Operations Research and Fuzziology 运筹与模糊学, 2020, 10(2), 122-138
Published Online May 2020 in Hans. http://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2020.102013

Hans汉斯

New Numerical Methods for Robust Regularization Problem Based on Alternating Direction Method of Multipliers

Fengrui Ji, Jie Wen*

College of Science, Nanjing University of Aeronautics and Astronautics, Nanjing Jiangsu
Email: *wenjie@nuaa.edu.cn

Received: Mar. 25th, 2020; accepted: Apr. 8th, 2020; published: Apr. 15th, 2020

Abstract

Linear regression is an important model in machine learning, but too large data will make the linear regression model fall into over-fitting problem. Regularization is the main method to solve the over-fitting problem. The standard regularization model is the method that linear regression co-operates with regularization to ensure the quality of classification and reduce the risk of falling into over-fitting. However, the classifier obtained by the standard regularization model will be unstable when the training samples are disturbed. To solve this problem, we propose two kinds of robust regularization models: 1) Stochastic robust regularization: combining the expectation of residual with the regularization; 2) Worst case robust regularization: combining the worst case residual with the regularization. Then, we use Alternating Direction Method of Multipliers (ADMM) algorithm to get the optimal solution of the robust regularization model. Numerical experiments show that the classifiers obtained by stochastic and worst-case robust regularization have good robustness when training disturbed data sets, but the classifiers obtained by standard regularization method fluctuate greatly.

Keywords

Linear Regression, Regularization, ADMM, Robustness

基于ADMM算法的鲁棒正则化问题求解方法

吉锋瑞，文　杰*

南京航空航天大学理学院，江苏　南京
Email: *wenjie@nuaa.edu.cn

收稿日期：2020年3月25日；录用日期：2020年4月8日；发布日期：2020年4月15日

*通讯作者。

文章引用: 吉锋瑞, 文杰. 基于 ADMM 算法的鲁棒正则化问题求解方法[J]. 运筹与模糊学, 2020, 10(2): 122-138.
DOI: 10.12677/orf.2020.102013

摘　要

线性回归是机器学习中的重要模型，但是过大的数据量会使线性回归模型陷入过拟合。正则化技术是解决过拟合问题的主要方法。标准的正则化模型是运用线性回归项和正则项协同作用以达到保证分类质量和降低陷入过拟合风险的目的。但是当训练样本存在扰动时，标准的正则化模型得到的分类器会因为数据集存在扰动而表现得不稳定。针对该缺点，本文提出了两种鲁棒正则化模型：1) 随机鲁棒正则化：将残差的数学期望和正则项结合；2) 最坏情况鲁棒正则化：将最坏情况的残差和正则项结合。然后利用交替方向法(交替方向法，ADMM)求得鲁棒正则化模型的最优解。数值实验显示：当训练存在扰动的数据集时，随机鲁棒正则化和最坏情况鲁棒正则化得到的分类器具有很好的鲁棒性，而标准正则化方法得到的分类器波动很大。

关键词

线性回归，正则化，ADMM，鲁棒性

1. 介绍

分类问题是机器学习中一类重要的问题。很多分类模型都可以很好地解决分类问题，其中线性回归就是一类重要的分类模型。线性回归[1]的数学模型如下：

$$\min_x \|Ax-b\| \tag{1.1}$$

其中 $A\in R^{m\times n}$ 和 $b\in R^m$ 是训练样本的数据，$x\in R^n$ 是回归系数。由于目标函数是凸函数，所以该问题是凸优化问题。当 $b\in R(A)$ 时，该问题的最优值是 0。在绝大部分时候，模型通过训练数据集学习经验或者知识时，得到的最优值往往不是 0，即 $b\notin R(A)$。所以 $b\notin R(A)$ 时的问题是非常有实用意义的。线性回归能够以较快的速度做出较高质量的预测，所以它被人们广泛地应用于各个领域。[2]通过多元线性回归模型分析了影响武汉市水资源承载力的主要因素，[3]通过线性回归模型讨论了温度与月售电量的关系，[4]通过线性回归模型分析了影响全国粮食产量的因素，得到了线性回归方程，并发现该方程具有较高的准确度。

当训练样本的维数远远小于训练样本的数据量时，模型(1.1)容易陷于过拟合。过拟合问题是模型(1.1)过分学习训练样本的经验或者知识，将训练样本的特殊特征作为普遍特征。它的表现形式是模型(1.1)训练得到的结果在训练集上表现很好，但是在测试集上表现不佳。正则化技术是解决过拟合问题的常用方法。正则化模型如下：

$$\min_x \|Ax-b\|+\lambda g(x) \tag{1.2}$$

其中 $g(x)$ 是正则项，$\|Ax-b\|$ 为保真项，$\lambda>0$ 是正则项参数。正则项参数平衡了残差保真项 $\|Ax-b\|$ 和正则项 $g(x)$ 的关系。λ 较小时，残差较小，分类准确率高，但是陷入过拟合的风险也就相对较大。λ 较大时，模型泛化能力会有所增强，但是残差较大，导致分类准确率较低。所以选择一个合适的 λ 是很重要的。当 $g(x)$ 是凸函数时，问题(1.2)是凸优化问题。常见的凸正则项有 LASSO [5]、Fused LASSO [6]、

Elastic Net [7]、L2 正则项[8]等等。由于它们都是凸优化问题，所以有很多优化方法求得最优解。正则项函数不仅仅局限于凸函数，常见的非凸正则项函数有 MCP 模型[9]、SCAD 模型[10]、桥回归模型[11]等等。由于这些模型都是非凸优化问题，所以求解较为困难，研究的相对较少。

当正则项函数是 L1 范数，度量残差 $Ax-b$ 的范数是 L2 范数时，问题(1.2)就是 LASSO 问题。由于 L1 范数会压缩回归系数，所以 LASSO 问题的回归系数具有稀疏性。LASSO 问题如(1.3)：

$$\min_x \frac{1}{2}\|Ax-b\|_2^2+\lambda\|x\|_1 \tag{1.3}$$

λ 平衡了模型的分类准确率和回归系数的稀疏度。λ 较大时，回归系数的稀疏度较大但是分类准确率较低。λ 较小时，分类准确率较高，但是回归系数的稀疏度较低，所以选择合适的 λ 相当重要。另外具有稀疏性的模型能够很好地对数据进行拟合，并且计算简单，所以被人们广泛地应用多个领域。随着人们对该领域的深入研究，发现押注稀疏性原理[12]：既然无法处理有效稠密问题，倒不如在稀疏问题上寻找有效的处理方法。由于 L1 正则项是凸函数并且能够保证回归系数具有稀疏性，所以被很多机器学习的模型用来解决过拟合问题。但是模型(1.3)在遇到存在扰动的数据时会陷入麻烦。对训练样本加 100 个随机扰动，每个扰动的幅度都是训练样本 10%。模型(1.3)通过训练原始训练样本求得一组回归系数。将求得的回归系数通过线性回归模型预测存在扰动的数据集类别，得到一组分类准确率。这组分类准确率构成一组数组，绘成图 1。

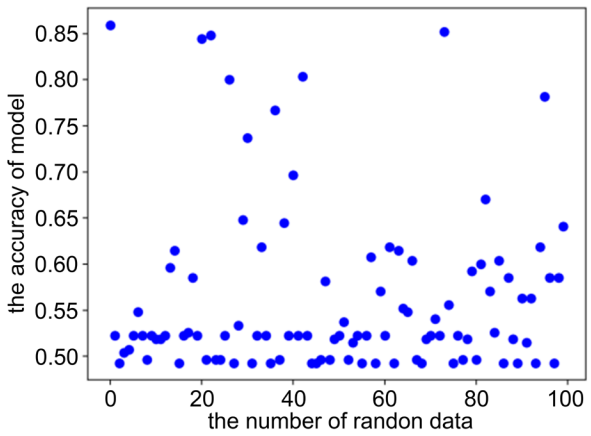

Figure 1. Training results of Model (1.3) when the training samples are disturbed
图 1. 当训练样本存在扰动时，模型(1.3)训练结果

通过图 1 发现当数据存在扰动时，LASSO 问题的分类准确率波动很大。最高分类准确率与最低分类准确率之差大约有 30%。由此可以看出，当数据集存在一定波动时，LASSO 问题的训练结果受到较大影响。

以 L1 正则项为例，发现标准的正则化方法不能够很好地训练受到扰动的训练样本。为了解决这个问题，本文提出了两种鲁棒正则化模型解决训练样本存在扰动的问题(本文所使用的正则项是 L1 正则项、L2 正则项、L1-L2 正则项和 Huber 正则项)。第一种模型是利用概率论的方法建立随机鲁棒正则化模型。该模型认为扰动是随机变量，所以控制拟合效果的方法由传统的 $\|Ax-b\|$ 替换为 $\frac{1}{2}E\|Ax-b\|_2^2$。通过调整正则项参数平衡 $\frac{1}{2}E\|Ax-b\|_2^2$ 和正则项之间的关系。第二种模型是通过定义最坏情况的误差建立最坏情况鲁棒正则化模型。该模型保障拟合效果的方法是考虑最坏情况下的残差。同样设定合理的正则项参数可以平衡最坏情况的残差和正则项。当正则项是 L1、L2、L1-L2 正则项时，本文将鲁棒正则化模型分为两

块优化问题进行求解。当正则项是 Huber 正则项时，本文将鲁棒正则化模型分为三块优化问题进行求解。数值实验显示鲁棒正则化模型能够保证分类准确率的同时也具有较强的鲁棒性。另外，数值实验还说明了鲁棒 Huber 正则化和 L1-L2 正则化模型分类准确率表现优于鲁棒 L1、L2 正则化模型。

本文的内容安排如下：第二章介绍正则化方法；第三章介绍两块 ADMM 算法和三块 ADMM 算法；第四章介绍随机鲁棒逼近正则化以及求解方法；第五章介绍最坏情况鲁棒正则化模型以及求解方法；第六章通过数值试验比较了鲁棒正则化方法和标准正则化方法；第七章对本文做出总结。

2. 基础知识

在大数据时代，数据的规模是十分巨大的。过大的数据量容易使机器学习中的模型陷入过拟合困扰。该模型会过分学习训练集上的知识，从而认为特殊性质就是普遍性质。正则化方法通过添加正则项的方法来降低模型陷入过拟合的风险。本文将介绍正则化方法以及常见的正则项。

2.1. 正则化方法

一般正则化模型如下：

$$\min_x \frac{1}{2}\|Ax-b\|_2^2+\lambda g(x) \tag{2.1}$$

其中 $\frac{1}{2}\|Ax-b\|_2^2$ 度量了残差 $Ax-b$ 的大小，g 从某种意义上度量了 x 的大小，$\lambda>0$ 是正则项参数。正则项参数平衡了保真项 $\frac{1}{2}\|Ax-b\|_2^2$ 和 $g(x)$。本文所使用的正则项函数 g 都是凸函数，所以(2.1)是凸优化问题。引入变量 y，令 $x=y$，问题则转化为等式约束下可分的优化问题：

$$\begin{aligned}&\min_{x,y} \frac{1}{2}\|Ax-b\|_2^2+\lambda g(y)\\ &\text{s.t. } x-y=0\end{aligned} \tag{2.2}$$

该问题的拉格朗日函数为：

$$L(x,y,u)=\frac{1}{2}\|Ax-b\|_2^2+\lambda g(y)+u^{\mathrm{T}}(x-y) \tag{2.3}$$

其中 u 是对偶变量，$\lambda>0$ 是正则项参数。

性质 1. 问题(2.1)的最优性条件是：

$$\begin{cases}0=A^{\mathrm{T}}\left(Ax^*-b\right)+u^*\\ 0\in\lambda\partial g\left(y^*\right)-u^*\\ 0=x^*-y^*\end{cases} \tag{2.4}$$

其中 ∂ 是次微分算子。若正则项函数 $g(y)$ 是可导函数，$\partial g\left(y^*\right)$ 则是 $\nabla g\left(y^*\right)$。

2.2. L1 正则化

当 $g(x)=\|x\|_1$ 时，模型(2.1)就是 LASSO，模型如(2.5)：

$$\min_x \frac{1}{2}\|Ax-b\|_2^2+\lambda\|x\|_1 \tag{2.5}$$

正则项参数 $\lambda>0$ 可以平衡分类准确率和回归系数的稀疏性。λ 较大时，回归系数较为稀疏，模型不易陷入过拟合，但是模型拟合效果较差。λ 较小时，则 λ 较大时相反。LASSO 回归有着广泛的应用，[13]使用 LASSO 回归对垃圾邮件过滤，[14]建立 LASSO 回归模型和计量经济学模型解释了实体金融市场与经济政策的内在关系，[15]将 LASSO 回归应用于三波段红外火焰探测器的具体识别中。

2.3. L2 正则化

当 $g(x)=\|x\|_2^2$ 时，模型(2.1)就是 Ridge 回归，模型如下：

$$\min_x \frac{1}{2}\|Ax-b\|_2^2+\lambda\|x\|_2^2 \tag{2.6}$$

正则项参数 $\lambda>0$ 可以平衡分类准确率和回归系数的规模。λ 较大时，回归系数规模较小，模型不易陷入过拟合，但是模型拟合效果较差。λ 较小时，则与之相反。[16]利用 ridge 回归分析了社会因素、经济因素以及政策制度因素对农地城市流转的影响，[17]利用 ridge 回归研究了对外贸易和外商直接投资对人力资本存量的影响。

2.4. L1-L2 正则化

当 $g(x)=\alpha\|x\|_1+(1-\alpha)\|x\|_2^2$ 时，模型如下：

$$\min_x \frac{1}{2}\|Ax-b\|_2^2+\lambda\left\{\alpha\|x\|_1+(1-\alpha)\|x\|_2^2\right\} \tag{2.7}$$

其中正则化参数 $\lambda>0$，$\alpha>0$。当 $\alpha=1$ 时，L1-L2 正则化就是 L1 正则化，当 $\alpha=0$ 时，L1-L2 正则化就是 L2 正则化。α 越大，则回归系数越稀疏。如果希望回归系数规模较小，则适当调小 α。这样可以选择合适的 λ 通过控制 α 的大小控制回归系数的稀疏性和规模。另外，不难发现 L1 正则项对小残差灵敏和 L2 正则项对大残差灵敏，但是 L1-L2 正则项通过组合 L1 正则项和 L2 正则项的方法克服上述两个正则项的缺点。

2.5. Huber 正则化

当 $g(x)=\text{Huber}(x)$ 时，问题(2.1)就是 Huber 正则化。Huber 正则化模型如(2.8)：

$$\min_x \frac{1}{2}\|Ax-b\|_2^2+\lambda\text{Huber}(x) \tag{2.8}$$

其中，

$$\text{Huber}(x)=\begin{cases}\frac{1}{2}x^2, & |x|\le 1\\ |x|-\frac{1}{2}, & |x|>1\end{cases} \tag{2.9}$$

Huber 正则项是通过定义分段函数的方法克服 L1 范数、L2 范数的缺点。该分段函数的特点是面对小误差时使用 L2 范数，面对大误差时使用 L1 范数，这与 L1-L2 正则项有本质不同，所以这两个正则项不是一个正则项。由于 $\text{Huber}(x)=\min_v \|v\|_1+\frac{1}{2}\|x-v\|_2^2$，所以问题(2.8)等价于：

$$\min_x \min_v \frac{1}{2}\|Ax-b\|_2^2+\lambda\left(\|v\|_1+\frac{1}{2}\|x-v\|_2^2\right) \tag{2.10}$$

这是一个凸的不光滑优化问题，使用一些非光滑技术可以求解出该问题的最优解。

3. ADMM 算法

ADMM 是一种可以高效求解等式约束下可分结构的凸优化问题的算法。该算法由于在统计学习、图像处理等数据分析处理领域有着良好的表现而备受关注。本节将讨论两块 ADMM 算法和三块 ADMM 算法。

3.1. 两块 ADMM 算法

当优化分体分成两个整体时，模型可以表示为：

$$\begin{aligned}&\min\ f_1(x)+f_2(y)\\&\text{s.t.}\ \ Ax-By=0\end{aligned}\tag{3.1}$$

其中 $x,y\in R^n$， $A,B\in R^{m\times n}$。该问题的增广拉格朗日函数为：

$$L_\rho(x,y,u)=f_1(x)+f_2(y)+u^{\mathrm{T}}(Ax-By)+\frac{\rho}{2}\|Ax-By\|_2^2\tag{3.2}$$

其中 $u\in R^n$ 是对偶变量， $\rho>0$ 是增广拉格朗日乘子。ADMM 算法子问题如下：

$$x_{k+1}=\arg\min_x\ f_1(x)+u_k^{\mathrm{T}}(x-y_k)+\frac{\rho}{2}\|Ax-By_k\|_2^2\tag{3.3}$$

$$y_{k+1}=\arg\min_y\ f_2(y)+u_k^{\mathrm{T}}(x_{k+1}-y)+\frac{\rho}{2}\|Ax_{k+1}-By\|_2^2\tag{3.4}$$

$$u_{k+1}=u_k+\rho(Ax_{k+1}-By_{k+1})\tag{3.5}$$

令 $u_k=\frac{u_k}{\rho}$， $u_{k+1}=\frac{u_{k+1}}{\rho}$，上述子问题则转化为：

$$x_{k+1}=\arg\min_x\ f_1(x)+\frac{\rho}{2}\|u_k+Ax-By_k\|_2^2\tag{3.6}$$

$$y_{k+1}=\arg\min_y\ f_2(y)+\frac{\rho}{2}\|u_k+Ax_{k+1}-By\|_2^2\tag{3.7}$$

$$u_{k+1}=u_k+(Ax_{k+1}-By_{k+1})\tag{3.8}$$

假设 1 [18]. 假设 $f_1:R^n\to R\cup\{\infty\}$， $f_2:R^n\to R\cup\{\infty\}$ (或者 f_1， f_2 的延伸函数)是闭凸真函数。

当 f_1 是真函数时，它的上图 $\text{epi}\,f_1=\{(x,t)\in R^n\times R\mid f_1(x)\le t\}$ 非空且不包含竖直的直线。因此 f 是真函数等价于 $C=\text{dom}\,f$ 是非空凸集，f 是 C 上的有限函数。另外当 f_1 是闭函数时，则函数 f 对应的下水平集 $\{x\mid x\in\text{dom}\,f,f\le\alpha,\alpha\in R\}$ 是闭集。当 f_1 是凸函数时，则有 $0\le\theta\le1$ 时，$f_1(\theta x_1+(1-\theta)x_2)<\theta f_1(x_1)+(1-\theta)f_1(x_2)$。综上，满足假设 1 的 f_1 可以通过上图简洁表示，即 $\text{epi}\,f_1=\{(x,t)\in R^n\times R\mid f_1(x)\le t\}$ 是非空闭凸集。假设 1 保证求解 x 和 y 的优化子问题存在最优解，即使 f_1 和 f_2 是不可导函数，也存在 x^* 和 y^* 极小化问题(3.6)和问题(3.7)。

假设 2 [18]. L_0 存在鞍点。

鞍点存在等价于原问题存在最优解，对偶问题存在最优解，强对偶成立即原问题的最优值与对偶问题的最优值相同。如果凸优化问题满足 Slater 条件，则对偶问题存在最优解 u^* 且强对偶成立。所以加上原问题存在最优解就可以保证原问题存在鞍点。

定理 1. 若优化问题(3.1)满足下述性质，则(3.1)存在鞍点：

1) 优化问题(3.1)是凸优化问题；

2) 满足 Slater 条件，即 $\exists x \in \mathrm{Relint}\, D$，使得 $Ax - By = 0$，其中 D 是(3.1)的可行域；

3) 原始问题存在最优解。

通过定理 1，可以判断出优化问题(3.1)是否存在鞍点，那么就可以验证假设 2 是否成立。当假设 1 和假设 2 成立时，定理 2 保证了 ADMM 算法的收敛性。

定理 2 [18]. 假设 1 和假设 2 成立，则 ADMM 算法满足如下性质：

1) 残差收敛：当 $k \to \infty$ 时，有 $r^k \to 0$，其中 $r^k = Ax^k - By^k$；

2) 目标函数收敛：当 $k \to \infty$ 时，有 $f_1\left(x^k\right) + f_2\left(y^k\right) \to p^*$；

3) 对偶变量收敛：当 $k \to \infty$ 时，有 $u_k \to u^*$，其中 u^* 是对偶最优点。

3.2. 三块 ADMM 算法

当优化问题可以分成三块时，模型可以表示为：

$$\begin{aligned} &\min\ f_1(x) + f_2(y) + f_3(z) \\ &\text{s.t.}\ \ Ax + By + Cz = 0 \end{aligned} \tag{3.9}$$

其中 $x, y, z \in R^n$， $A, B, C \in R^{m\times n}$。该问题的增广拉格朗日函数为：

$$L_\rho(x, y, z, u) = f_1(x) + f_2(y) + f_3(z) + u^{\mathrm{T}}(Ax + By + Cz) + \frac{\rho}{2}\|Ax + By + Cz\|_2^2 \tag{3.10}$$

其中 $u \in R^n$ 是对偶变量， $\rho > 0$ 是增广拉格朗日乘子。ADMM 算法迭代算法如下：

$$x_{k+1} = \arg\min_x\ f_1(x) + u_k^{\mathrm{T}}(Ax + By_k + Cz_k) + \frac{\rho}{2}\|Ax + By_k + Cz_k\|_2^2 \tag{3.11}$$

$$y_{k+1} = \arg\min_y\ f_2(y) + u_k^{\mathrm{T}}(Ax_{k+1} + By + Cz_k) + \frac{\rho}{2}\|Ax_{k+1} + By + Cz_k\|_2^2 \tag{3.12}$$

$$z_{k+1} = \arg\min_z\ f_3(z) + u_k^{\mathrm{T}}(Ax_{k+1} + By_{k+1} + Cz) + \frac{\rho}{2}\|Ax_{k+1} + By_{k+1} + Cz\|_2^2 \tag{3.13}$$

$$u_{k+1} = u_k + \rho\left(Ax_{k+1} + By_{k+1} + Cz_{k+1}\right) \tag{3.14}$$

文献[19]讨论三块 ADMM 算法的收敛性。三块 ADMM 算法不能够保证收敛，只有加强一些条件才能保证三块 ADMM 算法的收敛性。由于本文没有增加一些条件保证三块 ADMM 算法的收敛性，所以在进行数值实验的过程中三块 ADMM 算法在一些情况下不收敛。

4. 随机鲁棒正则化模型

当训练样本受到扰动时，标准正则化方法不能应对这种情况。从图 1 中可以发现标准的 LASSO 问题在这种情况下分类结果受到了较大的影响。为了增强分类模型的稳定性，本节利用概率论的方法提出随机鲁棒正则化模型。

4.1. 随机鲁棒正则化

随机鲁棒正则化模型如下：

$$\min_x \frac{1}{2}E\|Ax - b\|_2^2 + \lambda g(x) \tag{4.1}$$

其中 $A=\hat{A}+U$，$U\in R^{m\times n}$ 是均值为 0 的随机矩阵，$\hat{A}\in R^{m\times n}$ 是随机变量 $A\in R^{m\times n}$ 的期望，$\lambda>0$ 是正则项参数。由于 $g(x)$ 和 $E\|Ax-b\|_2^2$ 是凸函数，所以问题(4.1)是凸优化问题。保真项 $E\|Ax-b\|_2^2$ 是训练受到扰动的数据时残差的期望。λ 是权衡保真项和正则项之间的关系。λ 越大，模型陷入过拟合的概率就越低，但也会牺牲模型的准确率。相反的，λ 较小时，分类准确率会很高，但是模型陷入过拟合的概率也会较高。由于 $E\|Ax-b\|_2^2$ 的导数难以计算，所以直接求解问题(4.1)过程繁琐。性质 2 可以将该问题转化为一个相对容易求解的问题。性质 2 如下：

性质 2. $\frac{1}{2}E\|Ax-b\|_2^2=\frac{1}{2}\|\hat{A}x-b\|_2^2+\frac{1}{2}x^{\mathrm{T}}E\left(U^{\mathrm{T}}U\right)x$。

通过性质 2，问题(4.1)则转化为：

$$\min_x \frac{1}{2}\|\hat{A}x-b\|_2^2+\frac{1}{2}x^{\mathrm{T}}E\left(U^{\mathrm{T}}U\right)x+\lambda g(x) \tag{4.2}$$

令 $x=y$，问题则转化为：

$$\begin{aligned}&\min \frac{1}{2}\|\hat{A}x-b\|_2^2+\frac{1}{2}x^{\mathrm{T}}E\left(U^{\mathrm{T}}U\right)x+\lambda g(y)\\&\text{s.t. } x-y=0\end{aligned} \tag{4.3}$$

该问题的拉格朗日函数为：

$$L(x,y,z,u)=\frac{1}{2}\|\hat{A}x-b\|_2^2+\frac{1}{2}x^{\mathrm{T}}E\left(U^{\mathrm{T}}U\right)x+\lambda g(y)+u^{\mathrm{T}}(x-y) \tag{4.4}$$

其中 $u\in R^n$ 是对偶变量。

性质 3. 问题(4.1)的最优性条件：

$$\begin{cases}0=\hat{A}^{\mathrm{T}}\left(\hat{A}x^*-b\right)+E\left(U^{\mathrm{T}}U\right)x^*+u^*\\0\in\lambda\partial g\left(y^*\right)-u^*\\0=x^*-y^*\end{cases} \tag{4.5}$$

4.2. ADMM 算法求解随机鲁棒正则化

上节已经说明随机鲁棒正则化模型等价如下问题：

$$\begin{aligned}&\min \frac{1}{2}\|\hat{A}x-b\|_2^2+\frac{1}{2}x^{\mathrm{T}}E\left(U^{\mathrm{T}}U\right)x+\lambda g(y)\\&\text{s.t. } x-y=0\end{aligned} \tag{4.6}$$

问题(4.6)的增广拉格朗日函数是 $\frac{1}{2}\|\hat{A}x-b\|_2^2+\frac{1}{2}x^{\mathrm{T}}E\left(U^{\mathrm{T}}U\right)x+\lambda g(y)+u^{\mathrm{T}}(x-y)+\frac{\rho}{2}\|x-y\|_2^2$，ADMM 迭代算法如下：

$$x_{k+1}=\arg\min_x \frac{1}{2}\|\hat{A}x-b\|_2^2+\frac{1}{2}x^{\mathrm{T}}E\left(U^{\mathrm{T}}U\right)x+\frac{\rho}{2}\|x-y_k+u_k\|_2^2 \tag{4.7}$$

$$y_{k+1}=\arg\min_y \lambda g(y)+\frac{\rho}{2}\|x_{k+1}-y+u_k\|_2^2 \tag{4.8}$$

$$u_{k+1}=u_k+\left(x_{k+1}-y_{k+1}\right) \tag{4.9}$$

性质 4. ADMM 算法在求解问题(4.6)时算法是收敛的。

证明：令 $f_1(x)=\frac{1}{2}\left\|\hat{A}x-b\right\|_2^2+\frac{1}{2}x^{\mathrm{T}}E\left(U^{\mathrm{T}}U\right)x$，$f_2(y)=\lambda g(y)$。易知 $f_1(x)$ 和 $f_2(y)$ 是闭凸真函数，所以问题(4.6)满足假设 1。由于 $f_1(x)$ 和 $f_2(y)$ 都是凸函数，所以问题(4.6)是凸优化问题。又因为 $x,y\in R^n$，所以 $\exists x\in\mathrm{Relint}\,D$，有 $x-y=0$，所以问题(4.6)满足 Slater 条件。$f_1(x)$ 和 $f_2(y)$ 是闭凸真函数，所以原问题存在最优解 x^* 和 y^*。综上，问题(4.6)满足假设 2。由于问题(4.6)满足假设 1，假设 2，所以 ADMM 求解问题(4.6)时算法是收敛的。

问题(4.7)存在解析解 $x^*=\left(\hat{A}^{\mathrm{T}}\hat{A}+E\left(U^{\mathrm{T}}U\right)+\rho I\right)^{-1}\left(\hat{A}^{\mathrm{T}}b+\rho(y-u)\right)$。当 $g(y)=\|y\|_2^2$ 时，易知问题(4.8)存在解析解 $\frac{\rho x+\rho u}{\rho+2\lambda}$。当 $g(y)=\|y\|_1$ 时，利用分段讨论的方法可以得到问题(4.8)的解析解 $S_{\frac{\lambda}{\rho}}(x+u)$。

性质 5. 当 $g(y)=\|y\|_1$ 时，问题(4.8)的解析解 $S_{\frac{\lambda}{\rho}}(x+u)$。

证明：

Case 1. $y_j>0$：

$$\lambda+\rho\left(y_j-x_j-u_j\right)=0$$

$$y_j=x_j+u_j-\frac{\lambda}{\rho}$$

Case 2. $y_j<0$：

$$-\lambda+\rho\left(y_j-x_j-u_j\right)=0$$

$$y_j=x_j+u_j+\frac{\lambda}{\rho}$$

Case 3. $y_j=0$：

$$\rho\left(y_j-x_j-u_j\right)=0$$

$$y_j=x_j+u_j$$

其中 $S_b(x)$ 的定义如下：

$$S_b(x)=\begin{cases}x-b, & x>b\\ x+b, & x<-b\\ x, & |x|\le b\end{cases}$$

同理，当 $g(y)=\alpha\|y\|_1+(1-\alpha)\|y\|_2^2$ 时，问题(4.8)的最优解是 $S_{\frac{\lambda\alpha}{\lambda(1-\alpha)+\rho}}\left(\frac{\rho x+\rho u}{\lambda(1-\alpha)+\rho}\right)$。

当 $g(y)=\mathrm{Huber}(y)$ 时，由(2.5)可知鲁棒 Huber 正则化方法等价于：

$$\begin{aligned}&\min\frac{1}{2}\left\|\hat{A}x-b\right\|_2^2+\frac{1}{2}x^{\mathrm{T}}E\left(U^{\mathrm{T}}U\right)x+\lambda\left(\|v\|_1+\frac{1}{2}\|y\|_2^2\right)\\&\text{s.t. }\ x-v=y\end{aligned}\tag{4.10}$$

该问题的增广拉格朗日函数是

$$\frac{1}{2}\left\|\hat{A}x-b\right\|_2^2+\frac{1}{2}x^{\mathrm{T}}E\left(U^{\mathrm{T}}U\right)x+\lambda\left(\|v\|_1+\frac{1}{2}\|y\|_2^2\right)+u^{\mathrm{T}}(x-v-y)+\frac{\rho}{2}\|x-v-y\|_2^2\tag{4.11}$$

该问题的 ADMM 子问题如下：

$$x_{k+1}=\arg\min_x \frac{1}{2}\left\|\hat{A}x-b\right\|_2^2+\frac{1}{2}x^{\mathrm{T}}E\left(U^{\mathrm{T}}U\right)x+\frac{\rho}{2}\left\|x-y_k-v_k+u_k\right\|_2^2 \tag{4.12}$$

$$v_{k+1}=\arg\min_v \lambda\|v\|_1+\frac{\rho}{2}\left\|x_{k+1}-y_k-v+u_k\right\|_2^2 \tag{4.13}$$

$$y_{k+1}=\arg\min_y \lambda\frac{1}{2}\|y\|_2^2+\frac{\rho}{2}\left\|x_{k+1}-y-v_{k+1}+u_k\right\|_2^2 \tag{4.14}$$

$$u_{k+1}=u_k+\left(x_{k+1}-y_{k+1}-v_{k+1}\right) \tag{4.15}$$

问题(4.12)存在解析解 $\left(\hat{A}^{\mathrm{T}}\hat{A}+E\left(U^{\mathrm{T}}U\right)+\rho I\right)^{-1}\left(\hat{A}^{\mathrm{T}}b+\rho\left(y+v-u\right)\right)$，问题(4.13)存在解析解 $S_{\frac{\lambda}{\rho}}\left(x+u-y\right)$，问题(4.14)存在解析解 $\frac{\rho y+\rho v-\rho u}{\rho+\lambda}$。以上问题使用的是三块 ADMM 算法，所以不能够保证算法是收敛的。

5. 最坏情况鲁棒正则化

随机鲁棒正则化是利用概率论的方法确定保真项，利用正则化的思想避免模型陷入过拟合问题。本节将从另外一个角度构造保真项，并以此为基础构造最坏情况鲁棒正则化。

5.1. 最坏情况鲁棒正则化

随机鲁棒逼近是利用概率论的方法描述训练样本的不确定，本节将利用定义最坏情况的方法描述训练样本的不确定性，定义最坏误差如下：

定义 1. 最坏误差

$$e_{wc}\left(x\right)=\sup\left\{\|Ax-b\|\,\middle|\,A\in\mathbb{A}\right\} \tag{5.1}$$

其中 $\mathbb{A}$ 是不确定集，即描述 $A\in R^{m\times n}$ 所有情况的集合。由于范数是凸函数，所以任意不确定集 $\mathbb{A}$ 的最坏误差都是凸函数。

最坏误差的定义刻画了训练样本的不确定。将最坏情况鲁棒逼近与正则项结合，构成最坏情况鲁棒正则化，模型如(5.2)：

$$\min_x \frac{1}{2}e_{wc}\left(x\right)^2+\lambda g\left(x\right) \tag{5.2}$$

其中 $A=\hat{A}+U$，$\hat{A}\in R^{m\times n}$ 是随机变量 $A\in R^{m\times n}$ 的期望。正则项 $g\left(x\right)$，$\lambda>0$ 是正则项参数，保真项 $e_{wc}\left(x\right)=\sup\left\{\|Ax-b\|\,\middle|\,\|U\|\le a\right\}$。$e_{wc}\left(x\right)$ 是最坏情况下的误差。λ 平衡最坏情况下的误差和正则项之间的关系。由于最坏误差和 L1 范数是凸函数，所以该问题是凸优化问题。由于讨论 $\mathbb{A}$ 的所有情况过于复杂，所以本文讨论的最坏误差是 $e_{wc}\left(x\right)=\sup\left\{\left\|\hat{A}x+Ux-b\right\|\,\middle|\,\|U\|\le a\right\}$ $(U\in R^{m\times n})$。

性质 6. 在满足如下条件时，有 $e_{wc}\left(x\right)=\left\|\hat{A}x-b\right\|_2+a\|x\|_2$：

1) 向量 $\hat{A}x-b$ 的范数是 L2 范数；
2) 矩阵 U 的范数是 2 范数；
3) $\hat{A}x-b\neq 0$ 和 $x\neq 0$。

证明：

$$\left\|\hat{A}x+Ux-b\right\|_2^2=\left\|\hat{A}x-b\right\|_2^2+2\left(\hat{A}x-b\right)^{\mathrm{T}}Ux+\left\|Ux\right\|_2^2$$

当$\left(\hat{A}x-b\right)^{\mathrm{T}}$与$Ux$同向时，$\left\|\hat{A}x+Ux-b\right\|_2^2$取到最大值$\left(\left\|\hat{A}x-b\right\|_2+\left\|Ux\right\|_2\right)^2$。

由于$\|U\|\le a$，所以可知$\|Ux\|_2\le a\|x\|_2\to\sup\left\{\|Ux\|_2\,\middle|\,\|U\|\le a\right\}=a\|x\|_2$。

综上，$e_{wc}(x)=\sup\left\{\left\|\hat{A}x+Ux-b\right\|\middle|\,\|U\|\le a\right\}=\left\|\hat{A}x-b\right\|_2+a\|x\|_2$。

根据性质 6 最坏情况的鲁棒正则化模型可以转化为：

$$\min_x\frac{1}{2}\left(\left\|\hat{A}x-b\right\|_2+a\|x\|_2\right)^2+\lambda g(x)\tag{5.3}$$

令$x=y$，模型(5.3)可以转化为：

$$\begin{aligned}&\min\frac{1}{2}\left(\left\|\hat{A}x-b\right\|_2+a\|x\|_2\right)^2+\lambda g(y)\\&\text{s.t. }\ x-y=0\end{aligned}\tag{5.4}$$

问题(5.4)的拉格朗日函数是$\frac{1}{2}\left(\left\|\hat{A}x-b\right\|_2+a\|x\|_2\right)^2+\lambda g(y)+u^{\mathrm{T}}(x-y)$，其中$u\in R^n$是对偶变量。通过拉格朗日函数可以得到问题(5.4)的最优性条件。

性质 7. 问题(5.4)的最优性条件：

$$\begin{cases}0=\hat{A}^{\mathrm{T}}\left(\hat{A}x^*-b\right)+a^2x^*+\dfrac{\hat{A}^{\mathrm{T}}\left(\hat{A}x^*-b\right)}{\left\|\hat{A}x^*-b\right\|_2}\left\|x^*\right\|_2+\dfrac{x^*}{\left\|x^*\right\|_2}\left\|\hat{A}x^*-b\right\|_2+u^*\\0\in\lambda\partial g\left(y^*\right)-u^*\\0=x^*-y^*\end{cases}\tag{5.5}$$

5.2. ADMM 求解最坏情况鲁棒正则化

问题(5.4)的增广拉格朗日函数是$\frac{1}{2}\left(\left\|\hat{A}x-b\right\|_2+a\|x\|_2\right)^2+\lambda g(y)+u^{\mathrm{T}}(x-y)+\frac{\rho}{2}\|x-y\|_2^2$，其中$u\in R^n$是对偶变量，$\rho>0$是增广拉格朗日乘子。ADMM 子问题如下：

$$x_{k+1}=\arg\min_x\frac{1}{2}\left(\left\|\hat{A}x-b\right\|_2+a\|x\|_2\right)^2+\frac{\rho}{2}\|x-y_k+u_k\|_2^2\tag{5.6}$$

$$y_{k+1}=\arg\min_y\lambda g(y)+\frac{\rho}{2}\|x_{k+1}-y+u_k\|_2^2\tag{5.7}$$

$$u_{k+1}=u_k+\left(x_{k+1}-y_{k+1}\right)\tag{5.8}$$

性质 8. ADMM 求解问题(5.4)时算法是收敛的。

证明：令$f_1(x)=\frac{1}{2}\left(\left\|\hat{A}x-b\right\|_2+a\|x\|_2\right)^2$，$f_2(y)=\lambda g(y)$。易知$f_1(x)$和$f_2(y)$是闭凸真函数，所以问题(5.4)满足假设 1.由于$f_1(x)$和$f_2(y)$都是凸函数，所以问题(5.4)是凸优化问题。又因为$x,y\in R^n$，所以$\exists x\in\mathrm{Relint}D$，有$x-y=0$，所以问题(5.4)满足 Slater 条件。$f_1(x)$和$f_2(y)$是闭凸真函数，所以原问题存在最优解$x^*$和$y^*$，综上问题(5.4)满足假设 2。由于问题(5.4)满足假设 1，假设 2，所以 ADMM 求解问题(5.4)时算法是收敛的。

令 $\hat{f}(x)=\frac{1}{2}\left(\left\|\hat{A}x-b\right\|_2+a\|x\|_2\right)^2+\frac{\rho}{2}\|x-y_k+u_k\|_2^2$，有

$$\nabla\hat{f}(x)=\hat{A}^{\mathrm{T}}\left(\hat{A}x-b\right)+a^2x+\frac{\hat{A}^{\mathrm{T}}\left(\hat{A}x-b\right)}{\left\|\hat{A}x-b\right\|_2}\|x\|_2+\frac{x}{\|x\|_2}\left\|\hat{A}x-b\right\|_2+\rho\left(x-y_k+u_k\right) \tag{5.9}$$

得到(5.6)的导数(5.9)，可以利用很多经典的优化算法进行求解，但是大数据时代变量的维数和数据量会增加经典算法计算次数和存储空间。本文使用 L-BFGS 求解问题(5.6)的最优解。因为 L-BFGS 存储前 m 次迭代的少量数据，通过双回路递归的方法得到海塞阵逆矩阵的近似 H_k，所以会在一定程度上减少计算量。当 $g(y)$ 是 L1 正则项、L2 正则项、L1-L2 正则项时，问题(5.7)与问题(4.8)相同，此处不在赘述。当 $g(y)=\mathrm{Huber}(y)$ 时，该问题等价于

$$\begin{aligned}&\min\frac{1}{2}\left(\left\|\hat{A}x-b\right\|_2+a\|x\|_2\right)^2+\lambda\left(\|v\|_1+\frac{1}{2}\|y\|_2^2\right)\\&\text{s.t. }\ x-v=y\end{aligned} \tag{5.10}$$

该问题的增广拉格朗日函数是 $\frac{1}{2}\left(\left\|\hat{A}x-b\right\|_2+a\|x\|_2\right)^2+\lambda\left(\|v\|_1+\frac{1}{2}\|y\|_2^2\right)+u^{\mathrm{T}}(x-v-y)+\frac{\rho}{2}\|x-v-y\|_2^2$，其中 $u\in R^n$ 是对偶变量。该问题的 ADMM 子问题如下：

$$x_{k+1}=\arg\min_x\frac{1}{2}\left(\left\|\hat{A}x-b\right\|_2+a\|x\|_2\right)^2+\frac{\rho}{2}\|x-y_k-v_k+u_k\|_2^2 \tag{5.11}$$

$$v_{k+1}=\arg\min_v\lambda\|v\|_1+\frac{\rho}{2}\|x_{k+1}-y_k-v+u_k\|_2^2 \tag{5.12}$$

$$y_{k+1}=\arg\min_y\lambda\frac{1}{2}\|y\|_2^2+\frac{\rho}{2}\|x_{k+1}-y-v_{k+1}+u_k\|_2^2 \tag{5.13}$$

$$u_{k+1}=u_k+\left(x_{k+1}-y_{k+1}-v_{k+1}\right) \tag{5.14}$$

令 $\hat{f}(x)=\frac{1}{2}\left(\left\|\hat{A}x-b\right\|_2+a\|x\|_2\right)^2+\frac{\rho}{2}\|x-y_k-v_k+u_k\|_2^2$，有

$$\nabla\hat{f}(x)=\hat{A}^{\mathrm{T}}\left(\hat{A}x-b\right)+a^2x+\frac{\hat{A}^{\mathrm{T}}\left(\hat{A}x-b\right)}{\left\|\hat{A}x-b\right\|_2}\|x\|_2+\frac{x}{\|x\|_2}\left\|\hat{A}x-b\right\|_2+\rho\left(x-y_k-v_k+u_k\right) \tag{5.15}$$

得到(5.11)的导数(5.15)，利用 L-BFGS 可以求解出问题(5.11)的最优解 x^*。通过分段讨论的方法，可以求得问题(5.12)的解析解 $S_{\frac{\lambda}{\rho}}(x+u-y)$。由于问题(5.13)是光滑的凸优化问题，易知问题(5.13)存在解析解 $\frac{\rho y+\rho v-\rho u}{\rho+\lambda}$。

6. 数值试验

本节将描述训练数据得到的数值结果，模型在 2.2 GHz Intel Core i7 CPU，16 GB RAM 环境下进行训练。ε^{pri}，ε^{dual} 分别描述了 ADMM 算法对原始条件和对偶条件的可行性容忍程度。本文设置 $\varepsilon^{pri}=10^{-3}$，$\varepsilon^{dual}=10^{-3}$，使用 UCI 的相关数据通过 Python 3.7 实现 ADMM 算法求解鲁棒正则化问题，所使用的数据如表 1：

Table 1. University of California Irvine Repository training data
表 1. UCI Repository 训练数据

数据集	数据量	变量数	数据集	数据量	变量数
Heart	270	13	a6a	32,561	123
pop-failures	540	18	shuttle	43,500	9
Austrian	690	14	a9a	48,842	123
Pima	768	8	w8a	49,749	300
spambase	4601	57	ijcnn1	49,990	22
Mushroom	8124	22	Dota2	102,944	116

本文使用的数据包括 Heart，pop-failures，Austrian，Pima，spambase，Mushroom，a6a，shuttle，a9a，w8a，ijcnn1，Dota2。$U = t\hat{U}$，其中 $\hat{U}$ 是 $\left\|\hat{U}\right\| = 1$ 的随机矩阵。$t \sim U(-a, a)$ 的随机变量，生成了 100 个 U，这说明了存在振幅为 a 的扰动。通过 ADMM 算法求得结果见表 2。

Table 2. Residual expectation
表 2. 残差的期望

数据集	正则项类型	标准正则化	随机鲁棒正则化	最坏情况鲁棒正则化
Heart	L1 正则项	**35.190033**	8.150679	8.575951
	L2 正则项	321.228449	13.847582	8.528732
	L1-L2 正则项	**35.190033**	8.150679	8.526783
	Huber 正则项	57.195922	**7.880510**	**8.525760**
pop-failures	L1 正则项	**36.809833**	10.557134	11.244063
	L2 正则项	120.211912	10.660716	11.239929
	L1-L2 正则项	46.545529	10.520184	11.238676
	Huber 正则项	48.078604	**10.520138**	**11.238669**
Austrian	L1 正则项	**35.652731**	11.117176	12.453968
	L2 正则项	36.074329	11.119549	12.326205
	L1-L2 正则项	36.039073	**11.116835**	12.326204
	Huber 正则项	35.858963	11.116876	**12.325953**
Pima	L1 正则项	37.611661	11.606959	12.429837
	L2 正则项	**36.957709**	11.611515	12.405589
	L1-L2 正则项	40.141959	**11.605968**	12.405529
	Huber 正则项	38.009001	11.606118	**12.405528**
spambase	L1 正则项	**1136.923789**	31.880011	38.235519
	L2 正则项	1267.108529	31.880026	38.218157
	L1-L2 正则项	1175.378456	**31.880006**	**38.218143**
	Huber 正则项	1144.006229	**31.880006**	**38.218143**

Continued

Mushroom	L1 正则项	**151.205724**	32.965843	41.864909
	L2 正则项	527.258221	**32.965482**	38.566990
	L1-L2 正则项	164.807086	32.965658	38.564864
	Huber 正则项	164.807086	32.965644	**38.564858**
a6a	L1 正则项	**81.924991**	66.980143	79.655319
	L2 正则项	122.590718	66.980348	77.994063
	L1-L2 正则项	90.139718	66.979647	77.983832
	Huber 正则项	87.239068	**66.979606**	**77.983747**
shuttle	L1 正则项	533.6717801	181.8549372	201.6537234
	L2 正则项	534.6667871	181.8546774	**187.8732859**
	L1-L2 正则项	534.8815367	181.8540443	187.8732863
	Huber 正则项	**516.6994309**	**181.8540221**	187.8732983
a9a	L1 正则项	**181.780294**	161.873043	187.460652
	L2 正则项	291.347894	161.872388	**186.948356**
	L1-L2 正则项	189.040680	**161.872368**	186.962068
	Huber 正则项	185.614435	161.872399	186.962091
w8a	L1 正则项	60.950881	31.214066	33.686445
	L2 正则项	69.815335	31.190569	**31.449821**
	L1-L2 正则项	64.772441	31.203755	31.467771
	Huber 正则项	**33.222972**	**31.190307**	31.467499
ijcnn1	L1 正则项	265.710981	131.311997	204.534219
	L2 正则项	371.969304	**131.311828**	**153.685551**
	L1-L2 正则项	275.206242	131.311971	153.699805
	Huber 正则项	**159.980205**	131.311916	153.699895
Dota2	L1 正则项	**1440.002173**	**320.331273**	**320.848874**
	L2 正则项	2968.972226	**320.331273**	**320.848874**
	L1-L2 正则项	1442.382143	**320.331273**	**320.848874**
	Huber 正则项	1442.032602	**320.331273**	**320.848874**

通过表 2，可以发现当数据存在扰动时，标准的正则化方法已经不能够很好的解决分类问题，而鲁棒正则化模型能够较好地应对这种情况。由于 L1 正则项会压缩回归系数，所以 L1 正则化之后会导致分类准确率降低，残差增大。但是通过表 2 发现标准的 L1 正则化训练存在扰动的训练样本时表现竟然好于其它正则项。由此可以看出训练样本存在扰动时会破坏正则项一些较好的性质，让分类结果不尽人意。相比于标准的正则化模型，本文提出的鲁棒正则化模型不会破坏正则项的性质。通过表 2，不难发现鲁棒 L1-L2 正则化、Huber 正则化表现要优于 L1 正则化和 L2 正则化。另外，通过表 2 可以得到一个结论：随机鲁棒正则化的分类质量要优于最坏情况鲁棒正则化。导致这个结果的原因是两个模型考虑的因素不同，主要体现在随机鲁棒正则化模型的保真项是让残差的期望最小，而最坏情况鲁棒正则化的保真项是让最坏误差最小，所以随机鲁棒正则化残差的期望最小，分类结果最好。

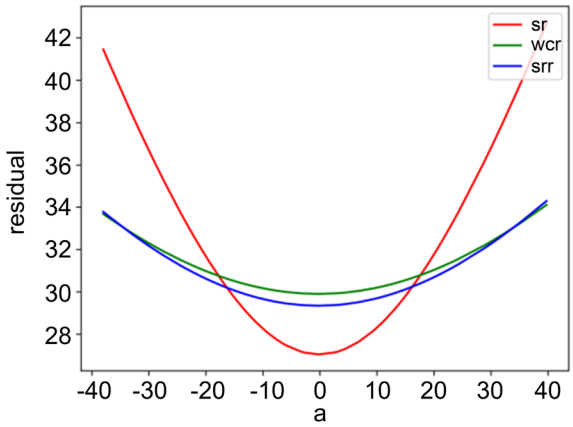

Figure 2. Training results of w8a Huber regularization
图 2. w8a Huber 正则化训练结果

图 2 中 sr 表示的是标准正则化模型得到的残差曲线，srr 表示的是随机鲁棒正则化模型得到的残差曲线，wcr 表示最坏情况鲁棒正则化模型得到的曲线。通过图 2，可以发现当扰动的振幅 $a = 0$ 时，标准正则化的残差最小。随着振幅 a 的增大，标准正则化得到的残差明显增大不少，但是随机鲁棒正则化和最坏情况鲁棒正则化模型的残差增长速度缓慢。这意味着随机鲁棒正则化模型和最坏情况鲁棒正则化模型的抗干扰能力比标准正则化模型要强。另外由于随机鲁棒正则化考虑的是残差的期望最小，所以在图 2 中随机鲁棒正则化得到的残差之和要小于最坏情况鲁棒正则化。但是当振幅 a 增大到一定程度时，最坏鲁棒正则化的残差会小于随机鲁棒正则化。方差可以表示数据的波动情况，表 3 给出了残差的方差。

Table 3. Variance of residuals
表 3. 残差的方差

数据集	正则项类型	标准正则化	随机鲁棒正则化	最坏情况鲁棒正则化
Heart	L1 正则项	444.295317	0.063045	**0.003982**
	L2 正则项	41424.888749	1.895289	**0.004407**
	L1-L2 正则项	444.295317	0.063045	**0.004435**
	Huber 正则项	1275.871281	0.043919	**0.004446**
pop	L1 正则项	416.833961	0.002358	**0.000236**
	L2 正则项	4972.827751	0.003744	**0.000241**
	L1-L2 正则项	701.468900	0.004525	**0.000242**
	Huber 正则项	751.418061	0.004521	**0.000242**
Austrian	L1 正则项	426.470664	0.188475	**0.003366**
	L2 正则项	438.250519	0.199743	**0.004387**
	L1-L2 正则项	437.809880	0.189704	**0.004387**
	Huber 正则项	432.563523	0.189184	**0.004391**
Pima	L1 正则项	359.405084	0.043336	**0.000547**
	L2 正则项	343.317120	0.058013	**0.000561**
	L1-L2 正则项	424.247165	0.045426	**0.000561**
	Huber 正则项	369.253361	0.045025	**0.000561**

Continued

spambase	L1 正则项	400399.963706	0.000764	**0.000683**
	L2 正则项	497672.069782	0.000766	**0.000693**
	L1-L2 正则项	428035.522928	0.000763	**0.000693**
	Huber 正则项	405420.896083	0.000763	**0.000693**
Mushroom	L1 正则项	8058.194044	4.123995	**0.297610**
	L2 正则项	104363.766444	4.134875	**1.105258**
	L1-L2 正则项	9662.772588	4.130733	**1.106304**
	Huber 正则项	7683.984106	4.130237	**1.106307**
a6a	L1 正则项	317.112070	8.185439	**0.007487**
	L2 正则项	2287.702348	8.223619	**0.017613**
	L1-L2 正则项	582.375799	8.224333	**0.017709**
	Huber 正则项	480.564331	8.223786	**0.017710**
shuttle	L1 正则项	69928.263642	289.713089	**26.161685**
	L2 正则项	70258.163243	290.135024	**80.918441**
	L1-L2 正则项	70329.080826	289.776428	**80.918437**
	Huber 正则项	63750.404986	289.763368	**80.918332**
a9a	L1 正则项	632.181843	18.334536	**0.000038**
	L2 正则项	11450.332269	18.363809	**0.000071**
	L1-L2 正则项	966.965557	18.370788	**0.000070**
	Huber 正则项	966.965557	18.367920	**0.000070**
w8a	L1 正则项	717.507328	1.811562	**0.152904**
	L2 正则项	1130.109044	2.290337	**2.010488**
	L1-L2 正则项	861.095374	1.955220	**1.798560**
	Huber 正则项	25.964530	2.495772	**1.801926**
ijcnn1	L1 正则项	11860.659302	0.757916	**0.001887**
	L2 正则项	31148.424444	**0.766460**	1.046655
	L1-L2 正则项	13244.481020	**0.760788**	1.043701
	Huber 正则项	1004.144961	**0.762319**	1.043686
Dota2	L1 正则项	531759.733515	0.001676	**6.95E−23**
	L2 正则项	531759.733515	0.001677	**2.74E−25**
	L1-L2 正则项	533819.513567	0.001676	**6.14E−27**
	Huber 正则项	533819.513567	0.001676	**2.58E−25**

通过表 3 的数据，可以发现标准的正则化方法得到残差的方差很大，鲁棒正则化模型得到残差的方差较小。表 3 说明标准的正则化方法得到的结果受到扰动的影响较大，在一些数据集上分类的波动是不能够被接受的。相反，鲁棒正则化模型得到的结果受到扰动的影响较小。由于最坏情况鲁棒逼近考虑的是优化最坏误差，所以收到扰动的影响最小，在一些样本集上训练结果波动几乎为 0。随机鲁棒正则化模型只考虑残差的期望最小，所以受到扰动的影响介于标准正则化和最坏情况鲁棒正则化之间。

7. 总结

当训练样本存在扰动时，本文提出随机鲁棒正则化模型和最坏情况鲁棒正则化模型。利用 ADMM 算法求解上述模型，并将鲁棒正则化模型与标准正则化模型进行比较。通过比较发现本文提出的算法具有比标准正则化模型更好的抗干扰能力。数值实验显示：1) 本文提出的模型能够成功对抗存在扰动的训练样本；2) L1-L2 正则项和 Huber 正则项表现优于 L1 正则项和 L2 正则项。

基金项目

国家自然科学基金青年科学基金项目：11701279。

参考文献

[1] Boyd, S. and Vandenberghe, L. (2006) Convex Optimization. Cambridge University Press, Cambridge.

[2] 陈威, 艾婵. 基于多元线性回归模型的武汉市水资源承载力研究[J]. 河南理工大学学报(自然科学版), 2017, 36(1): 75-79.

[3] 薛斌, 程超, 欧世其, 等. 考虑舒适温度区间和突变量的月售电量预测线性回归模型[J]. 电力系统保护与控制, 2017, 45(1): 15-20.

[4] 田秀芹. 基于多元线性回归的粮食产量预测[J]. 科技创新与应用, 2017(16): 3-4.

[5] Tibshirani, R. (1996) Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Methodological*), **58**, 267-288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

[6] Tibshirani, R. (2005) Sparsity and Smoothness via the Fused Lasso. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **67**, 91-108. https://doi.org/10.1111/j.1467-9868.2005.00490.x

[7] Hui, Z. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **67**, 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

[8] Tikhonov, A.N. (1977) Solution of Ill-Posed Problems. Winston, Washington DC.

[9] Zhang, C.-H. (2010) Nearly Unbiased Variable Selection under Minimax Concave Penalty. *The Annals of Statistics*, **38**, 894-942. https://doi.org/10.1214/09-AOS729

[10] Zhang, H.H., Ahn, J., Lin, X., *et al*. (2006) Gene Selection Using Support Vector Machines with Non-Convex Penalty. *Bioinformatics*, **22**, 88-95. https://doi.org/10.1093/bioinformatics/bti736

[11] Frank, L.L.E. and Fridman, J.H. (1993) A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-135. https://doi.org/10.1080/00401706.1993.10485033

[12] Hastie, T., Tibshirani, R. and Wainwright, M. (2016) Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press, Boca Raton, FL. https://doi.org/10.1201/b18401

[13] 徐征, 刘遵雄, 张贤龙. 基于套索(Lasso)的中文垃圾邮件过滤[J]. 华东交通大学学报, 2014(4): 130-135.

[14] 孙昕. 基于 Lasso 方法的中国股市时滞性回归分析[D]: [硕士学位论文]. 大连: 大连理工大学, 2017.

[15] 谭勇, 谢林柏, 冯宏伟, 等. 基于 LASSO 回归的红外火焰探测器的设计与实现[J]. 激光与红外, 2019, 49(6): 720-724.

[16] 高魏, 闵捷, 张安录. 基于岭回归的农地城市流转影响因素分析[J]. 中国土地科学, 2007, 21(3): 51-58.

[17] 罗良文, 阚大学. 对外贸易和外商直接投资对中国人力资本存量影响的实证研究——基于岭回归分析法[J]. 世界经济研究, 2011(4): 33-37.

[18] Boyd, S., Parikh, N., Chu, E., *et al*. (2010) Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends in Machine Learning*, **3**, 1-122. https://doi.org/10.1561/2200000016

[19] Cai, X.J., Han, D.R. and Yuan, X.M. (2017) On the Convergence of the Direct Extension of ADMM for Three-Block Separable Convex Minimization Models with One Strongly Convex Function. *Computational Optimization and Applications*, **66**, 39-73. https://doi.org/10.1007/s10589-016-9860-y