回归分析作业报告

郑昌乾
北京理工大学网络空间安全学院

2024 年 10 月 10 日

摘要

本文通过回归分析方法对不同数据集进行了建模与分析，旨在探究特征变量与目标变量之间的关系。主要采用了一元多项式回归和带正则化项的 Ridge 回归算法来处理两个不同的数据集。首先，针对人工生成的受扰动正弦函数数据集，通过多项式回归拟合不同阶数的多项式模型，分析了模型阶数与拟合精度的关系；其次，使用 Ridge 回归算法对 UCI 人脸红外热成像温度数据集进行建模，并通过调节正则化系数 λ，获取最佳模型参数，验证其预测效果。实验结果表明，多项式回归能够较好地拟合复杂的非线性关系，但在数据样本较小时易导致过拟合现象；Ridge 回归在控制模型复杂度和防止过拟合方面表现优异。最终，通过评估不同模型的预测性能，本文总结了各算法在不同应用场景中的优劣，并为回归分析在数据挖掘领域中的实际应用提供了理论依据与实践参考。

1 引言

数据挖掘（Data Mining）作为现代信息技术与大数据处理领域的重要研究方向，主要通过分析和挖掘大量数据集中的隐藏模式和规律，为预测、决策和知识发现提供支持。在数据驱动的应用场景中，如何有效地从大量数据中提取有价值的信息，已成为各行各业关注的热点问题。

本次数据挖掘课程作业的主题是对数据集进行回归分析，通过多种方法对不同数据集进行建模和分析，探索特征变量与目标变量之间的关系。在回归分析中，我们重点使用了一元多项式回归以及正则化回归中的 Ridge 回归对数据进行分析。这些回归方法能够处理复杂的函数关系，并通过引入正则化项有效地防止模型过拟合，提高模型的泛化能力。

本次作业分为以下几个部分：首先，我们根据给定要求人工生成了一个包含扰动的正弦函数数据集，用于一元多项式回归模型的训练与测试；其次，针对现如今 Ridge 回归在医疗领域的广泛研究与应用 [4][3][1]，我们从 UCI 数据库中选择了人脸红外热成像温度的多特征数据集，并对数据进行了一定的清洗和处理。最后，我们使用 Ridge 回归模型，通过调节正则化系数 ，获取正则化路径数据，并通过交叉验证确定最优的超参数 ，从而实现对目标变量的精确预测。并通过随机选取训练机和测试集进行验证，进行模型的评估。

通过本次作业的实践，我们不仅深入理解了回归分析的基本原理和算法实现过程，还探索了相关模型参数的预估方法，并且掌握了模型性能的评估指标（如均方误差 MAE 和均方根误差 RMSE）。这些指标能够帮助我们更好地衡量模型的拟合程度和预测效果。

本报告详细介绍了实验中使用的数据集及其预处理过程，回归分析的实验设计与结果分析，并对各回归模型的性能表现进行了总结和讨论。希望通过本次作业的完成，能够对数据挖掘中的回归分析有更深刻的理解，并为后续的实际应用提供理论基础和实践经验。

2 算法

2.1 数据预处理算法

在本次实验的数据预处理中，对于下载的数据集，我们选择 **Robust 标准化**作为我们数据预处理的主要算法。Robust 标准化是一种基于数据集的中位数 (Median) 和四分位差 (Interquartile Range, IQR) 的标准化方法。与传统的 Z-score 标准化不同，Robust 标准化对异常值 (Outliers) 更具鲁棒性（即稳定性），因此在数据集包含大量离群点时表现更优。Robust 标准化通过以下公式对每个特征 x 进行标准化处理：

$$x' = \frac{x - \text{median}(x)}{\text{IQR}(x)} \tag{1}$$

其中，x 是原始特征值，x' 是标准化后的特征值，$\text{median}(x)$ 表示特征 x 的中位数，$\text{IQR}(x)$ 是特征 x 的四分位差，定义如下：

$$\text{IQR}(x) = Q_3(x) - Q_1(x) \tag{2}$$

其中，$Q_3(x)$ 和 $Q_1(x)$ 分别表示特征 x 的第 3 四分位数（即 75th percentile）和第 1 四分位数（即 25th percentile）。该标准化方法能够有效地减少异常值对特征缩放比例的影响，从而更加稳健地反映数据的分布特性。

2.2 回归分析算法

根据要求，此次作业我们采用**多项式回归算法**和 **Ridge 回归算法**分别对相应的数据集进行回归分析，具体算法如下：

2.2.1 多项式回归算法

多项式回归 (Polynomial Regression) 是一种基于线性回归扩展的回归分析方法。与简单线性回归不同，多项式回归引入了特征的高次项（如平方项、立方项等），从而能够建模特征与目标变量之间更为复杂的非线性关系. 多项式回归的模型表示如下：

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_m x^m + \epsilon \tag{3}$$

其中：- y 为目标变量；- $\beta_0, \beta_1, \ldots, \beta_m$ 为回归系数；- $x, x^2, \ldots, x^m$ 分别是原始特征及其多项式特征；- ϵ 为误差项；- m 为多项式的阶数。

通过引入高次项，模型能够捕捉输入变量与目标变量之间的非线性关系。

多项式回归算法的步骤如下：

1. **数据转换：**将输入特征 x 转换为多项式特征，即 $[x, x^2, \ldots, x^m]$；
2. **线性回归：**使用线性回归对转换后的多项式特征进行拟合，确定回归系数 $\beta_0, \beta_1, \ldots, \beta_m$；
3. **模型预测：**利用训练得到的模型系数进行预测，得到目标值 y；
4. **误差计算：**通过均方误差（MSE）、均方根误差（RMSE）或其他误差指标，评估模型的预测性能。

2.2.2 Ridge 回归算法

Ridge 回归是一种带有 L2 正则化的线性回归算法，其目的是在最小化残差平方和的基础上，添加回归系数的平方惩罚项，从而减少模型的过拟合

风险。与普通的线性回归不同，Ridge 回归通过引入正则化项 $\lambda \sum \beta_i^2$，控制模型的复杂度。Ridge 回归的目标是最小化以下目标函数：

$$\min_{\beta} \left(\sum_{i=1}^{n} \left(y_i - \sum_{j=0}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right) \tag{4}$$

其中：

- y_i 是第 i 个样本的目标值，x_{ij} 是第 i 个样本的第 j 个特征；
- β_j 是回归系数；
- λ 是正则化超参数，用于调节正则化强度；
- $\sum_{j=1}^{p} \beta_j^2$ 是 L2 正则化项，它对所有回归系数施加惩罚。

通过调节 λ 的值，可以在模型复杂度和训练误差之间进行权衡。较大的 λ 值会减小模型系数，从而防止过拟合；较小的 λ 值则会使模型更加拟合训练数据。

Ridge 回归算法的步骤如下：

1. **数据准备：**将数据集划分为训练集和测试集，并对特征进行标准化处理，以确保所有特征具有相同的尺度；
2. **正则化系数选择：**通过交叉验证选择最优的正则化超参数 λ；
3. **模型训练：**使用 Ridge 回归模型对训练集进行拟合，确定回归系数 β_j；
4. **模型预测：**利用训练好的模型对测试集进行预测；
5. **误差评估：**计算均方误差（MSE）、均方根误差（RMSE）等指标，评估模型性能。

3 实验及结果分析

3.1 数据集

在本次实验中，根据实验所使用的模型，分别选取了以下两种数据集分别作为多项式回归分析算法和 Ridge 回归分析算法的数据集，具体介绍如下：

1. **受扰动的正弦采样数据 D_1：**使用正弦函数生成一个包含两个周期的数据集，设置振幅，从中均匀采样多个数据点，并对每个样本的目标变量 y_i 添加一个随机的扰动值，由此获得正弦采样数据集 D_1，具体的参数及选择如下：

   (a) **振幅 A：**此次实验中，我们设定振幅为 1

   (b) **随机扰动值 α：**此次实验中，我们设定每个采样的扰动值的范围为 1，并根据以下公式进行计算：

   $$y = y_0 + \text{random}(\alpha) \tag{5}$$

   其中 y_0 为原始采样值 random 函数会随机生成一个在 $[-\alpha, \alpha]$ 的值作为扰动

   (c) **采样值 ω** 本次实验的要求我们选择均匀采样 20 个数据点，在此基础上，为了探究不同采样率上生成数据的回归分析情况，我们又生成采样 50 个数据点和采样 100 个数据点的数据集作为回归分析的数据集进行对照分析。

2. **人脸红外热成像温度数据 D_2：**该数据来自于 UCI dataset repository 中的人脸红外热成像温度数据 [5]，该数据主要用于对人体口腔温度数据的评估。其包括从不同位置读取的关于患者的推断图像的温度，以及每个人测量的口腔温度。以此用于回归任务的分析，通过使用环境信息和热图像读数来预测口腔温度。

3.2 数据预处理

3.2.1 数据筛选

对于下载的体温数据，我们主要选取了 FLIR 设备所获取的人脸各个部分所得到的温度数据，并筛选了其中的以下类型的数据作为多维回归分析的输入特征 $x_1, x_2..., x_1 1$：

1. 左、右眼角周边的平均温度数据 $aveAllL$ 和 $aveAllR$
2. 左、右眼眶周围的平均温度数据 T_{RC1} 和 T_{LC1}
3. 左、右眼眦点周围的平均温度数据 $RCC1$ 和 $LCC1$

4. 额头中心、前、后、左、右周围的平均温度数据T_{FHCC1}、T_{FHRC1}、T_{FHLC1}、T_{FHBC1}、T_{FHTC1}

并且选取口腔区域的温度T_{OR}作为回归分析的目标值 y

3.2.2 数据标准化

针对本次实验所使用的人脸体温数据集，考虑到所选取的数据具有偏向性小，受各种因素影响大而容易出现极端值和离群点的特征，故对于本次实验该数据的，根据 2.1 节介绍的算法，我们对所筛选后的数据进行了 Robust 标准化，将其数据整体标准化为均值为 0，同时整体标准差在一定范围之内的数据

3.3 多项式回归分析实验

3.3.1 预期结果

由于本次多项式回归分析实验的数据来源于正弦函数的扰动采样，针对这一方面目前也有一定的相关研究 [2]，而在此次实验中，鉴于多项式回归分析的函数特点，我们选择采用泰勒展开进行预估，具体的公式如下：

$$\sin(\mathrm{x}) = x - \frac{x^3}{3!} + \frac{x^5}{5!} + ... + (-1)^{2n+1}\frac{x^{2n}}{(2n+1)!} \tag{6}$$

可以预期到的结果是：在不考虑数据扰动造成的过拟合现象下，预测误差 ϵ 与拟合阶数 m 之前呈负相关关系，同时对于奇数阶数，其更高一阶的拟合误差与其相差不大。因此对于不同阶数下的回归误差，整体呈随结束增加而阶梯减小的特点

3.3.2 实验过程

本次实验，我们主要采用 python 语言以及 sklearn 库编写程序，并通过以下步骤进行数据拟合以及误差分析：

1. 数据划分：对该数据的划分，我们通过调用 python 中的 $train_test_split$ 函数，将数据按照 $80\% : 20\%$ 的比例随机花粉划分为训练集 D_{train} 和测试集 D_{test}，并通过多次重复实验，进行不同的随机划分，以避免特殊划分造成的误导，从而获得普遍规律

2. 拟合分析：通过变换多项式的阶数 m，利用训练集 D_{train} 确定回归系数)。并利用测试集 D_{test} 进行误差分析，获取 MAE 和 RMSE 值。

   此处获取不同阶数 m 下的误差分析中，作业要求我们的 m 取值 1、2、3、4、5，但根据预期结果的内容，作业要求的 m 取值范围偏少，故在具体的实验中，我们将 m 的取值设置为 1 到 10。

3. 获得结论：利用 python 中的 matplotlib.pyplot 库，对拟合的函数图像以及误差分析获得的 MAE 与 RMSE 值进行比较以及作图。

3.3.3 实验结果

在多项式回归分析后，我们得到了不同阶回归下的拟合模型差异以及在不同条件下各种误差的差异：

1. 不同阶数 m 下回归模型的差异：参见图 1，可以发现，随着多项式模型阶数的增加，模型的复杂度增高，拟合曲线能够更好地贴合训练数据，捕捉更多数据中的非线性特征。一阶回归模型，即线性模型下，曲线非常简单，仅用一条直线拟合数据，明显无法很好地捕捉数据中的非线性模式。而随着阶数的增加，模型逐渐能够捕捉到数据的部分波动，但在一些高波动区域依然存在偏差。在五阶左右开始能够拥有相似的凹凸性，从而实现尽量贴合所有的训练点。

2. 不同阶数 m 和不同采样率 ω 下生成回归模型的误差的差异：参见图 2，可以得出以下结论：

   (a) 在阶数较低（$m < 8$）下，模型的误差（MAE 和 RMSE）在呈现阶梯下降的情况，整体上，当（$m = 7$）时，各个模型的损失情况均处于较低水平，此时模型的拟合效果较好，继续提高阶数到 10 时，误差开始缓慢增加，这可能是由于模型的过拟合现象导致的。同时对于相邻阶的训练的误差程度，对于奇数阶而言，其次一阶的误差均略微大于前一阶，这与此前所预测的结论相符。

   (b) 对于不同样本数下的模型，我们可以发现，当采样点较多（如 50 或 100 个采样点）的时候，模型阶数的差异与预期结果相似，可以认为在大样本数下，数据集的分布可以很好地代表训练集，训练模型的泛化程度也较高。而在采样点较少（20 个采样点）的时候，可以发现数据分布在$m = 4$ ，$m = 6$ 以及$m > 8$ 的时候均出

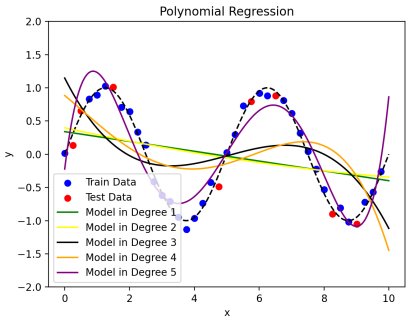
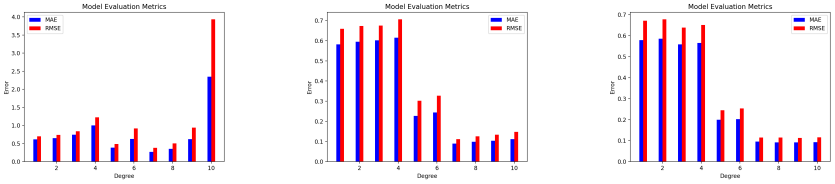


图 1: 不同阶数下的拟合模型

现了异常升高，由此可以判断模型在采样点数量较少的情况下对特殊点的拟合成都交叉，在多处异常点上下出现了过拟合情况，



(a) 20 个采样点下不同阶数回归模型的误差

(b) 50 个采样点下不同阶数回归模型的误差

(c) 100 个采样点下不同阶数回归模型的误差

图 2: 不同采样点数下多项式回归模型的误差分析

3.4 Ridge 回归分析实验

3.4.1 回归系数与正则化系数的关系

对于 Ridge 回归分析，正则化系数 λ 与各项的回归系数相关，当 λ 较低时，Ridge 回归退化为最小二乘回归，此时 λ 对回归系数没有额外的约束，λ 的值对各参数回归系数影响小，可能导致模型复杂度过高，容易过拟合。当 λ 值较高时，λ 的值对各参数回归系数影响大，致使模型回归系数变化大并且趋近于零，导致模型容易丧失回归能力。

根据 λ 的值对与回归系数的关系，可以估计 λ 的最优值约在回归路径中各个系数开始收敛的点附近。

3.4.2 确定回归系数

将进行清洗后的数据集 D_2 全部用作训练集，变换正则化系数 λ 的取值，由此确定各项的回归系数，并由此获取各项数据的回归系数与 λ 变化的正则化路径数据，由此估计 λ 的值。

在确定 λ 值之后，通过使用 RidgeCV 模型，进行交叉验证，获得 λ 值的最优值约为 20.09，由此确定回归系数。

3.4.3 回归分析

将数据按照 $80\% : 20\%$ 的比例随机划分为训练集 D_{train} 和测试集 D_{test}，并用选定的 λ 对训练集 D_{train} 进行训练，并且使用测试集 D_{test} 进行测试，进行回归分析。

实验按照要求，进行了 5 次重复实验，获取了多组 MAE 和 RMSE 值，并求得其平均值，用以分析结论。

3.4.4 实验结果

3.4.5 回归路径与 λ 值估计的关系

在变换 λ 值并获取回归系数后，我们利用多组数据，采用对数尺度绘制了 λ 在 $[10^{-3}, 10^{3}]$ 之间变换后的回归路径图，与采用 RidgeCV 进行交叉分析得到的 λ 值（20.09）相比较，可以发现最佳的 λ 值取值范围在 $[10^{0}, 10^{2}]$ 之间，在这一区间上，不同参数的回归系数均开始收敛，由此可证明此前部分对 λ 最优值的估计是正确的。

在确定回归系数后，根据实验可以确定多次 Ridge 回归的 MAE 和 RMSE 误差值，以及五次回归的平均值误差值。具体而言，在多次实验中，MAE 的误差值约在 0.45 左右，而 RMSE 的误差则在 0.6 上下浮动，且不同实验的差异也较大，由此可见，在此次回归实验中，我们选取的数据尽管进行了一定的数据标准化措施，但其原始数据存在一系列较大偏差值的情况，可能的原因是其数据来源（摄像机拍摄体温）本身就受各种因素影响，容易出现一些特殊的值。

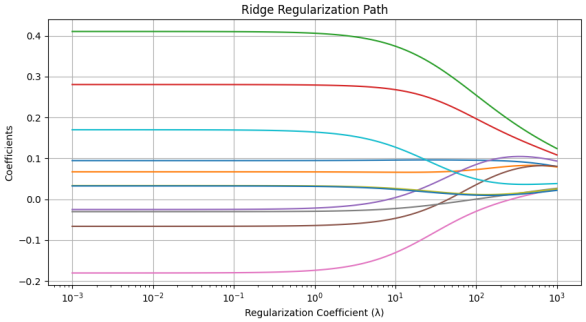


图 3: 对数尺度下的回归路径

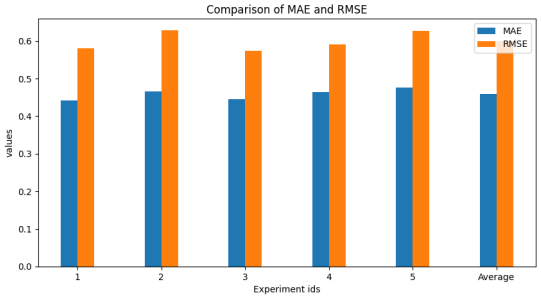


图 4: 多次 Ridge 回归的误差值和平均值比较

4 结论

4.1 多项式回归分析的结论

在本次实验中，我们针对生成的正弦函数数据集 D_1 进行了多项式回归分析，通过调整多项式的阶数 m （分别取值为 1, 2, 3, ..., 10），观察模型在训练集和测试集上的拟合效果。实验结果表明，随着多项式阶数的增加，模型在测试集上却出现了先减后增的趋势。这说明多项式回归能够通过引入高次项来捕捉非线性关系，但同时也更容易出现过拟合现象。

根据正弦函数的泰勒展开式，我们可以预测出数据拟合的精确程度呈阶梯下降的趋势，这在我们的具体实验中可以相印证。由此可见，对于已知波形或者生成模式的函数，通过一定的数学分析工具，我们可以拟合出一些超越函数的多项式表达式，由此较好地优化多项式函数的参数。同时，从另一个角度，我们也可以通过不同阶数下多项式拟合的精度变化情况反推出原始数据可能的拟合函数，从而优化拟合模型，提告拟合精度。

在拟合过程中，我们发现，当训练集和测试集较少的时候，不同阶数的变化波动较大，误差值在高阶上更容易出现异常的高值，而随着训练集和测试集数量的增加，高阶上的过拟合现象则逐渐消失。由此可见较少的训练数据是造成数据过拟合的重要原因。同时也表明，多项式回归分析具有一定的泛化能力，但对于较小数据集下抗过拟合的鲁棒性较低，有一定的局限性。

4.2 Ridge 回归分析的结论

Ridge 回归分析在引入了 L2 正则化项后，能够有效控制模型复杂度，从而缓解多元线性回归中存在的多重共线性问题。在人脸红外热成像温度数据集 D_2 上的实验表明，Ridge 回归通过调节正则化系数 λ，可以在模型复杂度与拟合误差之间实现较好的平衡。

在本次实验中，我们通过交叉验证确定了 Ridge 回归的最优正则化系数 $\lambda = 0.1$，此时模型在训练集和测试集上的均方误差（MSE）均表现较好。实验结果显示，当 λ 值较小时（如 $\lambda = 0.001$），模型的拟合效果接近普通的线性回归模型，表现出较高的训练集精度但测试集误差较大，说明模型过度拟合了训练数据。而当 λ 值过大时（如 $\lambda = 1$），正则化项对回归系数的限制过强，导致模型复杂度过低，训练集和测试集的误差均大幅增加。

因此，Ridge 回归能够通过正则化项对模型复杂度进行有效控制，从而减少模型的过拟合风险，并在存在多重共线性问题的数据集上获得稳定的

回归效果。在实际应用中，合理选择正则化系数 λ，可以使模型在训练误差和测试误差之间取得更好的平衡，从而提高模型的泛化能力。

4.2.1 总结与展望

通过本次回归分析实验，我们深入理解了多项式回归与 Ridge 回归的工作原理及其参数调整、数据选择、结论预估等任务。在实际应用中，应根据数据集的特征及任务目标，合理选择回归方法及模型超参数，以获得更好的回归效果。未来的工作中，我们可以尝试引入其他类型的正则化方法（如 Lasso 回归）以及更多的非线性特征转换方法（如基函数展开、样条回归等），以进一步提升模型的拟合能力与鲁棒性。

5 参考文献

[1] Talal Almutiri, Khalid Alomar, and Nofe Alganmi. “Predicting Drug Response on Multi-Omics Data Using a Hybrid of Bayesian Ridge Regression with Deep Forest”. In: *International Journal of Advanced Computer Science and Applications* 14.5 (2023).

[2] Igor Djurović et al. “Parameter estimation of coupled polynomial phase and sinusoidal FM signals”. In: *Signal Processing* 149 (2018), pp. 1–13. ISSN: 0165-1684. DOI: https://doi.org/10.1016/j.sigpro.2018.02.023. URL: https://www.sciencedirect.com/science/article/pii/S0165168418300793.

[3] Vadamodula Prasad and Tamada Srinivasa Rao. “Implementation of regularization method ridge regression on specific medical datasets”. In: *Int J Res Comput Appl Inf Technol* 3.2 (2015), pp. 25–33.

[4] Jeremy Rubin et al. “Ridge Regression for Functional Form Identification of Continuous Predictors of Clinical Outcomes in Glomerular Disease”. In: *Glomerular Diseases* 3.1 (Dec. 2022), pp. 47–55. ISSN: 2673-3625. DOI: 10.1159/000528847. eprint: https://karger.com/gdz/article-pdf/3/1/47/4116050/000528847.pdf. URL: https://doi.org/10.1159/000528847.

[5] Quanzeng Wang et al. “Infrared Thermography for Measuring Elevated Body Temperature: Clinical Accuracy, Calibration, and Evaluation”. In: *Sensors (Basel, Switzerland)* 22 (2021). URL: https://api.semanticscholar.org/CorpusID:245585208.