

基于机器学习的学生成绩预测研究

花卉^{*} 陈爱萍

金陵科技学院 计算机工程学院, 中国·江苏 南京 211169

摘要: 学生成绩是衡量学习效果和教学质量的重要指标, 准确的成績预测能够为个性化教学和学习干预提供有力支持。本研究通过学生在不同知识点上的平时成绩构建回归模型, 以预测期末成绩, 并识别出对期末成绩有显著影响的关键知识点。研究采用了数据预处理、特征选择、模型训练和评估等步骤, 结果表明, 平时成绩与期末成绩之间存在显著的线性关系。通过对比 Ridge 回归、Lasso 回归和 Elastic Net 回归, 研究不仅量化了各个知识点对期末成绩的影响强度, 还证明了这些模型具有较高的预测精度。这为教育工作者提供了基于数据的干预措施与个性化教学指导的科学依据。

关键词: 学生成绩预测; 机器学习; 回归分析; 深度学习

Research on Student Performance Prediction Based on Machine Learning

Hui Hua^{*} Aiping Chen

Jinling Institute of Technology, School of Computer Engineering, Nanjing, Jiangsu, 211169, China

Abstract: Student performance is a key indicator for assessing learning outcomes and teaching quality. Accurate performance prediction can provide strong support for personalized teaching and learning interventions. This study builds a regression model based on students' performance in various knowledge areas to predict final exam scores and identify key knowledge areas that significantly impact the final results. The study involves steps such as data preprocessing, feature selection, model training, and evaluation. The results show a significant linear relationship between students' performance in regular assessments and their final exam scores. By comparing Ridge regression, Lasso regression, and Elastic Net regression, the study not only quantifies the influence of each knowledge area on final exam performance but also demonstrates that these models have high prediction accuracy. This provides educators with a scientific basis for data-driven interventions and personalized teaching strategies.

Keywords: student performance prediction; machine learning; regression analysis; deep learning

0 前言

随着互联网技术和在线教育平台的快速发展, 教育逐渐向数字化、个性化方向转变。在这种背景下, 如何利用海量的学习数据进行分析和预测, 提升学生的学习效果和教学质量, 成为教育领域的研究热点^[1]。

学生成绩是衡量学习效果的重要指标, 准确预测学生成绩不仅能帮助教师优化教学策略, 还能为学生提供个性化的学习路径和及时的学习干预。然而, 传统的成绩评估方法往往依赖于学生的基础信息(如性别、年龄、入学成绩等), 难以全面反映学生的学习行为和表现。

近年来, 随着大数据技术和机器学习算法的迅速发展, 教育领域开始引入数据驱动的预测模型。许多研究都聚焦于基于数据挖掘和机器学习方法的学生成绩预测。马婉晴等^[2]以学生综合数据库为基础, 挖掘课程间的关联程度, 利用大数据技术对高校学生成绩进行分析和预测; 杨佳骏^[3]使用 Transformer 模型根据学生历史数据和风险指数对本学期课程成绩进行了推测; 潘花等^[4]利用机器学习中的多元线性回归方法建立了根据先修课程对数学建模成绩进行了

预测等; 李慧^[5]一种融合情感特征的在线学习成绩预测方法, 获得学习者的动态学习状态并预测在线学习成绩。

这类方法不仅具有较高的准确性, 还能揭示哪些特征对最终结果的影响最大, 也为个性化教学和教育决策提供了有力支持。因此, 研究基于机器学习的学生成绩预测具有重要的理论意义和应用价值。

1 问题描述

依托《Python 程序设计》课程, 旨在构建一个基于线性回归的预测模型, 预测期末成绩, 通过分析各知识点对期末成绩的贡献, 教师也可以明确哪些知识点在教学中最为关键, 这为教学的重点内容和课堂时间分配提供了科学依据。

在该问题中, 自变量(输入特征)包括每个学生多项知识点的平时成绩。因变量(目标变量)是学生的期末考试成绩。通过将平时成绩作为自变量, 期末成绩作为目标变量, 构建一个线性回归模型来捕捉自变量与因变量之间的线性关系。

2 线性回归模型

线性回归模型是一种用于分析自变量与因变量之间关

系的监督学习算法,其基本思想是通过拟合一条直线来描述自变量与因变量之间的线性关系^[6]。模型的目标是找到一组最佳的线性方程参数,使得自变量的线性组合能够最小化预测值与实际值之间的误差。线性回归假设因变量 Y 和自变量 X 之间存在线性关系:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \epsilon$$

其中: Y 是目标变量(期末成绩);

$X_1, X_2, \cdots, X_n$ 是自变量(平时成绩);

β_0 是截距;

$\beta_1, \beta_2, \cdots, \beta_n$, 是回归系数,表示每个自变量对目标变量的影响程度;

ϵ 是误差项,用于表示模型的预测值与实际值之间的误差。

线性回归模型训练的过程就是通过优化算法最小化损失函数,找到最优的参数 $\beta_0, \beta_1, \cdots, \beta_n$ 。模型一旦训练完毕,可以利用新的平时成绩数据来预测学生的期末成绩。该模型简单、计算高效,并且容易解释每个特征对最终预测的影响程度,特别适合用于成绩预测这类解释性强的预测任务。

3 数据获取和特征选取

根据 Python 课程学习要求,将学习内容划分为 23 个知识点。在确保数据统一性的基础上,基于 python123 平台,获取同一老师任教的三个班级 20 次作业和期末成绩数据共 21 项(如表 1 所示)。

表 1 2023-2024-2 学期 python 成绩数据

score1	score2	score3	score4	score5	...	score20	final_score
99	98	76	95	96.33		95	90
99	71	60.2	100	96.33		73	73
97	98	37.2	95	88.67		81	65
99	95	61.8	9	96		87	88
98	97	55.5	95	92		94	91
84	97	58.6	80	76.67		89	83

在基于机器学习的学生成绩预测中,数据清洗是关键步骤之一。原始数据常包含缺失值、异常值、重复数据等问题,直接影响预测模型的性能。对于缺失值通过删除、填充(设为 0 分)或预测等方式处理;其次,将每次成绩统一为百分制;最后去除重复数据,以确保数据的独立性和准确性。经过清洗,选择其中的 15 维 102 条数据作为实验数据。

4 实验过程

4.1 实验设计

本实验将数据集划分为训练集和测试集,采用 8 : 2 的比例进行划分。在训练过程中,线性回归模型通过最小二乘法来确定最优的回归系数。最小二乘法的目标是最小化预测值 $\hat{y}_i$ 与实际值 y_i 之间的误差平方和,即最小化以下损失函数:

$$\text{损失函数} = \sum_{i=1}^m (y_i - \hat{y}_i)^2$$

其中: $\hat{y}_i$ 是模型预测值, y_i 是实际值;

y_i 是第 i 个样本的实际期末成绩;

$\hat{y}_i$ 是模型预测的第 i 个样本的期末成绩;

m 是训练集样本的总数。

通过不断调整回归系数 β ,最小化该损失函数,找到最能反映训练数据的模型参数。在模型训练完成后,可以通过查看模型的回归系数来了解每个平时成绩特征对期末成绩的影响。回归系数越大,说明该平时成绩对期末成绩的影响越大。正回归系数表示该特征与期末成绩呈正相关,负回归系数表示呈负相关。

4.2 模型评估结果的分析

模型评估是机器学习过程中的关键步骤,旨在衡量模型在未见过的数据上的表现,判断其泛化能力。对于线性回归模型,本实验采用均方误差(Mean Squared Error, MSE)和决定系数(R^2 Score)模型预测结果与真实值之间的差异。

4.2.1 均方误差

均方误差用于衡量模型预测值与真实值之间的误差。具体来说,它是预测误差的平方和的平均值,即:

$$\text{MSE} = \frac{1}{m} \sum_{i=1}^m (y_i - \hat{y}_i)^2$$

MSE 提供了模型预测结果与真实值之间的误差大小,可以告诉我们模型在测试集上的预测误差的实际大小。较大的 MSE 值可能难以直观解读,对于更具可解释性的结果,通常可以使用均方根误差(RMSE),即 MSE 的平方根,如 $\text{MSE}=10$,计算出其 RMSE (MSE 的平方根),即 $\sqrt{10} \approx 3.16$,则可以解释为,模型的预测值和实际值平均相差约 3.16 分。

4.2.2 决定系数

决定系数是线性回归模型中常用的评价指标。它表示自变量(平时成绩)对因变量(期末成绩)变化的解释程度,取值范围为 0 到 1,公式为:

$$R^2 = 1 - \frac{\sum_{i=1}^m (y_i - \hat{y}_i)^2}{\sum_{i=1}^m (y_i - \bar{y})^2}$$

R^2 则衡量了模型对因变量的解释能力,帮助我们了解模型拟合的效果。如果 R^2 值接近 1,说明模型对数据有很好的拟合能力,如果 R^2 值过低,则说明线性回归可能不足以捕捉学生的平时成绩和期末成绩之间的复杂关系。

4.3 实验结果分析

在 Python 中,可以使用 scikit-learn 库中的 Linear Regression 类来实现这一过程。模型会在训练集上自动拟合,确定最优的回归系数。

4.3.1 均方误差和决定系数

通过实验,线性回归模型在学生成绩预测任务中的表

现为:

均方误差 (MSE): 35.98372679815176

决定系数 (R^2): 0.9397621706035655

可以看出, MSE 较低, 说明模型在测试集上的预测结果和真实值之间的差异不是很大, 模型的预测精度较高, 预测成绩在 6 分左右浮动; R^2 为 0.939 表示模型可以解释约 93.9% 的期末成绩的变化, 较高的 R^2 值也表明平时成绩能够很好地解释期末成绩的变化, 说明模型捕捉到了两者之间的线性关系。

4.3.2 回归系数

在此回归模型中, 每个系数代表一个特征 (平时成绩) 对期末成绩的影响大小。系数可以为正或负, 数值的绝对大小反映了影响的强弱。

其中, -0.1079, -0.1019, -0.0741, -0.3017, -0.3524 等特征的系数为负值, 说明这些特征对期末成绩有负向影响。当这些特征的值增加时, 期末成绩会有所下降。特别是第 9 个特征 (-0.3017) 和第 15 个特征 (-0.3524), 它们的系数比较大, 意味着这些特征对期末成绩的负向影响更强。

第 4 个特征的系数 1.9939 是一个非常显著的正系数, 这表明该特征对期末成绩有较大的正向影响。换句话说, 当第 4 个特征的成绩提高时, 期末成绩也会显著增加。

其他较小的正系数如 0.0202、0.0288, 表示这些特征也对期末成绩有正向影响, 但影响较小。

4.3.3 预测数据

从预测成绩和实际期末成绩的对比 (见图 1) 来看, 中高分段 (如 80 分及以上) 的样本预测较为准确, 说明平时成绩和期末成绩之间的关系可能在这个范围内表现较为线性; 低分段的预测偏差较大, 这可能是由于低分段学生的平时成绩与期末成绩的关系更加复杂, 且线性回归模型无法很好地捕捉这些复杂性。

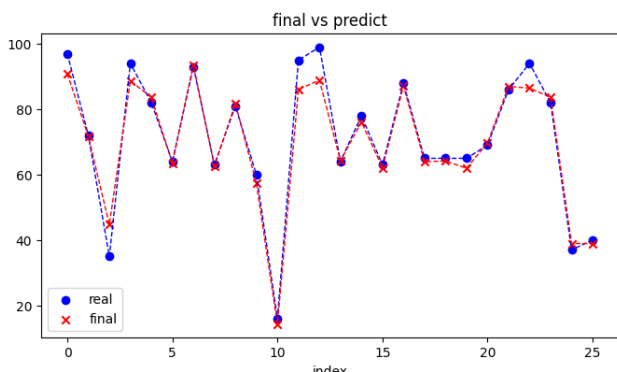


图 1 期末成绩与预测成绩对比视图

5 模型优化

根据计算出的方差膨胀因子 (VIF) (如表 2 所示), 可以看出实验中存在多重共线性问题, 可能导致模型回归系数不稳定。为了解决多回归模型中多个特征之间存在高度相

关性这一问题, Liu 等研究者建议使用正则化回归模型, 如 Ridge 回归和 Lasso 回归^[7]。Ridge 回归通过 L2 正则化减少了特征之间的多重共线性, 而 Lasso 回归则能通过 L1 正则化进行特征选择, 自动筛选出与目标变量关系密切的特征。

表 2 实验回归系数和膨胀因子

特征	对应课程知识点	回归系数	膨胀因子
特征 1	语言基础	-1.30631	2.4786
特征 2	turtle 画图	-2.85142	18.2733
特征 3	分支结构	-1.28828	2.27802
特征 4	for 循环	12.50122	6.894604
特征 5	while 循环	-0.03941	2.522627
特征 6	算法基础	-0.06961	14.26457
特征 7	函数应用	-0.33049	3.149893
特征 8	字符串基本操作	0.103364	2.470711
特征 9	字符串应用	-3.78329	9.768188
特征 10	列表基本操作	0.33107	4.03422
特征 11	列表应用	-0.83731	10.9509
特征 12	集合应用	0.162837	5.547569
特征 13	字典基本操作	0.558373	4.809324
特征 14	文件基本操作	0.570041	2.984485
特征 15	综合应用	-5.06043	37.92261

5.1 Ridge 回归

Ridge 回归通过引入正则化项, 来惩罚回归系数的大小, 使得模型的参数不会过大, 从而避免过拟合。其损失函数变为:

$$\text{损失函数}_{(R)} = \sum_{i=1}^m (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^n \beta_j^2$$

其中: λ 是正则化参数, 控制惩罚项的大小, 称为正则化强度;

$\lambda \sum_{j=1}^n \beta_j^2$ 是所有回归系数的平方和, 称为 L2 正则化项。

其中, 对于较小的 λ 回归更接近普通线性回归, 几乎没有正则化效果; 而较大的 λ 模型的回归系数被强烈抑制, 趋向于 0, 模型复杂度降低, 模型可能出现欠拟合。为了找到合适的 λ , 可以通过交叉验证或网格搜索进行超参数调优。

5.2 Lasso 回归

Lasso 回归通过引入 L1 正则化约束来解决多重共线性问题, 并自动执行特征选择。与普通线性回归相比, Lasso 回归在处理高维数据集时, 能够通过减少特征系数的数量, 防止模型过拟合, 并选择对目标变量影响较大的特征。

$$\text{损失函数}_{(L)} = \sum_{i=1}^m (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^n |\beta_j|$$

与普通线性回归不同, Lasso 回归在目标函数中引入了

L1 范数的惩罚项 $\sum_{j=1}^n |\beta_j|$ 。这一惩罚项具有的效果是使某些回归系数趋于 0，从而在回归模型中实现特征选择。

5.3 Elastic Net 回归: Lasso 和 Ridge 的结合

Elastic Net 是 Lasso 和 Ridge 的结合，它通过同时应用 L1 和 L2 正则化，既具备 Lasso 的特征选择能力，又能像 Ridge 那样处理多重共线性。Elastic Net 的损失函数为：

$$\text{损失函数}_{(E)} = \sum_{i=1}^m (y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^n \beta_j^2 + \lambda_2 \sum_{j=1}^n |\beta_j|$$

5.4 实验对比

5.4.1 均方误差和决定系数

通过实验结果来看（如表 3 所示），三种模型比线性模型都有较好的表现，Lasso 回归的表现最好，接着是 Ridge 回归，而 Elastic Net 回归的效果相对较差。以下是每个回归模型表现的分析。

表 3 四种模型的实验均方误差和决定系数

模型类型	均方误差	决定系数
Ridge	31.150030871198346	0.9478539214173533
Lasso	27.256787437824705	0.95513587224255565.4
Elastic Net	39.89711752449962	0.9332110380803095

Ridge 回归在处理数据时，所有特征的系数都会被缩小，但不会将任何特征的系数缩小到 0。这意味着模型保留了所有的特征，因此它能较好地处理存在多重共线性的问题。 R^2 值为 0.9479，表示模型可以解释约 95% 的期末成绩的变化，表现相对较好。

Lasso 回归通过 L1 正则化自动选择了对预测期末成绩最重要的特征，并将不重要的特征系数缩小为 0。MSE 为 27.26，相对于 Ridge 回归表现更好，说明 Lasso 能够更有效地缩减特征数，提高模型的解释能力。 R^2 为 0.9544，表明模型可以解释约 95.4% 的期末成绩变化，预测效果最优。

Elastic Net 结合了 L1 和 L2 正则化，因此它在某种程

度上既能够进行特征选择，又能够处理多重共线性问题。然而在本次实验中，Elastic Net 的效果相对较差。MSE 为 39.90， R^2 为 0.9332，说明它的模型对数据的解释能力不如 Ridge 和 Lasso 回归。

5.4.2 回归系数

从图 2 中可以看出三种模型在对平时成绩的回归系数上有显著差异。

特征 4 在三个模型中都有显著的正系数，表明它对期末成绩有非常大的正向影响。这个特征可能是预测期末成绩的关键特征，应该在教学中重点关注和强化。

特征 9 在所有模型中都有较大的负系数，表明它对期末成绩有显著的负向影响，Lasso 回归和 Elastic Net 都显示出这个特征的重要性，Lasso 尤其倾向于赋予更大的绝对值，进一步强调了该特征对模型的影响。

特征 2 在 Lasso 回归中，该特征的系数为 -2.40，表明它对期末成绩有较大的负面影响。Elastic Net 则将该系数缩小为 0，表示该特征在该模型中的贡献不大。Ridge 回归中，该特征的影响较小，但仍有负向影响。

特征 1 在 Ridge 和 Lasso 回归中表现为负系数，表示该特征对期末成绩有负向影响。然而，Elastic Net 将该特征的系数几乎缩小为 0，表明其对期末成绩的贡献较小。

特征 8 和特征 10 在 Lasso 和 Elastic Net 模型中系数为 0，表明这些特征被认为对预测结果的贡献不足，因此被 Lasso 和 Elastic Net 选择性忽略。

Ridge 回归在处理多重共线性时保留了所有特征的影响，但通过 L2 正则化缩小了系数。适合当所有特征都有贡献时使用。Lasso 回归通过 L1 正则化自动选择了关键特征，忽略了不重要的特征。它更适合特征稀疏的场景，并且在这个例子中，对第 2 个、9 个特征赋予了更高的权重。

Elastic Net 结合了 Ridge 和 Lasso 的优点，既能处理多重共线性，又能进行特征选择。它在这里将一些系数缩小到 0，但不像 Lasso 那样激进。

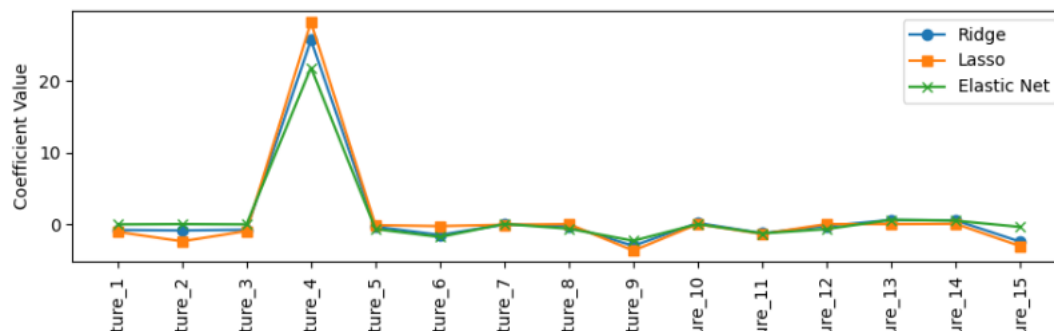


图 2 三种模型实验的回归系数

6 结论与未来展望

本研究通过对学生在不同知识点上的平时成绩进行分析，构建了一个线性回归模型，以预测学生的期末成绩。研究表明，平时成绩与期末成绩之间存在显著的线性关系，尤其是特征 4 和特征 9 对期末成绩有较强的预测能力。尽通分析，教师可以根据学生在不同章节的学习表现进行有针对性的干预与指导，帮助学生提升学习效果，最终提高期末成绩。这种数据驱动的教学分析能够有效支持个性化教育，优化教学策略。

通过对比 Ridge 回归、Lasso 回归和 Elastic Net 发现在预测期末成绩的任务中，若期望提高模型的解释性且特征数量较多，Lasso 回归或 Elastic Net 回归可能是最佳选择；如果特征间的关系较为复杂且希望保留所有特征，Ridge 回归则更适合。

尽管线性回归模型具有较好的解释性和预测能力，但在处理复杂的非线性关系时存在局限性，除了可以考虑交叉验证和特征选择以提升模型的精度外，还可以考虑使用更复杂的模型（如决策树、随机森林或神经网络）来处理非线性关系，并结合心理因素、学习习惯等非线性特征，构建更加精准的成绩预测模型。

参考文献：

[1] 张广海,祖璇,王杨.基于DBO-SVM的学生成绩三分类预测模型

[J].兰州文理学院学报(自然科学版),2024,38(2):56-61.

[2] 马婉晴,冯军,袁园.基于课程关联的高校学生成绩预测模型研究[J].浙江科技学院学报,2024,36(3):205-217.

[3] 杨佳骏,田圻,覃天.基于Transformer模型的学情预警系统[J].软件,2024,45(6):142-144.

[4] 潘花,仇海全,车金星.基于多元线性回归的数学建模成绩预测研究[J].南昌工程学院学报,2024,43(4):94-100.

[5] 李慧.融合情感特征的在线学习成绩预测研究[J].小型微型计算机系统,2023,44(7):1360-1366.

[6] 刘晓雲,刘鸿雁,李劲松,等.基于多元线性回归的学生成绩预测研究[J].计算机技术与发展,2022,32(3):203-208.

[7] Liu Y, Zhang J. Ridge and Lasso regression for handling multicollinearity in educational data[J]. Applied Mathematics and Computation,2020(385):125456.

作者简介：花卉（通讯作者）（1979-），女，中国江苏扬州人，讲师，从事计算机应用、图像处理等研究。

基金项目：全国高等院校计算机基础教育研究会计算机基础教育教学研究项目：基于师生赋能的 Python 程序设计课四维协同模式研究（项目编号：No.2024-AFCEC-071）；金陵科技学院 2024 年“数字赋能应用型高校高质量人才培养”课题：数字化赋能背景下大学计算机信息技术课程改革探索与实践（重点）（项目编号：No.SZH202413）。