高维数据惩罚回归算法

我们在实务中运用机器学习时，经常会遇到样本示例数据集中特征变量很多的情形，比如我们前面所述的商业银行对公客户违约问题，特征变量可能涉及客户的经营能力、盈利能力、偿债能力、发展潜力、现有负债及担保情况等方方面面，即便是对于从业经验丰富、专业水平极高的银行从业人员，在选取时也会不可避免的创建较多的特征变量，而且对公客户不比个人客户，样本数量相对较少，这时候就会产生高维数据及其带来的“维度灾难”问题。本章介绍的高维数据惩罚回归算法是解决这一问题的有效方法之一，包括岭回归、Lasso 回归和弹性网回归。

目录

CONTENTS

1 高维数据惩罚回归算法的基本原理
2 数据准备
3 变量设置及数据处理
4 岭回归算法
5 Lasso回归算法
6 弹性网回归算法
7 小结
8 习 题

PART 01

高维数据惩罚回归算法的基本原理

高维数据惩罚回归算法的基本原理

前面我们在第三章中讲到，开展机器学习时，经常会在训练样本时出现“过拟合”的现象导致模型的“泛化”能力下降，这一点对于高维数据体现的尤为明显。所谓高维数据是指样本示例数据集中的特征变量很多，维度就是特征变量的个数。高维数据会导致“维度灾难(Curse of Dimensionality)”的问题。“维度灾难”是指随着特征变量个数的增加，为了达到相同的预测效果，所需要的样本示例个数会呈指数型增加，而如果受条件限制，样本个数无法增加或增加不够充分，那么就会导致模型的预测能力下降。“维度灾难”的发生，在本质上是由于特征变量过多，导致对训练样本的拟合过于充分，虽然显著增强了对训练样本的拟合能力，但同时也引入了过多的“噪声”信息或随机性特征，而这些特征并没有出现在新数据或未来的数据中，从而导致模型的“泛化”能力不足。

高维数据惩罚回归算法的基本原理

所谓“大道至简”，奥卡姆剃刀定律指出，一个满足预测性能条件下尽量简单的模型，才能够有比较好的泛化能力。这一定律本质也反映了偏差和方差的权衡，特征变量过多或者模型过于复杂，偏差虽然会变小，但方差就会变大，为使得整体的均方误差变小，需要在偏差和方差之间找到一种平衡。

针对前述“维度灾难”问题，一般来说有三种方法可供选择：增加样本示例全集容量、降低维度、正则化等方法来解决。关于增加样本示例全集容量的方法很好理解，如果能够突破条件限制，找到更多合适的样本可供研究，显然是一种非常好的选择；关于降维方法，常用主成分分析，这一方法将在后面章节中详解。

高维数据惩罚回归算法的基本原理

本章主要讲解正则化的方法，或者称作惩罚回归算法。正则化都可以看做是损失函数的惩罚项，惩罚项的大小随着特征数量的增加而增加，惩罚回归的原理是，回归系数的选择应使残差平方和与惩罚项之和最小。所以在引入特征时，特征对模型拟合或预测的贡献必须足以抵消引入其带来的惩罚项的增加时，才会被保留下来，从而提高了引入特征变量的成本或者说应该以一种更加审慎的态度看待特征引入工作。正则化的具体方法包括岭回归、Lasso 回归和弹性网回归（Elastic Nets）。

岭回归

在特征变量之间存在严格多重共线性时，会有多个 OLS 估计量能够使得残差平方和最小且可决系数为 1，为了得到唯一回归系数，则需对系数的取值范围进行限制，增加惩罚项（正则项）。模型拥有的自由度越低，就越不容易过度拟合数据。岭回归使用 L2 正则化方法，引入惩罚项$\gamma_2||\beta||_2^2$，**该惩罚项基于回归系数大小的平方**。在岭回归方法下，回归系数β不再仅仅使得残差平方和$(Y-X\beta)^T(Y-X\beta)$最小，而是要**使得残差平方和与惩罚项之和最小**，其数学公式为：

$$\hat{\beta}^{\text{ridge}} = \text{argmin}_\beta((\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)+\gamma_2\|\beta\|_2^2)$$

其中的γ_2为调节系数，用来控制正则化（惩罚）的力度，取值范围为大于等于 0，惩罚的力度越大，就越不容易过度拟合数据。如果γ_2=0，则岭回归就是线性模型。如果γ_2非常大，则所有特征向量的系数都将无限接近于零。从“方差-偏差”的角度看，γ_2提升了模型的偏差，但是显著降低了方差，恰当的γ_2值将实现更优平衡。

由于岭回归对输入特征的大小非常敏感，所以在执行岭回归之前，需要对数据进行标准化处理（在 pyhton 中常常使用 StandardScaler()函数）。当然大多数正则化模型都应该做类似处理。

Lasso回归

Lasso 回归，又称最小绝对收缩和选择算子回归（Least Absolute Shrinkage and Selection Operator Regression，也称套索回归）的基本原理与岭回归基本相同，区别在于使用L1 正则化方法，引入惩罚项$\gamma_1\|\beta\|_1^{\square}$，**该惩罚项基于回归系数大小的绝对值**。在 Lasso 方法下，回归系数β 同样是**使得残差平方和与惩罚项之和最小**，其数学公式为：

$$\hat{\beta}^{lasso} = \mathrm{argmin}_{\beta}((\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)+\gamma_1\|\beta\|_1)$$

其中的γ_1为调节系数，用来控制正则化（惩罚）的力度，取值范围为大于等于 0。在最优调节系数选择时，常使用 K 折交叉验证方法（CV），通过实现交叉验证误差最小化的方式来获取。

弹性网回归

弹性网回归的基本原理与岭回归、Lasso 回归同样基本相同，区别在于同时使用 L2 正则化方法和 L1 正则化方法，其正则项是岭回归和 Lasso 回归的正则项的混合，引入的惩罚项为$\alpha\gamma_1||\beta||_1^{\cdot} + (1-\alpha)\gamma_2||\beta||_2^2$，实质上是将单一的惩罚项按照一定的权重分别分配给了 L2 正则化、L1 正则化，同时兼顾了两种正则化方法。其数学公式为：

$$\hat{\beta}^{el.net} = \mathrm{argmin}_{\beta}((\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta) + \alpha\gamma_1\|\beta\|_1 + (1-\alpha)\gamma_2\|\beta\|_2^2)$$

其中的α为权重分配比例，当$\alpha=0$时，弹性网络即等同于岭回归，而当$\alpha=1$时，即相当于 Lasso 回归；γ_1、γ_2分别为 L1 正则化、L2 正则化的调节系数。

惩罚回归算法的选择

岭回归只是收缩回归系数而不进行变量筛选，而 Lasso 回归的特色在于具有变量选择功能和变量空间降维功能，或者说具有自动执行特征选择并产生稀疏模型的能力，通过将不重要特征的权重设置为零实现大量冗余变量的去除，只保留与响应变量最相关的特征变量，在有效简化模型的同时可以保留数据集中最重要的信息。所以如果我们依据经济社会相关理论文献或者实践经验能够合理判断真正显著影响响应变量的特征变量较少，也就是说真实的模型本来就是稀疏模型，那么就应该倾向于选择 Lasso 回归算法，否则就应该选择岭回归算法。

如果特征数量显著超过训练样本集中样本示例数量，或者多个特征变量之间存在强相关时，Lasso 回归的预测表现可能不够稳定，那么使用弹性网回归这种折中的算法就更为合适，因为弹性网回归会倾向于将这些强相关的特征变量都选上。而且从前面的公式可以看出，当 $\alpha = 0$ 时，弹性网络即等同于岭回归，而当 $\alpha = 1$ 时，即相当于 Lasso 回归，所以弹性网回归算法事实上已经包含了岭回归、Lasso 回归两种算法，计算边界更广，预测能力更强一些。

PART 02

数据准备

数据准备

我们用一个手机游戏玩家体验评价影响因素建模实例来进行讲解。以手机玩家的体验评价得分（Y）为响应变量，构建起包括游戏流行程度（X_1）、游戏资源要求（X_2）、游戏花费成本（X_3）、游戏具体内容（X_4）和游戏广告植入（X_5）五个大方面特征变量的手机游戏玩家体验评价影响因素理论模型。

$$Y = f(X_1, X_2, X_3, X_4, X_5)$$

变量	子变量
□琚□□□掊	□琚□另掊着□粥殆□□雏圊民 popularity1
	□粥竦□着□粥殆□□雏圊民 popularity2
□琚□□□□	□琚着□锄□□着□粥殆□□雏圊民 resources1
	□琚着□□□□着□粥殆□□雏圊民 resources2
	□琚着□□□□着□粥殆□□雏圊民 resources3
□琚□□瑁貔	□琚□□□□着□粥殆□□雏圊民 spend1
	□琚蜉□□□着□粥殆□□雏圊民 spend2
	□琚□襟□□着□粥殆□□雏圊民 spend3
□琚洪殆瑚帚	□琚□□着□粥殆□□雏圊民 content1
	□琚镂炻着□粥殆□□雏圊民 content2
	□琚□迈忸着□粥殆□□雏圊民 content3
□琚蘧绿□鹤	卤劲□□蘧绿着□粥殆□□雏圊民 advertisement1
	□琚表蘧绿圊民着□粥殆□□雏圊民 advertisement2

数据准备

然后通过调查问卷的形式获取数据，最终形成的数据文件为“数据9.1”。“数据9.1”中设置了14个变量，即“V1-V14”分别用来表示玩家体验评价appraise、游戏知名度影响popularity1、玩家数量影响popularity2、游戏对硬件要求影响resources1、游戏对网速要求影响resources2、游戏对流量要求影响resources3、游戏金钱花费spend1、游戏时间花费spend2、游戏脑力花费spend3、游戏界面content1、游戏操控content2、游戏趣味性影响content3、启动界面广告影响advertisement1、游戏中广告影响advertisement2。

载入分析所需要的模块和函数

在进行分析之前，我们首先载入分析所需要的模块和函数，读取数据集并进行观察。

示例

参阅教材内容

数据读取及观察

大家首先需要将本书提供的数据文件放入安装python的默认路径位置，并从相应位置进行读取。

示例

参阅教材内容

PART 03

变量设置及数据处理

变量设置及数据处理

示例

参阅教材内容

PART 04

岭回归算法

使用默认惩罚系数构建岭回归模型

前面我们提到岭回归本质上是增加了L2正则项，而正则项中有调节系数，可以用来控制正则化（惩罚）的力度，取值范围为大于等于0，惩罚的力度越大，就越不容易过度拟合数据。调节系数（惩罚系数）的初始值为1，下面我们以默认的惩罚系数1为例进行模型拟合。

示例

参阅教材内容

使用留一交叉验证法寻求最优惩罚系数构建岭回归模型

惩罚系数的初始值为1，但是我们可以找到最优惩罚系数，并使用以得到更好的模型拟合效果，寻求方法有很多种，下面我们使用留一交叉验证法寻求岭回归最优惩罚系数。

示例

参阅教材内容

使用K折交叉验证法寻求最优惩罚系数构建岭回归模型

我们还可以使用K折交叉验证法寻求岭回归最优惩罚系数。

示例

参阅教材内容

划分训练样本和测试样本下的最优岭回归模型

示例

参阅教材内容

PART 05

Lasso回归算法

使用随机选取惩罚系数构建岭回归模型

前面我们提到Lasso回归本质上是增加了L1正则项，而正则项中也有调节系数，下面我们首先随机选取惩罚系数0.2为例进行模型拟合。

示例

参阅教材内容

使用留一交叉验证法寻求最优惩罚系数构建Lasso回归模型

与岭回归类似，在Lasso回归中，我们同样可以找到最优惩罚系数，并使用以得到更好的模型拟合效果，寻求方法有很多种，下面我们使用留一交叉验证法寻求Lasso回归最优惩罚系数。

示例

参阅教材内容

使用K折交叉验证法寻求最优惩罚系数构建Lasso回归模型

我们还可以使用K折交叉验证法寻求岭回归最优惩罚系数。

示例

参阅教材内容

划分训练样本和测试样本下的最优Lasso回归模型

示例

参阅教材内容

PART 06

弹性网回归算法

使用随机选取惩罚系数构建弹性网回归模型

前面我们提到弹性网回归本质上是增加了L1正则项、L2正则项，两个正则项中均有调节系数，下面我们首先随机选取L2正则项惩罚系数0.01、L1正则项惩罚系数0.5为例进行模型拟合。

示例

参阅教材内容

使用K折交叉验证法寻求最优惩罚系数构建弹性网回归模型

与岭回归、Lasso回归类似，在弹性网回归中，我们同样可以找到最优惩罚系数，并使用以得到更好的模型拟合效果，寻求方法有很多种，限于篇幅，我们仅使用K折交叉验证法寻求弹性网回归最优惩罚系数。

示例

参阅教材内容

划分训练样本和测试样本下的最优弹性网回归模型

示例

参阅教材内容

PART 07

小　　结

小结

对比岭回归、Lasso 回归和弹性网回归三种高维数据惩罚回归算法。就本例而言，总体上各种方法之间差别不大。

PART 08

习　题

习题

使用“数据9.2”数据文件。“数据9.2”是某电子商务平台上某网商企业记录的2021年全年销售数据。该网上企业的店铺名称叫做ZE果业，主要经营芒果、木瓜、橙子等新鲜水果以及松子、核桃、开心果等坚果。数据分析采用的样本数据包括交易金额、消费者年龄、消费者信用、注册时间、在线时间、整体好评率等数据。我们共设置了6个变量，即“V1~V6”分别用来表示交易金额、消费者年龄、消费者信用、注册时间、在线时间、整体好评率，请以交易金额（V1）为响应变量，以消费者年龄（V2）、消费者信用（V3）、注册时间（V4）、在线时间（V5）、整体好评率（V6）为特征变量，构建惩罚回归算法模型。

1、载入分析所需要的库和模块

2、数据读取及观察。

3、变量设置及数据处理。

习题

4、构建岭回归算法模型。

（1）使用默认惩罚系数构建岭回归模型；

（2）使用留一交叉验证法寻求最优惩罚系数构建岭回归模型；

（3）使用K折交叉验证法寻求最优惩罚系数构建岭回归模型。

（4）划分训练样本和测试样本下构建最优岭回归模型

感谢聆听

THANKS