的工作当中，希望能够通过不断的优化算法使得模型的拟合效果更好。；基于差分隐私的回归分析算法还有DPME算法、FP算法等等，如果本文能够实现这些算法，再与函数机制进行对比，这样会使得本文更全面，更具有说服力。

参考文献

[1] C. Dwork, "Differential privacy: A survey of results," in *International Conference on Theory and Applications of Models of Computation*, 2008, pp. 1-19.

[2] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography Conference*, 2006, pp. 265-284.

[3] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science,* vol. 9, pp. 211-407, 2014.

[4] A. Friedman and A. Schuster, "Data mining with differential privacy," presented at the Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, DC, USA, 2010.

[5] K. Nissim, S. Raskhodnikova, and A. Smith, "Smooth sensitivity and sampling in private data analysis," in *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, 2007, pp. 75-84.

[6] L. Wang and X.-F. MENG, "Location privacy preservation in big data era: a survey," *Journal of Software,* vol. 25, pp. 693-712, 2014.

[7] 李涛, "网络安全中的数据挖掘技术," *清华大学出版社,* 2017.

[8] 刘雅辉, 张铁赢, 靳小龙, 程学旗, "大数据时代的个人隐私保护," *计算机研究与发展,* p. 1, 2014.

[9] 钱萍, 吴蒙, "同态加密隐私保护数据挖掘方法综述倡," *计算机应用研究,* vol. 28, 2011.

[10] 王宝楠, "基于差分隐私拉普拉斯机制的线性回归分析研究," 安徽理工大学, 2016.

[11] 王宝楠, 方贤进, "基于差分隐私的线性回归分析," *电脑知识与技术,* pp. 26-29, 2016.

[12] 王平水, 马钦娟, "隐私保护 k-匿名算法研究," *计算机工程与应用,* vol. 47, pp. 117-119, 2011.

[13] 熊平, 朱天清, 王晓峰, "差分隐私保护及其应用," *计算机学报,* vol. 37, pp. 101-122, 2014.

[14] 张啸剑, 孟小峰, "面向数据发布和分析的差分隐私保护," *计算机学报,* pp. 927-949, 2014.

[15] 周志华, "机器学习," *清华大学出版社,* 2016.

[16] 邹鸿珍, "基于差分隐私的回归分析算法研究," 江西理工大学, 2016.

致谢

时光飞逝，转眼间四年的大学生活已经来到了尾声，在完成毕业设计的过程中，感谢方老师对我耐心的指导，指出了我的不足之处，并时刻关心课题进展。

在完成毕业设计的过程中也遇到了一些困难，方老师积极的为我提供资料，丰富我的知识，而且教会了我如何在大量的文献中寻找自己所需要的东西，并将这些知识能够为自己所用。通过这次毕业设计，我也有了很大的收获，由于自身知识面的狭窄以及能力方面的欠缺，在完成毕业设计的过程中对我进行了耐心的教导，帮我提高个人的能力，扩展自己的知识面。

我深刻的感觉到作为方老师学生的荣幸和自豪，我为有这样一位认真负责的老师而感到高兴。在此，我要对方老师表达诚挚的感谢与敬意，同时也要感谢学校四年来对我的培养。最后祝方老师身体健康、工作顺利，祝母校越来越好。