本 科 毕 业 论 文（设计）

落后生预警模型

学生姓名 朱如靖

学号 14020031146

指导教师 王勇

院、系、中心 信息科学与工程学院

专业年级 14级计算机科学与技术

论文答辩日期 年 月 日

**中 国 海 洋 大 学**

落后生预警模型

完成日期：\_\_\_\_\_\_\_\_\_\_

指导老师签字：\_\_\_\_\_\_\_\_\_\_\_

答辩小组成员签字：\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_

\_\_\_\_\_\_\_\_\_\_\_

落后生预警模型

摘要

越来越多的大学生因为在一学年内挂科太多或者没有获得足够的学分而留级甚至退学，在认识到这个严重的现状之后，我决定着手建立一个分析某位学生是否有留级或者退学风险的模型。这种预警模型出来的结果有两种，一种是有风险，另一种是没有风险。有风险指学生有留级或者挂科过多的危险。于是对中国海洋大学2013,2014,2015级的学生数据进行分析，其中数据信息包含了学生的学业成绩，学生智能卡消费信息，图书馆的打卡记录，借还书记录还有学生基本信息数据。本文打算采用机器学习中的监督习方式，利用所获得到的原始数据进行分析处理和学习。由于预测结果只有两种所以是一个二分类问题，分类算法有很多，包括K邻近算法，支持向量机，逻辑回归，决策树，朴素贝叶斯算法。因为不知道哪个模型会表现的优异，哪个表现糟糕，所以我们需要做的就是挨个尝试各种算法并最终找到相对比较准确的预测模型。该模型不仅可以针对在校学生的学业情况，还可以扩展到商业等其他领域，诸如可以预警公司员工是否有被解雇的危险等等。

关键词：落后生预警;机器学习;监督学习;数据挖掘

A Model for Prediction of At-risk Students by Using Educational Big Data

Abstract

More and more College students face the danger of failing to go up to the next grade or dropping out because they have failed too much subjects during a term or failed to get enough credits.Building a model which can predict whether a student will drop out the school is significant to lower the hanging rate. Students’ data about their behavior in the school ,such as the frequency of eating breakfast,can be [utilize](http://dict.cn/utilize)d to build a prediction model which reflect how much time they spend on their study. One method to build the model is Machine Learning.Python is a good language when it comes to the ML.There are three tools which will make the work easy.They are numpy,pandas and Scikit-Learn,which are popular in the area of ML.The most classic ML algorithms,such as SVM,Logistic and so on,will be introduced in the following.There are two kinds of results from this early warning model. One is risky, and the other is riskless. Risk refers to the risk that students will have to repeat or hang too much. Since there are only two prediction results, it is a binary classification problem. There are many classification algorithms, including K-nearest neighbor algorithm, support vector machine, logistic regression, decision tree, and naive Bayes algorithm. Because we do not know which model will perform well and which will perform poorly, all we need to do is try all kinds of algorithms and eventually find relatively accurate prediction models. The model can not only address the academic situation of students in school, but also can be extended to other areas such as business, such as can warn the company employees whether there is danger of being fired and so on.

Keywords:Prediction Model, Machine Learning, Supervised Learning, Data Mining.

目录

[摘要 1](#_Toc863041406)

[Abstract 2](#_Toc1571896983)

[1. 前言 5](#_Toc1779359396)

[2. 实际问题分析 6](#_Toc2146423458)

[3. 数据处理 6](#_Toc1805358372)

[3.1准备数据 7](#_Toc670732227)

[3.2 分析数据 8](#_Toc852049917)

[3.2.1 学生成绩表分析 8](#_Toc322166710)

[3.2.2入馆记录表和借书记录表 9](#_Toc1651244829)

[3.2.3学生信息表 10](#_Toc393626323)

[3.2.4智能卡打卡记录（如图3.8） 11](#_Toc1649836483)

[3.2.5合并数据 12](#_Toc848238571)

[3.2.6 学生分类标准 13](#_Toc1843124874)

[3.3数据预处理 14](#_Toc522330053)

[3.3.1数据归一化 14](#_Toc1218213059)

[3.3.2字符特征编码 15](#_Toc908313726)

[3.3.3划分测试级和训练集 15](#_Toc1025270258)

[3.4 特征选择 16](#_Toc484977892)

[3.4.1为什么要特征选择 16](#_Toc105325086)

[3.4.2如何进行特征选择 17](#_Toc2134832085)

[4. 模型建立 20](#_Toc110328727)

4.1算法选择

[4.2模型比较（比较结果参照表4-1） 20](#_Toc1628527824)

[4.2.1 KNN算法 20](#_Toc1945407230)

[4.2.2 Decision Tree 21](#_Toc274286307)

[4.2.3 Logistic Regression 22](#_Toc708990925)

[4.2.4 Naive Bayes 23](#_Toc1211090940)

[4.2.5 Support Vector Machines 23](#_Toc914568744)

[4.2.6模块持久化 24](#_Toc1657559613)

[4.3优化模型 25](#_Toc1486362409)

[5. 总结 26](#_Toc646228919)

[6. 致谢 28](#_Toc1617885616)

[7. 参考文献 29](#_Toc201920167)

1. 前言

人工智能在当今迅猛的发展，各行各业都在迅速向人工智能靠拢，谷歌的‘AlphaGo‘，无人驾驶技术以及小米的智能家居等等，都向我们证明了未来是属于人工智能的。2018年的谷歌开发者大会也向人们展示了谷歌未来的人工智能研究方向。然而机器学习是目前最接近人工智能的系统，试想一下如果没有机器学习作为辅助手段，那么创建人工智能将会是费事费力的。所以机器学习是人工智能的基础。如今机器学习的应用更是家常便饭，例如垃圾邮件的分类，图像识别技术，某商品的销售趋势预测等等数不胜数。可以说机器学习无处不在，然而机器学习必须建立在对大量数据的分析之上，通过对已有数据的分析和处理加之学习之后就能对新的数据作出相对正确的预测。

如今在有了学生智能卡系统和教务处系统的条件下，大量的学生数据可以被收集起来，研究者们认为这些数据可以反映学生学习的努力程度[1]和学生的学习成绩。

为了降低大学生的挂科率和留级率，许多学者利用学生在校表现等方面的数据来识别他们是否有学业预警的危险。他们认为可以成功预测学生在一门课程上的最终成绩[2]。预警模型可以提前提醒那些濒临挂科的学生来降低学校留级率。但是上述研究中存在普遍的问题就是预警模型只针对单独的一门课程而不是对所有的选修课程进行综合分析。另一个问题就是之前许多研究都是为网上课程设计的，现在缺乏对传统课程的合适研究模型.

所以在这篇论文中，预警模型是针对在校大学生的传统课程（非线上授课），找到学业落后的学生。使用中国海洋大学2013级本科生的数据，经处理后得到样本集，样本分为落后生和普通生两类，将其转变为二分类问题，（分类标准参见之后3.2.6章节）对多种算法运用交叉检验，针对落后生寻找符合要求的模型，然后展示了实验结果，并进行分析，最后对模型进行一定的改进。

1. 实际问题分析

在实际的学习生活中，学生的学习成绩的优异与否往往和他的一些行为有关，好的学生偏向于上自习次数多一点，早起的次数多一点。学生的专业同样也对学业成绩有一定的影响，例如某些专业的课程难度偏大，或课程任务偏重等等都会导致学生无法在最后取得理想的成绩。学生的个人信息也可能是一个值得挖掘的地方。学生的户籍反映了他是来自哪里的学生。众所周知，来自江苏，山东，河北等省份的学生素来有着良好的学习习惯（可能和高中的教育方式有关），所以他们可能在步入大学殿堂之后学习可能要比其他地区的学生要轻松一点，相反来自那些偏远教育落后地区的学生可能就不占太大的优势，有可能成为他们学业落后的一项原因。综上所述，现在的任务就是从数据中分析出可能影响学生成绩的特征，然后再找出这些特征的关联程度，根据数值特征进行机器学习建立理想的模型从而达到预测落后生的效果。

1. 数据处理

数据处理是数据挖掘中最为重要的一个环节，挖掘出的数据的价值取决于前期数据处理的好坏。如何把大量的原始数据处理成为我们需要的数据，例如从大量的学生智能卡消费信息中提取出学生的吃早饭信息，是最后建立成功模型的关键。由于numpy和pandas库提供了很好的数据处理方法，于是接下来就利用这两个库进行相应的数据处理。

3.1准备数据

从中国海洋大学数据中心取出数据，经分析2013级学生数据最为完整，于是以2013级学生为研究对象，为了运用数据挖掘算法完成预测任务，我们从原始数据中提取就餐数据、图书馆借书和入馆数据、成绩数据、学生基本信息，设计并计算特征，得到用于实验的数据集。因为部分数据涉及隐私，数据中学号均采取去敏操作。表3-1介绍了提取的原始数据，包括其记录数和具体描述。

表3-1：原始数据描述

|  |  |
| --- | --- |
| 表名 | 描述 |
| 基本信息表 | 基本信息表中的数据包括学生专业，入学年份，户籍所在地，生源地，性别，民族等学生个人信息。 |
| 成绩表 | 成绩表是用来判断某学生是属于落后生还是普通生的依据。通过学生的一学年所取得的总学分是否满足学校的要求来分类学生。除了标识学生类别之外，该表还包含了学年学期，专业，课程名称，学分等信息。一行数据信息代表了某学生在某学期修了某一门课程，这门课程的学分是多少，他是重修还是初修，最后他的期末成绩是多少。有一少部分的课程的最终成绩是等级制的，由于这些课程所占比重非常至少，重要性也不是很高，所以可以考虑删除该类难以处理的课程。 |
| 入馆记录 | 该表中主要包括两个信息：学生学号和入馆时间。理论上来说，学生进出图书馆的次数代表了该学生的学习热情程度，频率越大的学生越有可能取得优异的成绩。 |
| 借书记录表 | 记录的是学生借阅和归还图书的时间，包括学号，时间和借阅或归还的动作。借阅图书的次数也能在一定程度上反映学生的学习热情 |
| 智能卡打卡记录 | 学生在学校的消费记录都存储在校园智能卡的系统中。譬如学生在学校餐厅，超市的刷卡记录包括时间，金额和员工号都记录在智能卡的打卡记录当中。从该表中提取出学生的吃早饭记录是关键。 |

3.2 分析数据

3.2.1 学生成绩表分析

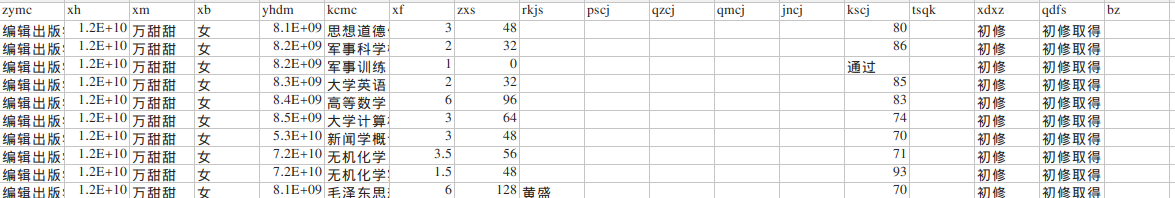


图 3-1

在’‘grades.csv‘’文件中存储的是学生成绩信息，图3.1是表中的详细信息，注意到图中有很多信息，包括专业、学号、姓名、性别、选课号、课程名、课程学分、课程学时、课程分数等等。我们将用这张表来判断学生是否为落后生，我的判定依据就是依照中国海洋大学相关标准：在中国海洋大学，学生在学期末不足12学分会受到“学业警示”，在学年末没有满足学分和成绩要求则会受到留级处分。根据成绩表，提取有用信息，包括学生学号，选修课程学分，选修课程成绩得到一个精简的表之后再删除除2013-2014学年之外的学期数据行，注意到课程成绩列中包含非数值型的分数（如“通过”），根据实际情况一般这些课程并不能正确反映学生的学习状况，所以过滤掉这些成绩。之后对学号进行分类[5]，计算每个学生每学期的平均分，2013-2014学年的加权平均分。最终得到的数据如图3.2所示。

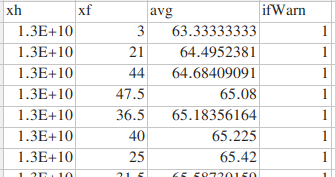


图 3‑2

加权平均分这个信息能够反映很多，如何评判一个学生的成绩好坏，一方面看他是否没有修够足够的学分，还有一方面是看他的加权平均分。年纪排名以及奖学金的评定都是以加权平均分为依据的。所以判定一个学生是否落后生可以参照他的加权平均分。

3.2.2入馆记录表和借书记录表

入馆记录表（如图3.3）借书记录（如图3.4）

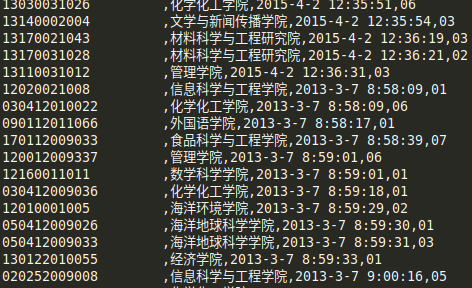


图 3-3

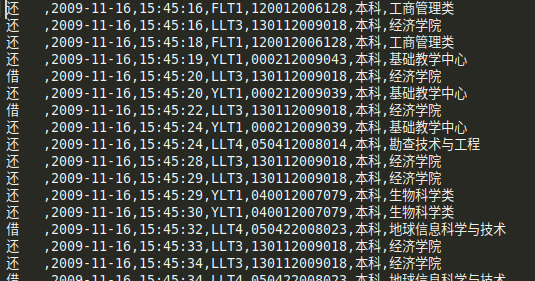


图 3-4

对于入馆记录，我最终要得到的是每名学生在2013-2014学年内去图书馆的次数，所以仍然需要删除所属学院这列无用信息。再看时间数据，他是精确到时分秒，其实没有必要，于是除保留日期之外删除时间。最终在新表的基础上仍然根据学号分类得到学生的入馆次数。如图3.5.

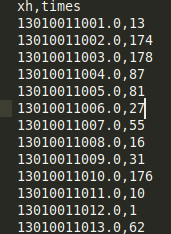


图 3‑5

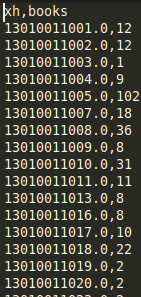


图 3-6

由于借书表与图书馆记录表差不多，进行相似操作得到最终数据表，图3.6。

3.2.3学生信息表

这个表的信息量有点大，所以我选择不展示该表。对信息表中提取我所需要的数据，包括学生学号，专业，民族，生源地如图3.7

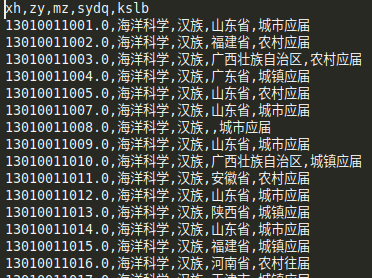


图 3‑7

3.2.4智能卡打卡记录（如图3.8）

是否每天坚持迟早饭是一个很好的习惯，身为一个大学生的我感同身受。那些很少吃早饭的学生一般偏向于睡觉到很晚因此错过了迟早饭的习惯，而那些坚持吃早饭的学生偏向于早起去上课或者自习。久而久之，行为体现在成绩上的差距就越来越明显。我把九点之前有过打卡记录的数据规定为吃早饭行为，统计每位学生在2013-2014学年之内的吃早饭总数如图3.9

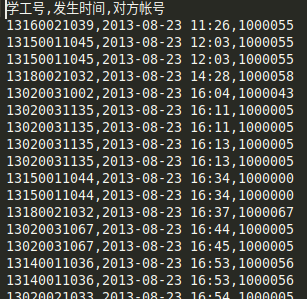


图 3-8

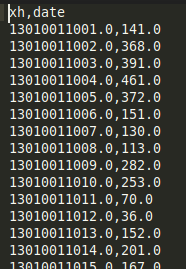


图 3-9

3.2.5合并数据

经过上面的数据清洗和数据聚合操作，得到了5张新表，这五张表中描述了学生的特征如表3-2

表 3‑2

|  |  |
| --- | --- |
| Features | Definitions |
| 专业 | 分为法学、工学、教育学、经济学、理学等 |
| 生源地 | 表示学生来自哪个省份 |
| CityOrTown | 学生是城镇户口还是农村户口 |
| TotalCount | 图书馆入馆总次数，即学生在大一学年总的入馆次数 |
| TotalBorrow | 借书总次数，即学生在大一学年总的借书次数 |
| TotalBreakfast | 学生在一学年之内的吃早饭次数 |
| avg | 学生在一学年之内的加权平均分 |
| TotalCredits | 学生在一学年之内所修的总学分 |
| 上学期是否受到学业警示 | 在中国海洋大学，学校会在每学期末予以学业警示，根据上学期的成绩和学分计算每个学生的学业警示情况。 |
| 上学期期末成绩 | 在成绩记录中有少数课程为等级制，这些课程大多无法体现学生学业表现，所以删除这些课程，保留百分制的课程，计算每个学生的期末成绩。在求得成绩后，考虑不同专业的平均成绩不同，所以需要对成绩进行标准化，将所有学生的成绩划归到同一个分布。 |

由于这些数据存在与不同的表格之中，所以我们对各个表进行合并操作，合并依据是相同的学号。

3.2.6 学生分类标准

依据中国海洋大学的有关标准，我们设计指标度量学生表现。在中国海洋大学，学生在学期末不足12学分会受到“学业警示”，在学年末没有满足学分和成绩要求则会受到留级处分。根据以上规定将学生分为落后生和普通生，衡量学生的学业表现，这样做有很高的实际应用价值，因为落后生往往学业成绩较差，这符合我们的研究目标，而且学生受到过多处分会被强制退学，降低退学率是学校工作的重要部分。

表 3‑3

|  |  |
| --- | --- |
| category | definition |
| 落后生 | 学年末受到学业警示或留级的学生为落后生，记作正类 |
| 普通生 | 除落后生外的学生为普通生，记为负类 |

3.3数据预处理

数据预处理是在前期数据处理的前提下把提取出的数据进一步处理来符合机器学习算法所需要的标准。数据和特征决定了机器学习的上限，而模型和算法的应用只是让我们逼近这个上限。所以可以看出数据处理的重要程度不亚于模型和算法，甚至有过之而无不及。所以在经过前面的数据处理之后，我们还要迎来最后一个处理环节。数据处理大约占据了整个数据挖掘和机器学习全流程百分之六十以上的时间和精力。

3.3.1数据归一化

经过数据分析之后，得到许多特征，由于他们的单位以及数值范围的不一致，如果不提前进行数据预处理可能会破坏预测模型的准确性。所以对原始数据进行归一化（又称规范化）操作。归一化的任务主要有两个[3]：

1 把数变为（0，1）之间的小数  
        主要是为了数据处理方便提出来的，把数据映射到0～1范围之内处理，更加便捷快速，应该归到数字信号处理范畴之内。  
2 把有量纲表达式变为无量纲表达式  
        归一化是一种简化计算的方式，即将有量纲的表达式，经过变换，化为无量纲的表达式，成为纯量。

经过归一化之后，有两个好处：

1. 提升模型的收敛速度

2. 提升模型的精度

常用的归一化方法有min-max标准化 和 z-score 标准化，这里选取第一种方式。min-max标准化（也称为离差标准化，是对原始数据的线性变换，使结果值映射到[0 - 1]之间）的公式：newvalue=(oldvalue-min)/(max-min)

3.3.2字符特征编码

在前面的章节里面，对连续型的数值特征进行了预处理，但是数据集中还存在其他的一些特征，例如专业类别，生源地，民族，农村还是城市，这些都是字符串型的数据，不能被我们机器所学习，所以想要利用这些数据，首先就必须给这些数据进行编码。

scikit-learn库里面提供了一种比较好的编码方式”one-hot”编码。首先定义一下独热编码：独热编码即 One-Hot 编码，又称一位有效编码，其方法是使用N位状态寄存器来对N个状态进行编码，每个状态都有它独立的寄存器位，并且在任意时候，其中只有一位有效。[12]

3.3.3划分测试级和训练集

由于手上处理好的数据是2013级学生的，这是全部数据，我们在根据这些数据建立模型之后要对模型进行正确率的分析，所以必须将全部数据划分为训练集（training set）和测试集（test set）。为避免数据划分带来的不平衡的影响，这里采用交叉验证法。交叉验证法的定义：”交叉验证法”[8]先将数据集D划分为k个大小相似的互斥子集每个子集都尽可能保持数据分布的一致性，即从D中通过分层采样得到。然后，每次用k-1个子集的并集作为训练集，余下的子集作为测试集；这样就可以获得k组训练/测试集，从而可以进行k次训练和测试，最终返回的是k个测试结果的均值。这种划分避免了划分不当带来的偶然性，使算法生成的模型更加的一般化。在这次落后生的模型当中，为了确定k的最佳值，依次枚举k从5到10带入交叉算法中验证，看效果最好的是哪一个。最终决定k选取5.

3.4 特征选择

前面的章节我们已经将所有的特征确定下来，接下来要做的就是如何利用这些特征进行学习。然而在学习之前，我们还有一项重要的事没有做，那就是特征选择，又称特征工程。

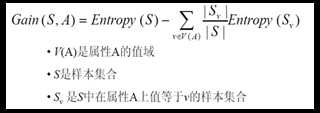
3.4.1为什么要特征选择

我们能用很多属性描述一名学生，例如加权平均分，借书次数，吃早饭次数等等，但有经验的人往往只需要看看他的加权平均分，和他一学期所修的总学分就可以大致知道他是否是一名学习优异的学生。换言之，对一个学习任务来说，给定属性集，其中有些属性很关键，另一些则可能没什么用，相关性弱的特征会给数据集带来冗余信息甚至噪声，干扰模型训练，对最终结果产生不利影响，同时过多的属性会造成工作冗余和资源浪费。通过特征选择，达到降维的目的，将多维的样本通过变换或者映射的途径降至低维空间，最终剔除相关性小的或者冗余度高的特征来再次降维。

3.4.2如何进行特征选择

特征选择算法有好多，例如过滤式选择和包裹式选择。[6]我采用了第一种方式。过滤式的主要思想是：通过给每一维的特征标识权重，即对每一维特征“评分”，评判出来的分数称为该特征的重要性，然后根据分数进行排序。这里的分数又称为信息增益，信息增益可以代表每个个特征对分类任务的作用大小。接下来的任务就是计算每个特征的信息增益[7]（见表3-3），增益的值越高，说明它和分类结果关系越大。

信息增益的数学公式：



每个特征的信息增益结果如表3-4

表 3‑4

|  |  |
| --- | --- |
| Feature | InfoGain |
| 上学期期末成绩 | 0.201 |
| 专业排名 | 0.082 |
| 专业类别 | 0.034 |
| 一学年吃早饭次数 | 0.120 |
| 一学期所修总学分 | 0.026 |
| 一学年入馆次数 | 0.009 |
| 一学年借书次数 | 0.031 |
| 生源地 | 0.097 |
| 加权平均分 | 0.163 |

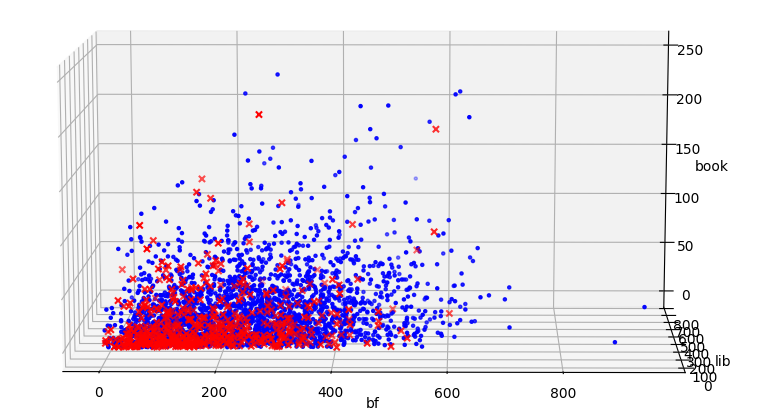


图 3‑10

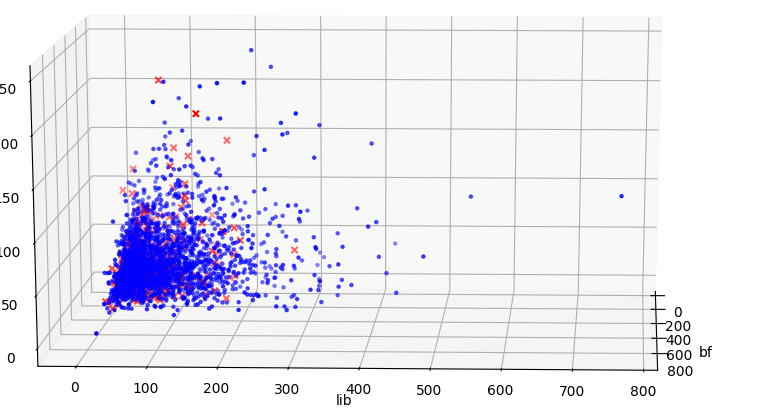


图 3-11

图3-10，3-11中红色的点代表落后生，蓝色的点代表普通生。这是一个三维图。

bf:一学年内吃早饭次数

book:一学年内借书次数

lib:一学年内入馆次数

从3-10图中可以大致看出，红色的点聚集的地方偏向于吃早饭次数，入馆次数，借书次数比较小的区域,图3-11中可以看出蓝色的点聚集在吃早饭次数比较多然而入馆次数和借书次数就很随机了。

1. 模型建立

4.1算法选择

选择合适的算法进行机器学习是最为关键的一步。判断一个分类器好坏的依据就是它的预测正确率（预测结果正确的个数 / 测试样本总数）[4]。在前面3.3.3节中，交代了交叉验证算法。将样本集合分为5个等数量子集，每次取其中4个集合为训练集，剩下一个作为测试集。取五次测试集正确率的平均值即为该算法模型的最终正确率。这种验证方法显然是需要大量的计算时间成本，但是它却是最好用的判断一个分类器好坏的方法之一。如果某个模型的结果不尽如人意，就继续进行算法选择。本论文探讨的问题是给定一个学生的一些平时的行为和学期末的成绩来判断他是落后生还是普通生，这是属于监督学习中的分类问题。分类算法也有好多。由于一开始我不能确定哪个算法的效果最好，所以只能选择目前学术界流行的几个分类算法进行依次实验。本文使用的机器学习算法包括K邻近算法，朴素贝叶斯算法，逻辑回归，支持向量机，决策树算法。在提前学习了这些算法的基本原理之后，利用scikit-learn库进行相应的建模操作。

4.2模型比较

4.2.1 KNN算法

算法流程：

1. ：计算当前学生在图上所对应的点与已知类别的点的距离
2. ：将各个学生按照距离递增的顺序排序
3. ：选取与当前学生点距离最相近的k个点
4. ：计算这k个点所在类别出现的频率，

（5）：返回前k个点出现频率最高的类别作为当前学生的预测分类

K邻近算法对于k的选取要求比较高，一旦k值小了，那么落后生的模型复杂度就会高，这会使我们的模型产生过拟合问题；如果k值大了，则模型简单，学习的近似误差就会增大。想要寻找合适的k值，就需要采用交叉验证法，反复实验。

从scikit-learn库中的sklearn.neighbors模块中导入KNeighborsClassifier建立knn模型。其中最重要的就是n\_neighbors参数，它代表算法中的k值需要交叉验证找到最为合适的值。（比较结果参照表4-1）

4.2.2 决策树（Decision Tree）

决策树顾名思义是一个树的结构。从根节点开始，每次判断学生的一个特征，根据判断结果来选择下降的子节点，递归该过程直到走到了叶子节点，得出分类结论。

决策树的工作流程一般划分为三个部分：

1. ：特征选择，步骤和3.4节介绍的一样，通过计算每个特征的信息增益
2. ：决策树生成，在这个步骤当中，我最关心如何选择最优划分属性集。通常来说，随着划分过程不断的深入，希望决策树包含的样本最大可能的属于同一种类别，即属于落后生或者普通生，这样就可以保证节点的“纯度”变高。
3. ：剪枝，主要是为了减少模型过拟合。在决策树学习过程当中，为尽可能正确分类学生样本，节点划分过程将不断的重复，这就形成了决策树分支过多的现象。

从scikit-learn库中导入tree模块，利用DecisionTreeClassifier函数生成分类器。（比较结果参照表4-1）

4.2.3逻辑回归（Logistic Regression）

算法流程：

1. ：寻找预测函数（线性函数）

X:学生特征向量

θ线性回归参数

1. ：构造代价函数[9]
2. 利用梯度下降算法使代价函数最小化并求得回归参数（θ）使用逻辑回归建立的落后生预警模型计算代价相比于其他算法要小的多，但是模型容易欠拟合，分类精度可能不高。

为解决过拟合带来的影响，分析过拟合问题往往源自过多的特征   
解决方法   
1）减少特征数量（减少特征会失去一些信息，即使特征选的很好）   
• 可用人工选择要保留的特征，例如信息增益值比较大的一些相关特征   
2）正则化   
• 保留之前经过特征选择章节的所有特征，但减少线性回归参数θ的大小

从scikit-learn库中导入sklearn.linear\_model模块，通过LogisticRegression函数构造分类器，最后得到逻辑回归模型的正确率。（比较结果参照表4-1）

4.2.4朴素贝叶斯（Naive Bayes）

朴素贝叶斯与概率论结合的比较紧密。朴素贝叶斯用的最广的就是邮件分类，一般人们用它来鉴别垃圾邮件。这里我也用贝叶斯算法来对学生分类。

朴素贝叶斯一般划分为三个流程：

第一步：准备工作，分为数据分析和数据处理工作，形成训练样本集。这一部分的任务主要是输入所有学生的信息（包括所有特征信息）的待分类数据，输出结果是训练样本和特征属性。这一步骤的完成情况对整个分类模型有着很重要的意义，这是因为分类器的好坏程度在很大程度上由特征属性、特征属性划分及训练样本质量决定。

第二步：训练工作，这一步骤是机器操作，根据朴素贝叶斯的算法机器自己完成生成分类器工作。这个步骤的任务就是通过学生的成绩，行为等数据生成分类器，主要工作是计算落后生和普通生在训练样本中的出现频率及每个特征属性划分对不同类别的条件概率，并将结果记录。

第三步：应用工作。在第二步骤的基础上我们已经产生了分类器，现有任务就是把学生测试集带入分类器中产生分类结果，最后计算分类器的正确率，评价该算法的优劣。

在实际操作中发现该模型对输入数据的准备方式较为敏感，所以尽量将训练集数据格式规范。

从scikit-learn库中导入sklearn.naive\_bayes 模块，利用GaussianNB函数构造分类器。（比较结果参照表4-1）

4.2.5 支持向量机（Support Vector Machines）

在现实生活中，SVM普遍应用在文本，图像的分类。由于我们的学生特征数量很多，所以必定是一个高维度的问题，因此SVM很适合我们的模型。SVM本身是一个二类分类器，所以非常适合我们落后生模型的建立。SVM的原理较为复杂，其中好多结合高等数学和概率统计的部分，推导过程也较为复杂，所以这里就不详细介绍学习过程，简单介绍一下如何利用sklearn库来实现操作。导入库中的svm模块，采用它的SVC分类器进行建模。（比较结果参照表4-1）

4.2.6模块持久化

经过前面章节的研究与实践，我们已经把落后生预警模型实现了，最后的工作就是如何把我们实现的模型给保存到硬盘上存储起来。这就是机器学习的最后一个步骤--模块持久化。同样，scikit-learn库也给了我们很好的算法支持，我们只需要调用该库中的函数便可以实现模块持久化。

导入sklearn.externals模块中的joblib对象，调用该对象的dump方法将上面实现的所有分类器保存在磁盘上以备之后的预测工作。

表 4‑1

|  |  |
| --- | --- |
| 分类器 | 正确率 |
| K邻近 | 76.59% |
| 决策树 | 86.23% |
| 逻辑回归 | 92.79% |
| 朴素贝叶斯 | 89.51% |
| 支持向量机 | 90.57% |

从表4-1看出来，五个分类器中，表现最好的是逻辑回归，表现最差的是k邻近算法。一般情况下，算法复杂度越大，对细节的识别能力越强，同时也容易得到过拟合的结果。所以没有绝对优秀的算法，这也是机器学习与其他研究不同之处。因为你无法得到一个准确的答案，可能每个模型的效果都差不多，也可能会有天壤之别，所以必须每个模型都做一下尝试才能选择适合落后生预警模型的算法。

4.3优化模型

在每个算法建立的分类器中，这些分类器都不是最优的，过多过少的会出现过拟合（能很好的拟合训练集，但不能很好的拟合测试集，又叫High Variance）和欠拟合（连训练集都不能很好的拟合，又叫High Bias）的问题。如图4.1所示。

图4.1的第一个曲线就是欠拟合情况，第二个是非常理想的情况，第三个就是欠拟合。

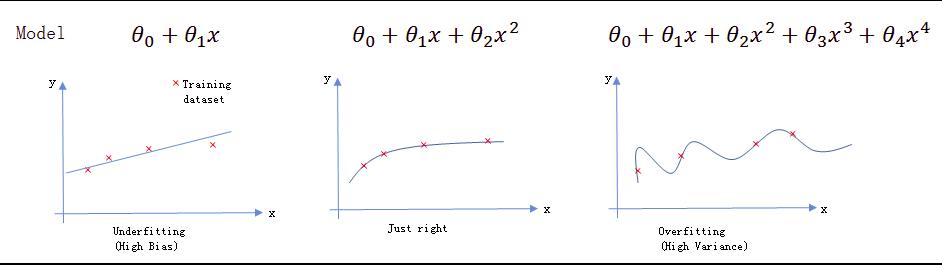


图 4‑1

通常为解决优化问题，有几个

（1）增加多项式元素（比如将特征平方后叠加到原特征上，相当于增加了非线性的输入）

（2）减少特征维度（从已有的特征中挑选出一部分）: 正如特征选择那章节所说，祛除相关性小的学生特征，保留重要的特征，如学生上学期的期末成绩。

（3）增加学生的训练集:在原始数据当中，不仅包括了2013级学生的数据，还包括了14,15级的，适当增加学生个数。

（4）增加新特征:有时还可以适当挖掘出一些新的特征

1. 总结

2013级学生的数量一共是3288名左右，在经过前期数据清洗之后，去掉那些信息不够完整的同学信息，最后只剩下了2967名完整学生信息。经统计其中符合落后生标准的达到了9.3%。可以看出问题的严重性，平均每十名学生当中就会有一名学生受到学业警示。其中统计数据表名了，部分专业，例如法学，机械设计制造及其自动化，通信工程，计算机科学与技术等，挂科率和学业警示率比其他专业高出好多，所以这些专业是重点预警对象，需要学校老师及辅导员重点照顾。

通过数据分析出来的一些特征当中，原本以为与学生成绩相关度很大的特征却 其实信息增益不是很大，最突出的就是学生的一年去图书馆次数。结合常识，大家可能会以为去图书馆越多的学生成绩可能相对要好一点，这没有错，但是那些去图书馆次数少的，甚至寥寥无几的学生未必成绩就会差。仔细分析一下在海大的四年时光，这是有依据的。由于中国海洋大学的崂山校区占地面积非常大，所以学生宿舍共分为三个宿舍区，东区，北区和南区，而学校图书馆的位置靠近南区宿舍，反而北区和东区宿舍离图书馆距离非常远，所以东，北区的学生偏爱去教学区上自习，而南区学生则首选图书馆，因此造成了部分学生的图书馆入馆次数远不及其他同学，但是他们的成绩依然很优秀。所以入馆次数在对学生进行分类时大部分只能起到积极的作用（即可以帮助判定该学生可能是优秀学生），但是不能代表一名学生是落后生。所以在特征选择的时候，要充分考虑到这种情况，使最终的模型更加精确。但是向其他一些特征，例如吃早饭次数受外界条件的影响就小一点。所以在讨论特征对结论的决定程度时候一定要充分考虑到实际情况对分类的影响。

在前面的章节介绍了，主流的分类器有k邻近，支持向量机，逻辑回归，朴素贝叶斯和决策树，这些分类器在特定的问题下表现会比较好，针对落后生预警模型，不同的分类器的分类效果是不同的。其中关键是不同算法的“归纳偏好”不同导致的。“奥喀姆剃刀”（Occam’s razor）[11]（切勿浪费较多东西去做用较少的东西同样可以做好的事情）是一种常用的，自然科学研究中最基本的原则。当特征是多维的时候，我们尽量要剔除那些作用很小的特征，这样不仅简化我们的模型，降低了计算的时间和空间复杂度，而且还能提高模型的精确度。因此学生的户籍信息来自城镇还是乡村对学生是否会被学业警示关联性不大，所以删除此类数据保留譬如学生一学年的加权平均分等重要信息。

最终，我选择Logistic Regression 的模型投入实际应用。考虑模型预测落后生的人数不宜过多，不然会超过教学资源的负荷，Logistic Regression会得到大约247人，准确率76.1%，具有很高的效率。虽然Logistic Regression是较简单的算法，但它的结果符合我们的要求，选择Logistic Regression也符合奥卡姆剃刀原则，所以选择Logistic Regression的模型应用到之后的工作中。

对于实际中的预测工作，首先收集相应数据，按特征定义构造数据集，然后运用选择的模型对学生进行分类，教育者们可以对预测出的落后生进行针对性的帮扶工作。

在未来的工作，会进行更深层次的研究。由学校制定的学业警示和留级标准确定类别，这种方式存在一定的问题，因为这些标准是人工制定的，在样本空间中不一定存在天然的边缘，限制了实验结果的上限，这启示教育工作者应该寻找更加合理的分类标准，这应该在未来的工作中占有一席之地。再者数据不一定会像研究中那样的理想化，比如图书馆并不是学生学习的唯一场地、餐厅也不是唯一的选择，需要增加收集数据的方式，使信息能更加全面。

1. 致谢

历时五个月，从论文选题到搜集资料，从[开题报告](https://www.baidu.com/s?wd=%E5%BC%80%E9%A2%98%E6%8A%A5%E5%91%8A&tn=SE_PcZhidaonwhc_ngpagmjz&fenlei=mv6quAkxTZn0IZRqIHckPjm4nH00T1d9PHRYnjbLPjbvmWPhnWRv0ZwV5Hcvrjm3rH6sPfKWUMw85NP7gdKWyh9GuA7EUMICmdq8uLK9uvdJX6K1TL0qnfK1TL0z5HD0IgF_5y9YIZ0lQzqlpA-bmyt8mh7GuZR8mvqVQL7dugPYpyq8Q1DdPHm1njmsPs" \t "/home/jokin/文档\\x/_blank)、写初稿到反复修改，期间经历了喜悦、聒噪、痛苦和彷徨，在写作论文的过程中心情是如此复杂。如今，伴随着这篇毕业论文的最终成稿，复杂的心情烟消云散，自己甚至还有一点成就感。  
 本课题在选题及研究过程中在得到王勇老师的亲切关怀和悉心指导下完成的。他严肃的科学态度，严谨的治学精神，精益求精的工作作风，深深地感染和激励着我。从课题的选择到项目的最终完成，王老师都始终给予我细心的指导和不懈的支持。王勇老师不仅在学业上给我以精心指导，同时还在思想、生活上给我以无微不至的关怀，在此谨向王老师致以诚挚的谢意和崇高的敬意。

1. 参考文献

[1]Campbell, J. P., DeBlois, P. B., & Oblinger, D. G. (2007). Academic analytics: a new tool for a new era. Educause Review, vol. 42, no. 4, pp. 40–57.

[2]Lackey, L. W., Lackey, W. J., Grady, H. M., & Davis, M. T. (2003). Efficacy of using a single, non-technical variable to predict the academic success of freshmen engineering students. Journal of Engineering Education, 92(1), 41e48.

[3]胡宣东.《规范化数据处理理论与应用》[C].中国高等学校自然科学学报研究会.科学编辑论文集(2).北京:北京师范大学出版社，2008:09-14

[4] S. B. Kotsiantis Supervised Machine Learning: A Review of Classification

Techniques[C]

[5]william wesley Mckinney . Python for Data Analysis Data Wrangling with Pandas, NumPy, and IPython [M] O'Reilly Media 2012:11-01

[6]周志华.机器学习[M].北京：清华大学出版社.2016:-1-1

[9]peter Harrington. Machine Learning in Action. Manning Publications.2012:4-19

[10]Xindong Wu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang ·Hiroshi Motoda · Geoffrey J. McLachlan · Angus Ng · Bing Liu · Philip S. Yu ·Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg Top 10 algorithms in data mining[J]

# [11]  罗邺编译. 奥卡姆剃刀[M] 北京民航图书发行部.2005:1-1