1 对论文中的 lasso 做出自己的理解

1.1 引言

线性回归模型一般只对低维度的数据适用，即$n > p$的情况，并且特征变量之间还不能存在相关程度高，即多元共线的关系。而为了解决多元共线的问题，提出了岭回归，其在线性回归模型的基础上加上了l_2范数作为罚项，虽然l_2范数在计算时相对方便一些，但并不能实现将某些参数收缩至 0，换言之，不能实现变量选择。

Lasso 是针对岭回归不能解决变量选择的缺陷提出的，其全称是 Least Absolute Selection and Shrinkage Operator,即最小绝对值选择与收缩算子。

1.2 相关先修知识

在精读 Lasso 论文的过程中，我遇到了几个不甚熟悉的概念，于是暂时暂停了论文的精读，转而去搜寻这些概念的含义及其应用，掌握必要知识背景后，才能更好地理解 Lasso 论文所要表达的知识要义。

1.2.1 Ordinary Least Square(OLS)

OLS 表示普通最小二乘，作为一种常见的数学优化方法，其核心思想是通过对残差平方和的最小化来进行估计。

假设数据样本 x、y 之间呈线性关系，即期望采用一元线性回归模型

$$\boldsymbol{y = ax + b}$$

来确定 X、Y 之间的具体函数关系。

$$\begin{cases} a + b = 1 \\ 2a + b = 5 \\ 3a + b = 6 \end{cases}$$

目的是期望通过确定参数 a、b,进而确定 x、y 之间的函数关系。但是显然无法直接求解该方程组。因此只能转换一个思路：找到一条直线，使其与各个样本点之间的残差和最小。虽然这条直线不能保证经过全部的样本点，但一般能够很好地描述出 x、y 之间的关系。由于残差的正负问题需要添加绝对值，然而在实际计算中绝对值不便处理，故直接对残差进行平方来保证非负性。

$$S = \sum_{i=1}^{n} (\hat{y_i} - y_i)^2$$

S 表示真实值和估计值之间误差的平方和，则当 S 最小时，参数 a、b 即为所求值，即

$$\arg\min_{a,b}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2=\arg\min_{a,b}\sum_{i=1}^{3}(ax_i+b-y_i)^2$$

根据样本数据有

$$S=(a+b-1)^2+(2a+b-5)^2+(3a+b-6)^2$$

根据极值理论可知，当 S 取极值时，S 对 a 和 b 的偏导均为 0，故有

$$\begin{cases}\frac{\partial S}{\partial a}=2(a+b-1)+2\cdot 2(2a+b-5)+2\cdot 3(3a+b-6)=0\\ \frac{\partial S}{\partial b}=2(a+b-1)+2(2a+b-5)+2(3a+b-6)=0\end{cases}$$

根据上式，可直接解出参数 a、b

$$\begin{cases}a=2.5\\ b=-1\end{cases}$$

1.2.2 范数

范数的本质是距离，存在的意义是为了实现比较。例如，在一维实数集合中，随机取两点：6 和 7，显然 7 大于 6。但在二维实数空间中，随机取两点(1，1)和(0，4)，此时无法直观地比较两点的大小，因为它们不是可以比较的实数，于是引入范数这个概念，将(1，1)和(0，4)通过范数分别映射到实数$\sqrt{2}$和 4，这样就能够直接比较两点的大小。因此范数其实是一个函数，将不能比较的向量转换为可以比较的实数。

l_1范数的定义

根据l_p范数的定义，当$p=1$时，任意向量x的l_1范数为

$$||\boldsymbol{x}||_1=\sum_i|x_i|$$

等于向量中所有元素绝对值之和。

l_2范数的定义

l_2范数表示向量或矩阵的元素平方和开根号，即

$$\|\boldsymbol{x}\|_2=\sqrt{\sum_i x_i^2}$$

l_1范数：正则项与稀疏解

在机器学习的诸多方法中，假设给定了一个较小的数据集来做训练，往往会遇到过拟合的情况，即训练得到的模型可能将数据中隐含的噪声和毫无关系的特征也表征出来。

为了避免类似的过拟合问题，一种解决方法是在机器学习模型的损失函数中加入正则项，例如用l_1范数表示的正则项，只要使得l_1范数的数值尽可能变小，就能够让我们期望的解变成一个稀疏解，即解的很多元素为0。

如果我们想解决的优化问题是损失函数$f(x)$最小化，那么，考虑由$\boldsymbol{l_1}$**范数**构成的正则项后，优化目标就变为

$$\min f(\boldsymbol{x}) + ||\boldsymbol{x}||_1$$

只要优化模型的解x的l_1范数比较小，那么这个解就是稀疏解，并且稀疏解可以避免过拟合。其中，“稀疏”一词可以理解为x中的大多数元素都是 0，只有少量的元素是非 0 的。

1.3 Lasso 详细介绍

1.3.1 Lasso 的稀疏性假设

尽管世界如此复杂，但有用的信息却非常有限。套用到常见的统计学模型中，例如考虑一个线性回归模型，存在一个因变量y，但却有不计其数的自变量x，在此基础上假设仅存在有限个x的回归系数不为0，其余的全为0，即这些回归系数为0的自变量与y并不存在显著的关系。找到其中重要的x，对我们理解数据有重要的意义。

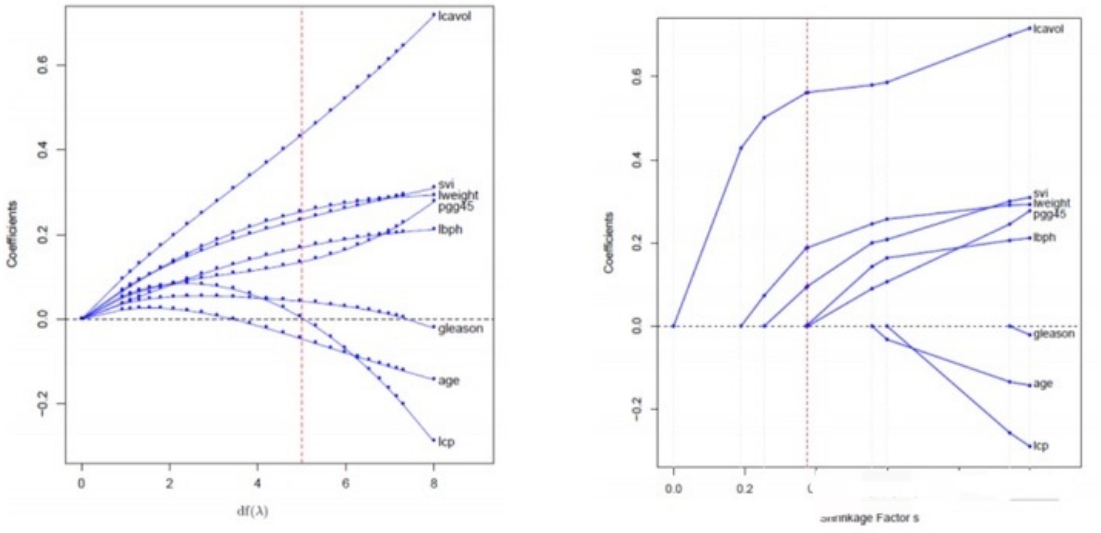
例如生物学家想要研究基因对于某类疾病的影响，面对上万个可能的基因，生物学家们倾向于认为只有其中的一小部分对于该类疾病有着显著的影响。在分析基于和疾病的关系时，对于放入的 10000 个可能的基因，我们认为这 10000 个基因在回归模型中的回归系数只有少量的不为0。

1.3.2 Lasso 原理

与岭回归类似，从模型上看，Lasso 无外乎是加入了l_1范数惩罚项的优化问题；把拉格朗日对偶形式改为了

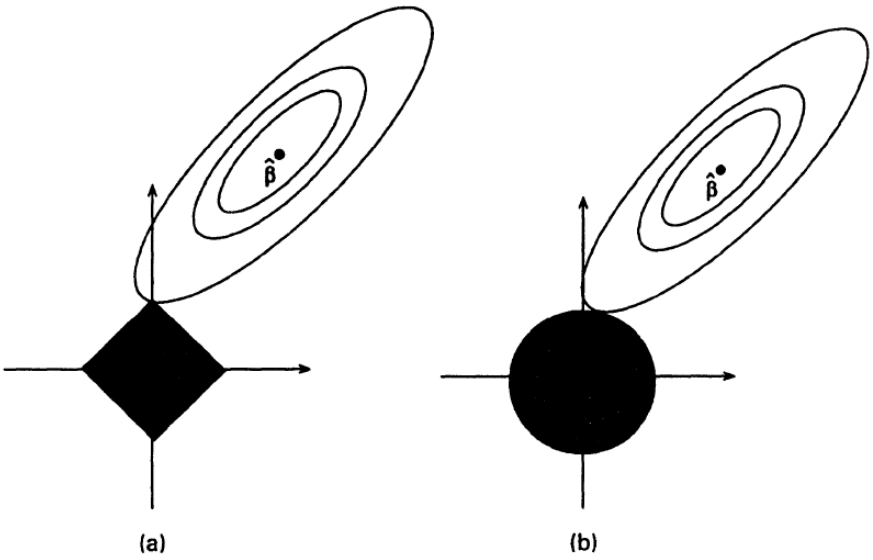
$$\min_{\beta_0,\beta}\left\{\frac{1}{2N}\|y-\beta_0 1-X\beta\|_1\right\},\|\beta\|_1 \le t$$

将它得到的解称为 Lasso 估计。它对应的原始的优化形式中的λ称为收缩算子。观察岭回归和 Lasso 的实验结果（坐标从右往左看）



两者都可以让回归系数慢慢的趋于 0，但 Lasso 回归针对不同的自变量，会使其收敛的速度不一样：有的变量很快趋于 0，有的却速度较慢。因此一定程度上 LASSO 回归非常适合于做变量选择。相比之下，看左侧岭回归的图能够观察到，虽然它也会有这个趋势，但是严格趋于 0 时一定是要求它的λ值要非常大。

实际上，Lasso 回归和岭回归在线性回归模型上增加的罚项，可以看作是对参数取值的限制。在岭回归中，这个限制是$\|\beta\|_2^2 \leq t$，直观理解下，其边界是一个圆；而在 Lasso 中，限制为$\|\beta\|_1 \leq t$，边界则是一个正菱形。由于目标函数也是二次的，我们可以划圆形等高线表示参数β的取值，在同一个等高线上的β都使目标函数取值相同。



而我们都希望损失函数的值越小越好，因此参数β的取值尽量满足在越小的同心圆等高线越好，又由于β取值的限制，可知取同心圆等高线与限制区域相切的位置即为最优化的解。

1.3.3 Lasso 为什么能进行变量选择

观察上图可知，岭回归的圆是光滑的，很容易做到相切，但不容易做到使得回归系数值取 0 的时候相切。相较而言，Lasso 回归的正菱形区域有顶点，因此更容易在顶点处相切，此时某个变量取值为 0。

这也说明了 Lasso 回归能够使某些自变量的值缩减为 0，即进行了变量的选择，在高维度数据的问题上受到广泛的应用。

1.3.4 Lasso 与岭回归的对比

lasso 回归相比于岭回归，会比较极端。它不仅可以解决过拟合问题，而且可以在参数缩减过程中，将一些重复的没必要的参数直接缩减为零，也就是完全减掉了，达到提取有用特征的作用。但毕竟l_1范数不是连续可导的，因此 lasso 回归的计算过程复杂，这也是 Lasso 回归待优化的一个问题。

1.4 Lasso 最优解

由于 Lasso 的损失函数不是连续可导的，因此无法使用梯度下降法对其求解，但可以使用坐标轴下降法和最小角回归法。

1.5 历史角度看 Lasso

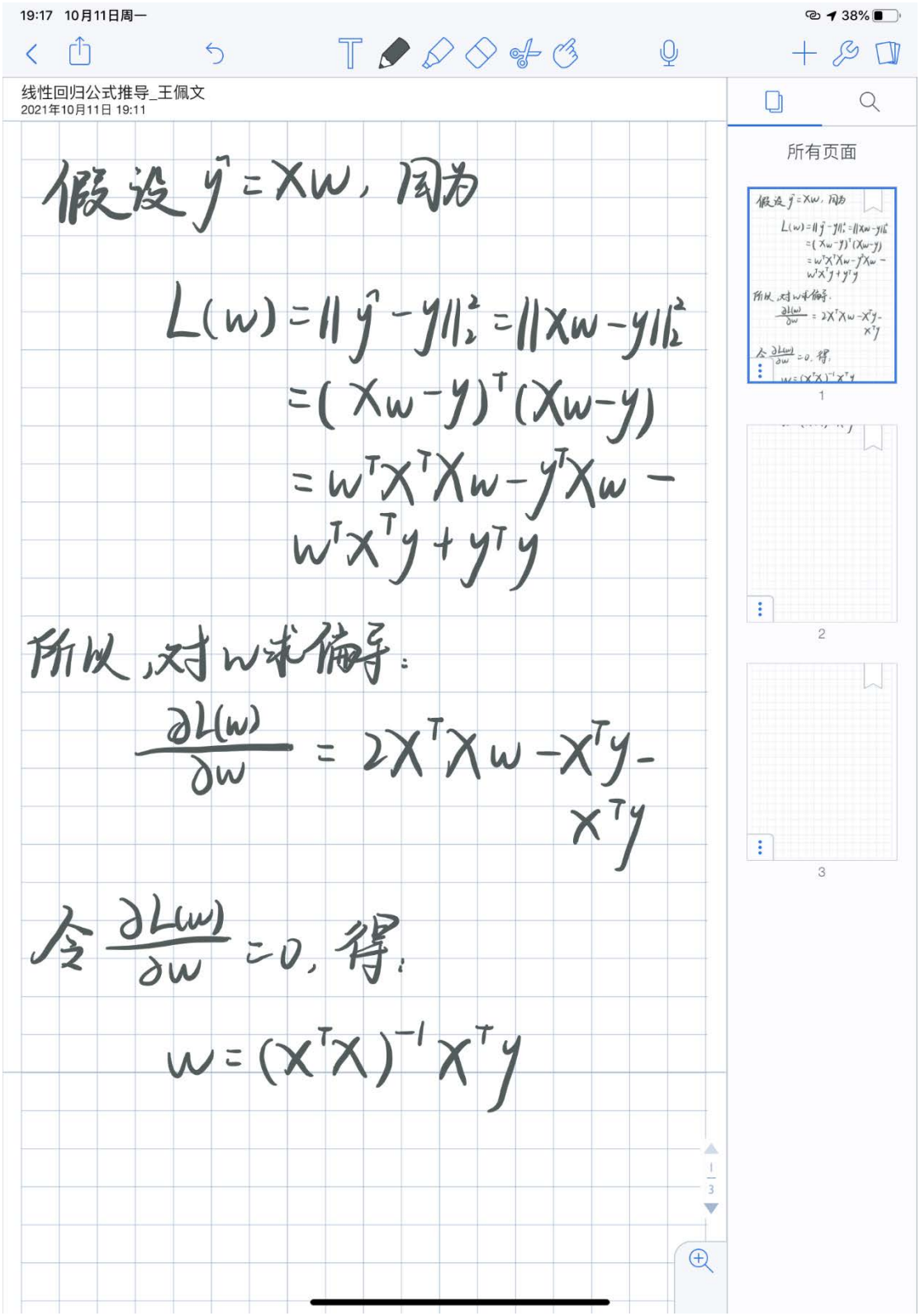
通过历史的角度看统计学方法，更有利于我们理解 Lasso 提出的意义。个人认为 Lasso 确实是基于了一些前人的工作，但这篇论文的重要性更多的是在统计学科建设上。

Lasso 方法最早由 Robert Tibshiran 于 1996 年提出，论文发表在“统计四大”之一的皇家统计学会期刊上。Lasso 出现以前，统计领域的文章都是预测和解释两方面并行。Lasso 论文发表后，人们才发现可以通过牺牲一定的偏差，而交换更高的均方误差，打破了传统的统计学一向的思维定式。

2 对论文中的重要过程的推导

2.1 线性回归的公式推导

由于 Lasso 是在线性回归模型的基础上加上了l_1范数，因此首先对线性回归进行公式推导，推导如下：



2.2 Lasso 回归的公式推导

Lasso 回归在线性回归的经验风险的基础上加上表示模型复杂度的正则化项l_1范数，所以线性回归是经验风险最小化，Lasso 回归是结构风险最小化。

Lasso回归公式推导_王佩文
2021年10月11日 19:25

Lasso回归的原理

多元线性回归:

$$\beta^{\rightarrow} = \arg\min_{\beta^{\rightarrow}} \sum_{i=1}^{n} (y_i - X_i'\beta^{\rightarrow})^2, \text{其中:}$$

$$\beta^{\rightarrow} = (\beta_1^{\rightarrow}, \beta_2^{\rightarrow} \cdots \beta_n^{\rightarrow})$$

岭回归:

$$\beta^{\rightarrow} = \arg\min_{\beta^{\rightarrow}} \left[\sum_{i=1}^{n} (y_i - X_i'\beta^{\rightarrow})^2 + \lambda \left(\sum_{i=1}^{n} \beta_i^2 \right) \right]$$

(λ为正常数)

Lasso回归:

$$\beta^{\rightarrow} = \arg\min_{\beta^{\rightarrow}} \left[\sum_{i=1}^{n} (y_i - X_i'\beta^{\rightarrow})^2 + \lambda \sum_{i=1}^{n} |\beta_i| \right]$$

记 $L = (y - X\beta^{\rightarrow})'(y - X\beta^{\rightarrow}) + \lambda |\beta^{\rightarrow}|$

为 $\dfrac{\quad}{\quad}$

易知：$\lambda \to 0$ 时，Lasso 回归与多元线性回归完全相同；

$\lambda \to \infty$ 时，$\beta^{(1)} = 0_{(p\times 1)}$

另外：$\dfrac{\partial L}{\partial \beta^{(1)}} = \begin{cases} -2X'y + 2X'X\beta^{(1)} + \lambda = 0, & \beta^{(1)} > 0 \text{ (i)} \\ -2X'y + 2X'X\beta^{(1)} - \lambda = 0, & \beta^{(1)} < 0 \text{ (ii)} \end{cases}$

(1) 对于 $\beta^{(1)} > 0$ 时，

$$\frac{\partial L_i}{\partial \beta_i} = -\beta_i^{(1)} + \beta_i + \lambda = 0$$

① 对于 $\beta^? > 0$ 时，

$$\frac{\partial L_i}{\partial \beta_i} = -\beta_i^? + \beta_i + \lambda = 0$$

$$\beta_i = \beta_i^? - \lambda,$$

由于 $\beta_i > 0$，因此当且仅当 $\beta_i^? - \lambda$ 非负时成立.

② 对于 $\beta^? < 0$ 时，

$$\frac{\partial L_i}{\partial \beta_i} = -\beta_i^? + \beta_i - \lambda = 0$$

$$\beta_i = \beta_i^? + \lambda$$

---

3 相关进一步引用该论文并作出了有效改进的追踪（两个）

3.1 Lasso 的缺陷

虽然 Lasso 回归具有许多优势，能够实现变量选择，但它同时也存在一些缺陷。例如：变量选择时会出现不一致性；急于将某些参数缩减为 0，因此可能会导致过程中产生一些错误，使得最终的模型偏差较大；并且当 n 很小时至多只能选出 n 个变量；而且不能做群体选择。

3.2 Elastic Net

3.2.1 Elastic Net 介绍

针对 Lasso 算法急于将某些参数缩减为 0，这其中可能会导致一些相关性较强的特征参数也化为 0，导致特征丢失，进而导致最终的模型偏差较大的问题，因此提出 Elastic Net，其中文名称是弹性网络。

Elastic Net 是一种线性模型，可以看作是对岭回归和 Lasso 回归的折中，因为它在目标函数里同时使用了l_1范数和l_2范数作为惩罚项。这样的组合既与 Lasso 类似，学习了一个稀疏的模型，同时也保持了岭回归的正则属性。其功能也是解决模型训练过程中的过拟合问题。其目标函数是

$$J(\theta)=MSE(y,\hat{y};\theta)+r\alpha\sum_{i=1}^{n}|\theta_i|+\frac{1-r}{2}\alpha\sum_{i=1}^{n}\theta_i^2$$

3.2.2 变量选择

在变量选择方面，当多个特征和另一特征相关时，Elastic Net 非常有用。Lasso 回归倾向于随机选择其中一个，而 Elastic Net 更倾向于全部选择，其对所选变量的数量没有限制。

因此 Elastic Net 既做到了岭回归没有实现的变量选择问题，又能够解决 Lasso 回归使一些相关性较强的重要特征参数也缩减为 0 的问题。

3.2.3 实际应用

在实际问题中，在进行正则化的过程中，通常要先使用岭回归，原因是岭回归计算更加精准，但并不具有变量选择的功能。

因此当特征变量的数量非常多时，应考虑使用 Elastic Net，原因是 Elastic Net 结合了岭回归的计算的优点，同时又结合了 Lasso 回归特征选择的优势。

3.3 Adaptive Lasso

3.3.1 Adaptive Lasso 介绍

针对 Lasso 回归在变量选择时会出现不一致性的缺陷，Hui Zou 于 2006 年在《The Adaptive Lasso and Its Oracle Properties》中提出了 Adaptive Lasso，翻译为中文是自适应 Lasso。

Adaptive Lasso 是一个两阶段的方法。在第一步中，得到了一个初始估计量。然后，必须求解一个带有加权惩罚的惩罚优化问题。初始估计量不需要是一致的，但它应该是零一致的。在偏正交条件下，由一元回归得到一个简单的零相容初值估计。

与 Lasso 相比，Adaptive Lasso 的理论优势在于它具有 oracle 属性。与具有 oracle 特性的 SCAD 和桥接方法相比,Adaptive Lasso 的优点是计算效率高。在给定初始估计量的情况下，Adaptive Lasso 估计的计算是一个凸优化问题，其计算代价与 Lasso 相同。实际上，Adaptive Lasso 的整个正则化路径可以用与 LARS 算法的最小二乘解相同的计算复杂度来计算。因此，Adaptive Lasso 是分析高维数据的有效方法。

3.3.2 Adaptive Lasso 原理

Lasso 回归的损失函数加上了l_1范数，能够进行变量选择，但在变量选择时会出现不一致的情况，而 Adaptive Lasso 针对这一点在 Lasso 的基础上进行了相关的改进，具体体现在基于 Lasso 增加了一个权重ω，因此其损失函数为

$$||y - \boldsymbol{X}\beta||^2 + \lambda \sum_{i=1}^{p} w_j|\beta_j|$$

权重的选择可以依赖数据驱动，通过交叉验证的方式得到。