基于多元线性回归的学生成绩预测研究

刘晓雲[1],刘鸿雁[1],李劲松[2],王冠帮[1]
(1. 渤海大学 教育科学学院,辽宁 锦州 121000;
2. 渤海大学 信息科学与技术学院,辽宁 锦州 121000)

摘 要:随着数据挖掘技术在教育领域的深入应用,使得成绩预测成为改进教学质量的重要手段之一。对学生进行成绩预测,可以督促学生提高学习效率以及鞭策教师改进教学质量,更好地完善教学,达到最佳效果。但在目前研究中,虽然对成绩预测应用已十分广泛,但是多是基于学生全部成绩对某门课程成绩的预测,忽略了成绩预测的时效性。因此提出基于多元线性回归方法构建一年级成绩预测毕业成绩的预测模型。以某学校计算机应用专业的学生课程成绩为研究对象,构建相应的多元线性回归预测模型。通过大量实验以及检验证明,利用一年级成绩预测毕业成绩可行,并且构建的成绩预测模型具有极高的预测精度,可以为改进教学方案提供参考信息,有助于提高学校的教学质量和学生的学习效果。

关键词:成绩预测;教育数据挖掘;线性回归;教学质量;显著性检验
中图分类号:TP305;G420 **文献标识码**:A **文章编号**:1673-629X(2022)03-0203-06
doi:10.3969/j.issn.1673-629X.2022.03.034

Research on Student Achievement Prediction Based on Multiple Linear Regression

LIU Xiao-yun[1],LIU Hong-yan[1],LI Jin-song[2],WANG Guan-bang[1]
(1. School of Education Science,Bohai University,Jinzhou 121000,China;
2. School of Information Science and Technology,Bohai University,Jinzhou 121000,China)

Abstract:With the deep application of data mining technology in the field of education,achievement prediction has become one of the important means to improve the teaching quality. The prediction of students' performance can urge students to improve their learning efficiency and urge teachers to improve their teaching quality,so as to better improve teaching and achieve the best results. However,in the current research,although the application of grade prediction has been very extensive,most of them are based on the whole score of students to predict the grade of a certain course,ignoring the timeliness of grade prediction. Therefore,a prediction model based on multiple linear regression is proposed to predict the graduation scores of the first grade. A multivariate linear regression prediction model was established based on the course performance of students majoring in information and computing science in a certain school. Through a large number of experiments and tests,it has been proved that it is feasible to use the grades of freshmen to predict the graduation grades,and the prediction model built has a very high prediction accuracy,which can provide reference information for improving the teaching scheme,and help to improve the teaching quality of the school and the learning effect of students.

Key words:achievement prediction;educational data mining;linear regression;teaching quality;significance test

0 引 言

随着中国经济的快速发展,人才需求越来越大,教育也越来越受到社会的关注。为了保证教学质量,国家也不断颁布新的教育整改政策,数据挖掘技术也逐渐深入地应用到了教育领域,例如关联规则、多元线性回归、聚类分析、分类预测等等。其中成绩预测可以督促学生,使学生及时调整自己的学习方法,改变学习策略,并且使教师及时改进教学策略,所以成绩预测是提升学生成绩的重要手段。它也成为了教育数据挖掘领域的一个热点研究课题[1]。

对学习成绩进行预测分析对提高教学质量有着十分重要的作用,一些国内外学者对此已经开展了相关研究。尤佳鑫利用多元线性回归方法,预测了云环境下的学生学业成绩[2]。徐铭希采用多种机器学习算法

收稿日期:2021-03-29 **修回日期**:2021-07-29
基金项目:辽宁省自然科学基金项目(2019-ZD-0503);辽宁省教育科学研究项目(WJ2020004,LJ2020003)
作者简介:刘晓雲(1998-),女,硕士研究生,研究方向为现代教育技术研究;刘鸿雁,副教授,硕导,研究方向为数字图像处理、模式识别与智能计算、智能学习、教育技术与课程整合。

对学生成绩进行预测并构建最优模型[3]。赵光等人利用多元线性回归方法,构建大学英语四级考试成绩预测模型[4]。张晓等人通过多元线性回归,分析了基础课程对专业课程的影响[5]。汪慧利用多元线性回归方法,建立通过影响电子技术的 6 门课的成绩预测该门课的模型[6]。虽然国内外学者已经开展相关的成绩预测研究,但多是利用现有全部成绩预测某科成绩。利用一年级预测毕业成绩较少,未能充分发挥成绩预测的及时性。

目前普遍认为,一个人的学习成绩是符合一定趋势的,并且一年级时期开展的课程,包括基础课和通识课,对毕业总体成绩也有着一定的影响。其中如解析几何这样的专业基础课程,对后面其他专业课的学习有着直接的影响。因此利用一年级预测毕业成绩具有可行性和可预测性。

回归分析是研究统计规律的方法之一。应用回归分析评价考试成绩不仅能分析各种因素对考试成绩的影响大小,还能对成绩进行合理的预测[7-8]。鉴于多元回归分析的以上优点,所以建立多元回归模型不仅可以帮助教师改进教学方法,还可以帮助学生及时调整自己的学习方法,以便得到更好的成绩,为提高教学质量提供了保障。

1 回归分析

1.1 线性回归

线性回归有很多实际用途。分为以下两大类:如果目标是预测或者映射,线性回归可以用来对观测数据集的和 X 的值拟合出一个预测模型。当完成这样一个模型以后,对于一个新增的 X 值,在没有给定与它相配对的 y 的情况下,可以用这个拟合过的模型预测出一个 y 值。

给定一个变量 y 和一些变量 $X_1,X_2,\cdots,X_p$,这些变量有可能与 y 相关,线性回归分析可以用来量化 y 与 X 之间相关性的强度,评估出与 y 不相关的 X,并识别出哪些 X 时子集包含了关于 y 的冗余信息。

1.2 多元线性回归

多元回归分析是指在相关变量中,将一个变量视为因变量,其他一个或多个变量视为自变量,建立多个变量之间线性或非线性的数学模型数量关系式,并利用样本数据进行分析的统计分析方法。另外,也要讨论多个自变量与多个因变量的线性依赖关系的多元回归分析,称为多元多重回归分析模型。通常影响因变量的因素有多个,这种多个自变量影响一个因变量的问题,可以通过多元回归分析来解决。在线性回归分析中,多元线性回归比一元线性回归具有更大的实用意义[9-10]。

多元线性回归分析的基本任务如下:根据因变量与众多自变量的实际观察值建立因变量对多个自变量的多元线性回归方程;评定各个自变量对因变量影响的相对重要性以及测定最优多元线性回归线性方程的偏高度等[11-13]。许多多元非线性回归问题可以通过多元线性回归来解决,所以多元线性回归具有广泛的应用。

1.3 多元线性回归模型

设变量 Y 与变量 $X_1,X_2,\cdots,X_p$ 间有如下的线性关系:

$$Y=\beta_0+\beta_1X_1+\cdots+\beta_pX_p+\varepsilon \tag{1}$$

其中,β_0 是回归常数,$\beta_1,\beta_2,\cdots,\beta_p$ 是总体回归参数,当 $p=1$ 时,称公式(1)为一元线性回归模型,$p\geqslant 2$ 时,称之为多元线性回归模型。ε 为随机误差,且服从 $\varepsilon\sim N(0,\sigma^2)$ 分布。

参数 β 的估计方法最常用的是最小二乘估计法(ordinary least square,OLS),其目标函数为最小化:

$$Q(\beta)=\sum_{i=1}^{n}\|y_i-x_i\beta\|^2 \tag{2}$$

因在解决实际问题时,矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 通常都是奇异的。所以当 $\boldsymbol{X}'\boldsymbol{X}$ 是非奇异矩阵时,表明变量之间不完全相关,而这时得到的最小二乘估计为:

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{3}$$

从而可得回归模型为:

$$\hat{\boldsymbol{Y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{4}$$

1.4 多元线性回归模型的检验

由建立的多元线性回归模型以及已得到的回归系数,要对整个回归方程进行拟合检验,可以采用 R^2 检验。

判定系数 R^2 的定义为:

$$R^2=1-\frac{\mathrm{SSE}}{\mathrm{SST}} \tag{5}$$

其中,SSR 表示回归平方和,其定义如公式(6),反映了由于 x 与 y 之间的线性关系引起的 y 的变化部分;SST 表示总离差平方和,其定义如公式(7),反映因变量的 n 个观察值与其均值的总离差;SSE 表示残差平方和,其公式如公式(8),反映除了 x 对 y 的线性影响之外的其他因素对 y 变差的作用,是不能由回归直线来解释的 y 的变差部分。

$$\mathrm{SSR}=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2 \tag{6}$$

$$\mathrm{SST}=\sum_{i=1}^{n}(y_i-\bar{y})^2 \tag{7}$$

$$\mathrm{SSE}=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2 \tag{8}$$

三者之间的关系满足:

$$SST = SSR + SSE \tag{9}$$

R^2 反映的是回归直线对数据的拟合优度,取值在[0,1]之间。R^2 趋近于 1,说明回归方程拟合得越好,相反,R^2 趋近于 0,说明回归方程拟合得越差。

2　基于多元线性回归的学生成绩预测研究

鉴于 SPSS 软件是目前教育研究领域使用最为广泛的统计软件之一,具有界面美观、操作简洁的特点,因此该文在实验数据处理中使用了 SPSS 软件[14],用其对实验数据进行单次实验。而预处理和统计分析部分基于 Matlab 系统完成。

2.1　数据预处理

2.1.1　数据收集

实验数据选用某学校计算机应用专业一年级共 55 名学生的课程成绩。由于部分课程涉及分流培养,因此实验数据仅使用 17 门课程。

2.1.2　数据处理

(1)为保护学生隐私,将原始学生姓名用编号替代以及将性别、学号等身份信息隐藏,只保留所需的成绩、课程名称等基本信息。

(2)为了使数据结果更具有合理性、普遍性,除去极端学生成绩的影响,因此去掉低于平均成绩大于 $X+3\sigma$ 或小于 $X-3\sigma$ 的学生,最后剩下 53 名学生的课程成绩。

(3)实验数据中的部分课程成绩采用等级制进行的赋分(优秀、良好、中等、及格、不及格),对这类数据前期进行了转换和处理,转换原则为“优秀”对应 95 分,“良好”对应 85 分,“中等”对应 75 分,“及格”对应 65 分,“不及格”对应 59 分。

(4)为避免数据属性的影响,对所有实验数据都进行了归一化[0,1]处理,最终获得的部分实验数据如表 1 所示。

表 1　部分学生成绩

学号	解析几何	数学分析 1	数学分析 2	高等代数 1	高等代数 2	…
1	76.0	66.3	73.9	74.9	76.0	…
2	75.8	70.9	63.5	84.6	75.0	…
3	64.2	71.5	61.0	63.6	60.0	…
4	85.4	72.7	78.7	91.2	74.6	…
5	73.4	64.2	64.6	85.7	90.2	
…	…	…	…	…	…	…

2.2　建立多元线性回归模型及其分析

2.2.1　实验原理与结果

平均绝对误差(mean absolute error,MAE)是所有单个观测值与算术平均值的偏差的绝对值的平均,所以选用简便、直观的平均绝对误差作为评估成绩预测模型的预测误差指标[15],其计算公式如下所示。

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\frac{|\,Score_p - Score\,|}{Score} \tag{10}$$

其中,N 为样本个数;Score 和 $Score_p$ 分别为原始成绩和模型预测成绩。MAE 值越小,模型预测误差越小,预测越准确。

该文随机从 53 名学生中选出 3 名、5 名、10 名和 20 名作为测试样本(训练样本数量即为 50 名、48 名、43 名和 33 名),并分别进行 100 次随机选择。然后对得到的 MAE 值取其平均值,得到的最终平均预测性能结果如表 2 所示。

表 2　训练和测试样本 MAE 详情

实验	训练样本数量	训练 MAE	测试 MAE
1	50	0.012 1	0.019 9
2	48	0.012 0	0.019 1
3	43	0.011 4	0.020 4
4	33	0.009 4	0.023 0

统计结果表明,训练和测试误差都小于 1.9%,说明构建的预测模型具有较高的预测精度,已证明利用一年级预测毕业成绩可行。此外,从表中也可看出构建的模型性能对训练样本需求较低,更利于推广。

2.2.2　单次实验结果分析

为了更加清楚地展现实验结果,分别选用上述四种实验的某一次实验结果进行具体分析。利用 SPSS 软件进行分析,令四年总体平均成绩为因变量,17 门课程成绩为自变量。

(1)实验 4。

实验 4 的 33 个训练样本得到的标准线性回归方程为:

$$\hat{Y} = 0.291 + 0.037X_1 + (-0.042)X_2 + 0.009X_3 + 0.188X_4 + 0.093X_5 + 0.131X_6 + (-0.014)X_7 + 0.009X_8 + (-0.021)X_9 + 0.025X_{10} + (-0.005)X_{11} + 0.085X_{12} + 0.192X_{13} + 0.140X_{14} + 0.290X_{15} + 0.092X_{16} + (-0.108)X_{17} \quad (11)$$

表 3 模型摘要

模型	R	R 方	调整后 R 方	标准估算的错误	R 方变化量	F 变化量	自由度 1	自由度 2	显著性 F 变化量
1	0.946[a]	0.894	0.838	1.533 69	0.894	15.937	17	32	0.000

注:a. 预测变量:(常量),001、002、003、004、005、006、007、008、009、010、011、012、013、014、015、016;

b. 因变量:四年最终平均分。

表 4 多元回归模型概要

目标	信息选择方法	信息标准	准确性/%
四年最终平均分	向前步进	-18.208	97.3

注:信息标准用于与模型进行比较。具有较小信息条件值的模型拟合度得更好。

对所建立的实验 4 的线性回归模型进行 R^2 检验,从表 3 可以看出,R^2 的值为 0.894,接近 0.9,趋近于 1,说明模型的拟合度很高。从表 4 可以看出,模型的准确性为 97.3%(>95%),进一步说明模型的拟合度高。

通过模型预测出剩余 20 个测试样本的预测值,如表 5 所示。预测差值最高不超过 3.5 分,平均误差为 1.43%,预测性能精度较高。

表 5 实验 4 真实值和预测值对比

测试样本	四年最终平均分	预测值	差值(绝对值)	误差/%
1	83.88	83.47	0.41	0.48
2	82.28	81.48	0.80	0.97
3	77.59	76.18	1.41	1.81
4	77.82	78.87	1.05	1.34
5	89.12	89.36	0.24	0.26
6	82.72	81.00	1.72	2.08
7	87.70	85.07	2.63	2.99
8	87.05	83.90	3.15	3.61
9	85.08	82.94	2.14	2.51
10	83.91	82.67	1.24	1.48
11	85.95	86.29	0.34	0.40
12	85.63	85.37	0.26	0.30
13	89.43	88.96	0.47	0.52
14	87.60	86.15	1.45	1.65
15	86.00	86.25	0.25	0.29
16	83.13	83.21	0.08	0.10
17	82.27	82.89	0.62	0.75
18	85.84	89.15	3.31	3.86
19	81.02	80.64	0.38	0.46
20	82.16	84.48	2.32	2.82
平均值	——	——	1.21	1.43

(2)实验 3。　　回归方程为:

类似地,实验 3 的 43 个训练样本得到的标准线性

$$\hat{Y} = 0.358 + (-0.056)X_1 + (-0.062)X_2 + (-0.024)X_3 + 0.131X_4 + 0.101X_5 + 0.162X_6 + (-0.007)X_7 + (-0.014)X_8 + 0.008X_9 + 0.023X_{10} + 0.011X_{11} + 0.092X_{12} + 0.163X_{13} + 0.079X_{14} + 0.264X_{15} + 0.127X_{16} + (-0.073)X_{17} \quad (12)$$

通过模型预测出剩余 10 个测试样本的预测值,如　　0.9% ,预测性能精度较高。表 6 所示。预测差值最高不超过 1.5 分,平均误差为

表 6　实验 3 实际值与预测值对比

测试样本	四年最终平均分	预测值	差值(绝对值)	误差/%
1	85.95	86.35	0.40	0.46
2	85.63	85.58	0.05	0.06
3	89.43	88.26	1.17	1.30
4	87.60	86.10	1.50	1.71
5	86.00	86.11	0.11	0.13
6	83.13	82.09	1.04	1.25
7	82.27	81.81	0.46	0.56
8	85.84	84.73	1.11	1.30
9	81.02	80.36	0.66	0.81
10	82.16	83.44	1.28	1.56
平均值	——	——	0.78	0.90

(3)实验 2。　　程为:

实验 2 的 48 个训练样本得到的标准线性回归方

$$\hat{Y} = 0.334 + (-0.052)X_1 + (-0.037)X_2 + (-0.026)X_3 + 0.149X_4 + 0.096X_5 + 0.164X_6 + (-0.010)X_7 + (-0.013)X_8 + (-0.004)X_9 + 0.017X_{10} + 0.001X_{11} + 0.087X_{12} + 0.167X_{13} + 0.093X_{14} + 0.284X_{15} + 0.119X_{16} + (-0.073)X_{17} \quad (13)$$

通过模型预测出剩余 5 个测试样本的预测值,如　　0.97% ,预测性能精度较高。表 7 所示。预测差值最高不超过 1.4 分,平均误差为

表 7　实验 2 实际值与预测值对比

测试样本	四年最终平均分	预测值	差值(绝对值)	误差/%
1	83.13	81.90	1.23	1.48
2	82.27	81.91	0.36	0.44
3	85.84	86.50	0.66	0.76
4	81.02	80.55	0.47	0.58
5	82.16	83.54	1.38	1.67
平均值	——	——	0.82	0.97

(4)实验 1。　　程为:

实验 1 的 50 个训练样本得到的标准线性回归方

$$\hat{Y} = 0.329 + (-0.062)X_1 + 0.002X_2 + 0.164X_3 + 0.094X_4 + 0.132X_5 + 0.001X_6 + (-0.010)X_7 + 0.010X_8 + 0.010X_9 + 0.038X_{10} + (-0.011)X_{11} + 0.088X_{12} + 0.185X_{13} + 0.094X_{14} + 0.268X_{15} + 0.107X_{16} + (-0.084)X_{17} \quad (14)$$

通过模型预测出剩余 3 个测试样本的预测值,如表 8 所示。预测差值最高不超过 1.2 分,平均误差为 0.61%,预测性能精度较高。

表 8 实验 1 实际值与预测值对比

测试样本	四年最终平均分	预测值	差值(绝对值)	误差/%
1	85.84	86.36	0.52	0.06
2	81.02	80.76	0.26	0.32
3	82.16	83.35	1.19	1.45
平均值	——	——	0.66	0.61

通过这四个实验的单次实验表明,结果与训练样本数量关系不大,可行性较强。并且构建的预测模型具有较高的精度,可以为学校改进教学方案,提高教学质量提供一定的参考信息,具有重要的意义。

3 结束语

成绩预测是提高教学质量的重要辅助工具之一,但是目前多是基于全部成绩进行研究。因此该文提出利用多元回归方法构建通过一年级成绩预测毕业成绩的预测模型,并以某学校计算机应用专业的学生课程成绩为研究对象开展研究。大量实验结果表明可以利用一年级成绩预测毕业成绩,并且该文构建的预测模型具有较高的准确度。该研究可以为教学的改进提供依据,为老师对学生采取帮扶措施提供参考。但学生成绩预测是一个比较复杂的课题,本次研究只考虑了成绩因素,因此在下一步的研究中会考虑学科背景、素质测评等更多因素,构建更加精确的预测模型。

参考文献:

[1] 林鹏飞,何秀青,陈甜甜,等. 深度学习视阈下 MOOC 学习者流失预测及干预研究[J]. 计算机工程与应用,2019,55(22):258-264.

[2] 尤佳鑫,孙 众. 云学习平台大学生学业成绩预测与干预研究[J]. 中国远程教育,2016(9):14-20.

[3] 徐铭希. 机器学习在学生成绩预测中的应用[J]. 电子制作,2019(2):42-44.

[4] 赵 光,王栓宏,孙 珩. 大学英语四级考试成绩预测模型构建与实证分析[J]. 中国西部科技,2015,14(4):94-95.

[5] 张 晓,李晓戈. 基于多元线性回归的学生成绩分析[J]. 计算机与数字工程,2020,48(9):2089-2092.

[6] 汪 慧. 多元线性回归在大学成绩预测中的应用[J]. 保山学院学报,2020,39(5):84-88.

[7] STACEY J,BUSH C,DIPASQUALE T. The hidden blood loss in proximal femur fractures is sizeable and significant[J]. Journal of Clinical Orthopaedics and Trauma,2021,16:239-243.

[8] 孙 毅,刘仁云,王 松,等. 基于多元线性回归模型的考试成绩评价与预测[J]. 吉林大学学报:信息科学版,2013,31(4):404-408.

[9] MOUSTRIS K P,NASTOS P T,LARISSI I K,et al. Application of multiple linear regression models and artificial neural networks on the surface ozone forecast in the greater Athens area,Greece[J]. Advances in Meteorology,2012,2012:978-988.

[10] 张俭鸽,李颖颖. 基于多元线性回归预测模型的 sensor 态势研究[J]. 计算机技术与发展,2011,21(9):229-232.

[11] PARK M H,JU M,JEONG S,et al. Incorporating interaction terms in multivariate linear regression for post-event flood waste estimation[J]. Waste Management,2021,124:377-384.

[12] 徐龙飞,郁进明. 不同优化器在高斯噪声下对 LR 性能影响的研究[J]. 计算机技术与发展,2020,30(3):7-12.

[13] 顾金池. 学生成绩影响因素分析与预测研究——基于多元回归和决策树模型[J]. 管理观察,2019(25):156-157.

[14] MAAOUANE M,ZOUGGAR S,KRAJAČIĆ G,et al. Modelling industry energy demand using multiple linear regression analysis based on consumed quantity of goods[J]. Energy,2021,225:120270.

[15] KHASHEI M,HAMADANI A Z,BIJARI M. A novel hybrid classification model of artificial neural networks and multiple linear regression models[J]. Expert Systems with Applications,2012,39(3):2606-2620.