《人工智能与机器学习》

课程期末大作业

题 目 饮酒与在校表现的关系

姓 名 杜艳红

学 号 212019081000001

二○二○ 年 6 月 11 日

### 任务介绍

学生成绩好坏与多方面因素都有着关系，研究学生在校表现与那些因素有关。研究数据通过对中学数学和葡萄牙语课程的学生进行调查而获得的。它包含许多有关学生的有趣的社交，性别、学习信息，也包含饮酒量，家庭人数，参与课外活动等信息。具体数据集如下：

表格 1 数据集

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **age** | **address** | **famsize** | **Pstatus** | **Medu** |
| 15->22 | U | LE3 | T | 0->4 |
|  | R | GT3 | A |  |
| **failures** | **schoolsup** | **famsup** | **paid** | **activities** |
| 1->3 | yes/no | yes/no | yes/no | yes/no |
| **romantic** | **famrel** | **freetime** | **goout** | **Dalc** |
| yes/no | 1->5 | 1->5 | 1->5 | 1->5 |
| **reason** | **guardian** | **Fedu** | **nursery** | **absences** |
| home | mother | 0->4 | yes/no | 0->93 |
| reputation | father | **Walc** | **health** |  |
| course | other | 1->5 | 1->5 |  |
| other |  |  |  |  |

G1 – 第一学期成绩 (0->20)

G2 - 第二学期成绩 (0->20)

G3 – 最终成绩 (0->20)

通过上述数据集，分析出以下问题：

1. 喝酒会影响学生的成绩吗，如果不会，那有什么别的影响吗？
2. 尝试利用这些数据，挖掘学校成绩与各种因素之间的关系。
3. 看看是否可以根据其他变量来预测学生的成绩！

### 开发环境

本次实验基于python语言开发，版本对应为Python 3.6.3，使用PyCharm（2018.1.4）工具进行编写，第三方依赖库以导入包的方式给出：

1. **import** pandas as pd
2. **from** keras.models **import** Sequential
3. **from** keras.optimizers **import** Adam
4. **from** keras.layers **import** Dropout
5. **import** numpy as np
6. **from** keras.callbacks **import** EarlyStopping, TensorBoard, ModelCheckpoint
7. **from** keras.layers.core **import** Dense, Activation
8. **import** matplotlib.pyplot as plt

### 处理流程概述

对于“饮酒与在校表现的关系”的课题，处理流程图如下：



图 1 处理流程图

1. 数据处理

由于数据中有文字描述的参数，在做数据分析时很难使用，故将文字部分数据化，以进行下一步处理。数据化处理如表格2所示。

表格 2 参数数字化

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **文字** | **数字化** | **文字** | **数字化** | **文字** | **数字化** | **文字** | **数字化** |
| GP | 1 | F | 1 | T | 1 | U | 1 |
| MS | 2 | M | 2 | A | 2 | R | 2 |
| teacher | 1 | teacher | 1 | home | 1 | LE3 | 1 |
| health | 2 | health | 2 | reputation | 2 | GT3 | 2 |
| at\_home | 5 | at\_home | 5 | other | 4 | mother | 1 |
| other | 4 | other | 4 | yes | 1 | father | 2 |
|  |  |  |  | no | 2 | other | 4 |

G1 – 第一学期成绩 (0->20)

G2 - 第二学期成绩 (0->20)

G3 – 最终成绩 (0->20)

并且，其使用标签数据由第一学期成绩、第二学期成绩以及最终成绩的均值给出。

1. 求数据相关

参数之间是否有相关性是数据分析的前提，故将数据进行相关处理，如下图所示。

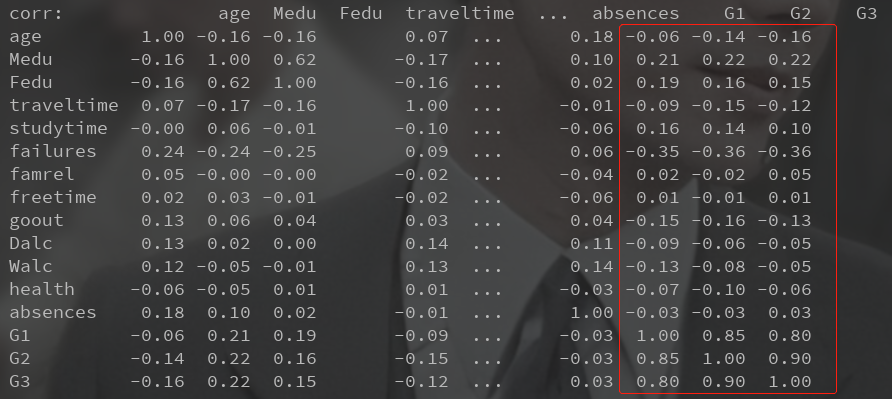


图 2 参数部分相关系数

可以看到成绩与failures呈现很强的负相关。故失败次数极大的影响到了一个学生在校的表现情况。而针对任务中的问题，喝酒会对学生的成绩有负的影响，但总体来说，影响不大，如图3所示，喝酒对一个学生的自由时间是正相关的，故喝酒的学生的空余自由时间会相对不喝酒的学生而言更少。

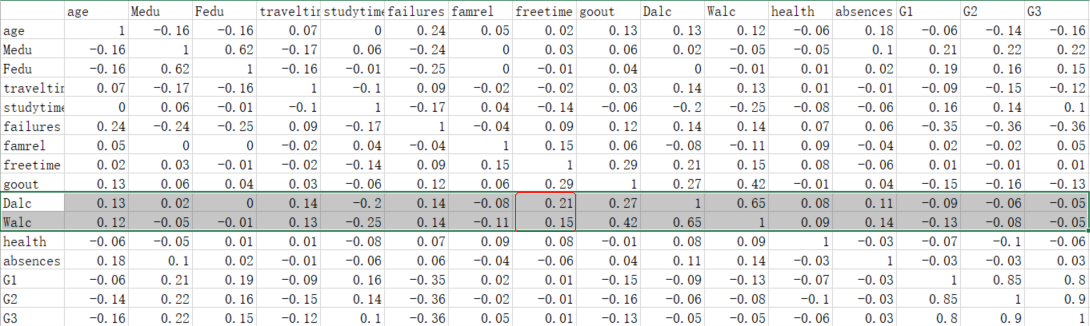


图 3 参数见的相关系数

1. 主成分分析

主成分分析法旨在将多个影响参数提取其中影响最大的因素，也可预测输出，但这里仅为分析学生成绩与各元素之前的关系。

用过主成分分析，得到各因素间的特征值分别为：3.2682，2.5414，1.7271，1.5784，1.4794，1.4196，1.3391，1.1629，1.1236，1.0859，1.0414，0.9991，0.9610，0.9286，0.8669，0.8046，0.7899，0.7427，0.7341，0.7020，0.6606，0.6253，0.5887，0.5667，0.5025，0.4843，0.4198，0.3531，0.2977，0.2052。

最终得到主成分回归方程为：

y=17.220951+-0.259419\*x1+0.090749\*x2+-0.210433\*x3+-0.040955\*x4+0.066354\*x5+0.003794\*x6+0.295266\*x7+0.284796\*x8+-0.230113\*x9+-0.219751\*x10+0.059567\*x11+-0.242279\*x12+-0.096326\*x13+0.062449\*x14+-0.362732\***x15**+-0.130088\*x16+-0.226946\*x17+-0.190473\*x18+-0.232649\*x19+-0.344863\*x20+-0.819955\*x21+-0.136088\*x22+0.220727\*x23+0.005350\*x24+-0.025834\*x25+-0.067750\*x26+-0.094306\*x27+-0.080328\*x28+0.027036\*x29+-0.014831\*x30

通过主成分回归方程可以看出，x15对成绩的影响最大，显现负相关（也就是failures），这与我们用相关系数分析得到的结果是一致的，其次x1（学校）,x7（母亲的教育水平）,x8（父亲的教育水平）,x20（参加幼儿园）同样对成绩也有较大的影响。

### 模型设计

针对第三个问题，设计神经网络模型对成绩进行预测，输入数据datasets\_251\_561\_student-por.csv作为训练数据，datasets\_251\_561\_student-mat.csv作为校验及测试数据。在前面说过其标签数据由第一学期成绩、第二学期成绩以及最终成绩的均值给出。

模型设计结构如下（仿真得出）。

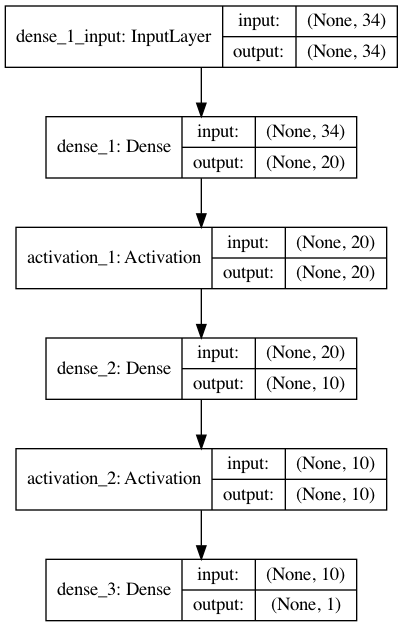


图 4 训练模型

### 模型训练

1. 模型的节点数不大，故dropout不易太大。
2. 网络层不是越多越好，多了有时反而学的不好
3. 激活函数影响也很大，在本任务中使用到了relu激活函数

### 模型评测

1. Train\_loss(mse)曲线：

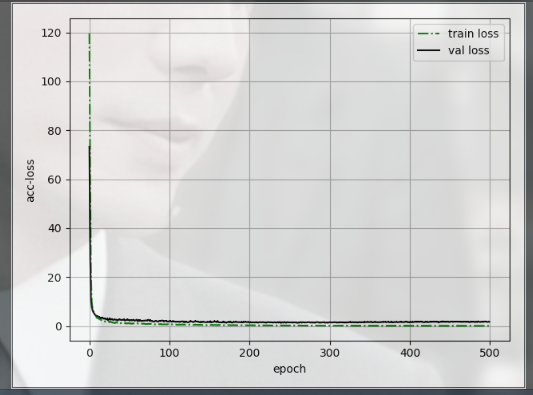


图 5 loss曲线

训练loss（MSE）： 0.0673

测试loss（MSE）：1.962847428985789

1. 预测结果

表格 3 预测结果与真实值对比

|  |  |
| --- | --- |
| 预测值 | 真实值 |
| 7.817593 | 5.666667 |
| 7.8377066 | 5.333333 |
| 9.573396 | 8.333333 |
| 14.570314 | 14.666667 |
| 9.950404 | 8.666667 |
| 14.639854 | 15 |
| 12.366623 | 11.666667 |
| 7.962138 | 5.666667 |
| 16.539501 | 17.666667 |
| 14.374212 | 14.666667 |
| 10.324086 | 9 |
| 12.083011 | 11.333333 |
| 13.965824 | 14 |
| 11.188502 | 10.333333 |
| 14.956823 | 15.333333 |
| 14.122387 | 14 |
| 13.664101 | 13.666667 |
| 10.515331 | 9.333333 |

### 问题总结

1. 在训练模型时，参数的调优是及其重要的，即使初始值也对模型的训练有着重要的作用，故在训练时，需建立在理论基础之上，设置模型的参数。
2. 训练数据与测试数据有所区别，一个为数学课程一个为葡萄牙语课程，故测试结果一定会存在泛化现象，这里需要改进。
3. 因为训练数据量过小，若训练次数过多会造成过拟合现象，故在训练中加入Dropout适当减少这种现象。

### 参考文献

1. 程颖涛. 基于深度学习的自然语言处理中问题分析的 研究[D]. 西安:西安邮电大学, 2018.
2. 周敬一, 郭燕, 丁友东. 基于深度学习的中文影评情感分 析[J]. 上海大学学报(自然科学版), 2018, 24(5):703-712. ZHOU Jingyi, GUO Yan, DING Youdong. Sentiment analysis of Chinese movie reviews based on deep learning[J]. Journal of Shanghai University(Natural Science Edition), 2018, 24(5):703-712.