

上机实验报告

|  |  |
| --- | --- |
| 题目： | 运用逻辑回归预测员工离职 |

|  |  |
| --- | --- |
| 姓 名： | **李梓焓** |
| 学 号： | **1700011735** |
| 组 别： | **无** |
| 实验日期： | **2019.09.29** |

**摘要** 本实验运用逻辑回归，对员工离职情况进行拟合，并给出相应的预测结果，然后计算准确度和AUC，从而得出结论：（1）是否离职与satisfaction\_level（满意度）呈明显的负相关关系，Work\_accident（是否发生事故）和salary （薪水高低）也与离职呈较为明显的负相关，其余变量有一定的负相关，但效应较弱或不太明显；另一方面，是否离职与last\_evaluation（最新考核评估）呈较为明显的正相关关系。（2）多次测试来看，逻辑模型对训练集和测试集的预测准确率大致接近0.80，且准确率波动不是很大，而AUC大致在0.57~0.70之间，且波动较大，超过0.10，从这一点看，预测结果差强人意，但提升空间不是很大，且稳定性有待商榷。

**关键词** 逻辑回归，分类器，预测

1. **引言**

逻辑回归虽然有“回归”字样，本质上却是一类分类算法。所谓分类，就是用一个（或若干个）超平面将数据剖分为两部分（或多个部分），每一部分均含有一定的特征。而逻辑回归的思路，就是建立起线性函数，然后用Logistic函数进行转换，从而得到一个范围在0~1之间的数值。如果大于0.5，属于分类1，反之属于分类0。

现在，我们面对的最大问题，便是求解最佳回归系数。与其他回归模型一样，逻辑回归也存在过拟合问题，因此，也需要引入正则化的方法避免过拟合，即通过添加惩罚项，从而对损失函数的某些参数进行限制，以提高模型的泛化能力。本次实验采用二次惩罚项，此时按Scikit-learn的定义，损失函数为：

只要求解出相应的和（或写做），就可以求出最优模型。常用的求解和方法为梯度下降法，计算过程中不断计算函数的梯度，由此推算出下一步迭代时和的变化量，最终得到损失函数的最小值和相应的和。

本次上机实验采用Pandas从csv文件读入数据，并采用自己的规则，将非数值变量转换为数值变量，然后Scikit-learn自带的LogisticRegressionCV，对样本进行逻辑回归和测试，并分别用accuracy\_score()和roc\_auc\_score()计算准确率和AUC。

**2 实验部分**

**2.1 仪器**

2.1.1硬件

Surface Pro（第5代，处理器参数：Intel® Core™ i5-7300U CPU @ 2.60GHz，2.71 GHz，2个内核，4个逻辑处理器；内存容量：8.00 GB）

2.2.2软件

操作系统：Windows 10家庭版，版本1903

开发环境：Visual Studio 2019 Community（已预装64位Miniconda 3，含64位Python 3.7.4、Conda 4.7.12、NumPy 1.16.5、Scikit-learn 0.21.2、Pandas 0.25.1）

2.2.3训练和测试数据

HR\_comma\_sep.csv: 含约1.5万员工的数据，读入后按1:4划分训练和测试数据。

**2.2 实验过程**

2.2.1训练和测试数据的读入

利用Pandas携带的read\_csv()方法，从HR\_comma\_sep.csv读入数据。其中，sales和salary这两列数据为非数值变量，需要我们寻找一个合适的转换方法。为此，我们对sales和salary这两列数据进行统计，结果表1、表2。

**表1** sales列数据分析

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| sales类型 | 总人数 | 离职数 | 离职比 | 离职比与平均离职比差值 | 最大绝对值归一化后差值 |
| accounting | 767 | 204 | 0.27 | 0.03 | -0.30 |
| hr | 739 | 215 | 0.29 | 0.05 | -0.56 |
| IT | 1227 | 273 | 0.22 | -0.02 | 0.17 |
| management | 630 | 91 | 0.14 | -0.09 | 1.00 |
| marketing | 858 | 203 | 0.24 | -0.00 | 0.02 |
| product\_mng | 902 | 198 | 0.22 | -0.02 | 0.20 |
| RandD | 787 | 121 | 0.15 | -0.08 | 0.90 |
| sales | 4140 | 1014 | 0.24 | 0.01 | -0.07 |
| support | 2229 | 555 | 0.25 | 0.01 | -0.12 |
| technical | 2720 | 697 | 0.26 | 0.02 | -0.19 |
| total | 14999 | 3571 | 0.24 | 0 | 0 |

**表2** salary列数据分析

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| salary类型 | 总人数 | 离职数 | 离职比 | 离职比与平均离职比的差值 | 最大绝对值归一化后差值 |
| low | 7316 | 2172 | 0.30 | 0.06 | -0.34 |
| medium | 6446 | 1317 | 0.20 | -0.03 | 0.20 |
| high | 1237 | 82 | 0.07 | -0.17 | 1.00 |
| total | 14999 | 3571 | 0.24 | 0 | 0 |

由表1、表2可知，sales和salary对离职率有较为明显的影响，例如management（管理）和RandD（研发）的离职率明显低于平均离职率，而hr（人力资源）的离职率明显高于平均离职率；又如低薪资的人员的离职率明显高于中薪资和高薪资的人员。因此，我们考虑用最大绝对值归一化后的离职比差值，作为sales和salary对应的数值。

接下来，我们用train\_test\_split()划分训练和测试数据，其中test\_size = 0.8，保证训练和测试数据按1:4划分；random\_state = 1551，保证每次输出的一致性，方便调试。

2.2.2训练逻辑回归模型并验证

使用Scikit-learn库中的LogisticRegressionCV（带交叉验证的逻辑回归），对训练数据采用逻辑回归，得到相应的逻辑回归模型，然后用该模型对训练数据进行预测，并计算准确率和AUC，最后将预测结果与训练集输出至dLogisticR2a.dat，一并输出的还有该模型对训练集预测的准确率和AUC。

2.2.3对测试集预测

采用2.2.2训练的逻辑回归模型，对测试数据进行预测，并计算准确率和AUC，然后将预测结果与训练集输出至dLogisticR2b.dat，一并输出的还有该模型对训练集预测的准确率和AUC。

2.2.4输出模型参数

将2.2.2训练的逻辑回归模型的参数输出到dLogisticR2c.dat，这些参数包括：正则化系数的倒数C、不同变量的（斜率）系数coef[i]、截距intercept。

2.2.5模型准确度和稳定性考察

为考察逻辑回归模型在不同训练集上的表现，在这一部分，我们在划分训练和测试数据时，不使用固定的随机种子，然后重复2.2.1-2.2.4的流程10次，每次均计算模型对训练集和测试集预测的准确率和AUC。

**3 数据与结果分析**

**3.1固定训练集和测试集后模型参数和预测效果**

random\_state设定为1551，按2.2.1-2.2.3进行实验，得到模型参数，如表3所示，其中coef[i]分别对应satisfaction\_level、last\_evaluation、number\_project、average\_montly\_hours、time\_spend\_company、Work\_accident、promotion\_last\_5years、sales、salary：

**表3** 训练后的逻辑模型参数

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| C | coef[1] | coef[2] | coef[3] | coef[4] | coef[5] | coef[6] | coef[7] | coef[8] | coef[9] | intercept |
| 8.07 | -2.09 | 0.50 | -0.15 | 0.00 | 0.13 | -0.71 | -0.31 | -0.29 | -0.61 | -0.08 |

由表3可知，是否离职与satisfaction\_level（满意度）呈明显的负相关关系，即满意度越高，越不容易离职，这说明员工对于工作的满意程度，极大影响员工未来是否考虑离职。此外Work\_accident（是否发生事故）和salary （薪水高低）也与离职呈较为明显的负相关，这可能是因为发生事故容易影响员工的身体状况，进而影响绩效，最终导致员工被“炒鱿鱼”，也可能是因为发生事故后对员工的处罚影响到员工对上级和对公司的心理印象，进而选择离职；而过低薪水既影响当前员工的物质生活水平，又有可能打击员工的工作积极性，导致对公司的满意度降低。promotion\_last\_5year（五年内是否升职）、sales（岗位）、number\_project（项目数）与离职之间存在负相关，但效应不明显。

另一方面，是否离职与last\_evaluation（最新考核评估）呈较为明显的正相关关系，即最新考核评估分越高，越容易离职，大概与能力较强、表现优秀的员工更希望通过跳槽获得更多的发展机会？而time\_spend\_company（工作年限）与离职有一定正相关，但效应不明显。average\_montly\_hours（平均每月工作时长）对离职的效应不明显，应进一步考察不同情景的工作时间与离职的关系（到底有多少时间是在完成项目，还是开会，抑或是摸鱼？）。

接下来我们比较一下训练集和测试集的准确率和AUC，结果如表4所示：

**表4** 训练集和测试集的准确率和AUC

|  |  |  |
| --- | --- | --- |
| 数据来源 | 准确率 | AUC |
| 训练集 | 0.7936 | 0.6412 |
| 测试集 | 0.7841 | 0.6279 |

由表4可知，训练出来的逻辑模型在训练集上的表现比在测试集上要好，其中准确率在训练集和测试集上均接近0.80，还算差强人意，但是仅靠修改正则化参数进一步优化模型，已经没有太大提升空间，更好的选择是分析各变量是否满足独立假设，以舍弃不必要的变量，以及采用更好的分类算法，如基于随机森林的分类算法等；而AUC在训练集上为0.6412，在测试集上为0.6279，略有下降。根据简书上的文章《让你彻底记住什么是ROC/AUC（看不懂你来找我）》（<https://www.jianshu.com/p/82903edb58dc>），AUC在0.5~0.7时预测效果较差，不过用于预测股票已经很不错。（不过我严重怀疑这样的模型用于实际还远远达不到要求……）

**3.2模型准确度和稳定性考察**

不设置random\_state，然后重复2.2.1-2.2.4的流程10次，每次均计算模型对训练集和测试集预测的准确率和AUC，结果如表5所示。

**表5** 随机重复测试的准确率和AUC

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 测试次数 | C | 训练集准确率 | 训练集AUC | 测试集准确率 | 测试集AUC |
| 1 | 0.5594 | 0.7883 | 0.6248 | 0.7779 | 0.6111 |
| 2 | 2.7826 | 0.7963 | 0.6441 | 0.7961 | 0.6416 |
| 3 | 256.0993 | 0.8049 | 0.6753 | 0.7971 | 0.6576 |
| 4 | 0.5594 | 0.8129 | 0.6912 | 0.8048 | 0.6776 |
| 5 | 14.6273 | 0.7876 | 0.6548 | 0.8028 | 0.6672 |
| 6 | 1.0865 | 0.7923 | 0.6471 | 0.7909 | 0.6489 |
| 7 | 3.4240 | 0.8013 | 0.6870 | 0.8066 | 0.6733 |
| 8 | 9.1403 | 0.7936 | 0.6395 | 0.7861 | 0.6306 |
| 9 | 0.0377 | 0.7893 | 0.5826 | 0.7731 | 0.5725 |
| 10 | 3.2849 | 0.8109 | 0.6898 | 0.8085 | 0.6791 |
| 平均值 | - | 0.7977 | 0.6536 | 0.7944 | 0.6460 |
| 极差 | - | 0.0253‬ | 0.1086‬ | 0.0354‬ | 0.1066‬ |
| 极差/平均值 |  | 0.0317 | 0.1662 | 0.0446 | 0.1650 |

由表5可知，逻辑模型对训练集的预测准确率大致在0.787~0.811之间，平均值0.7977，极差0.0253；AUC大致在0.582~0.692之间，平均值0.6536，极差0.1086。而对测试集的预测准确率大致在0.773~0.809之间，平均值0.7944，极差0.0354；AUC大致在0.572~0.680之间，平均值0.6460，极差0.1066。

以上数据说明：（1）该模型的预测准确率和AUC差强人意，但提升空间不是很大，想要提升至0.90，必须从优化数据和算法入手；（2）准确率绝对值变化不是很大，但与平均值相比，仍有一定波动；（3）AUC绝对值变化有点大，已超过0.10，说明逻辑模型的AUC受训练数据影响较大，稳定性有待商榷。

**3.3可能的改进方向**

笔者目前想到的可以提升分类准确度和AUC的措施：（1）对数据集的各变量进行独立性检验，剔除不必要的相关变量，避免相关变量对模型的影响；（2）逻辑回归常常出现欠拟合现象，因此考虑换用其他分类算法。

**3.4实验结论**

（1）是否离职与satisfaction\_level（满意度）呈明显的负相关关系，Work\_accident（是否发生事故）和salary （薪水高低）也与离职呈较为明显的负相关，其余变量有一定的负相关，但效应较弱或不太明显；另一方面，是否离职与last\_evaluation（最新考核评估）呈较为明显的正相关关系。

（2）多次测试来看，逻辑模型对训练集和测试集的预测准确率大致接近0.80，且准确率波动不是很大，而AUC大致在0.57~0.70之间，且波动较大，超过0.10，从这一点看，预测结果差强人意，但提升空间不是很大，且稳定性有待商榷。