优达学城数据分析师纳米学位项目 P5

安然提交开放式问题

说明：[你可以在这里下载此文档的英文版本](https://s3.cn-north-1.amazonaws.com.cn/static-documents/nd002/Enron+Submission+Free-Response+Questions.pdf)。

机器学习的一个重要部分就是明确你的分析过程，并有效地传达给他人。下面的问题将帮助我们理解你的决策过程及为你的项目提供反馈。请回答每个问题；每个问题的答案长度应为大概 1 到 2 段文字。如果你发现自己的答案过长，请看看是否可加以精简！

当评估员审查你的回答时，他或她将使用特定标准项清单来评估你的答案。下面是该标准的链接：[评估准则](https://review.udacity.com/?&_ga=1.206413269.797580181.1464528540#!/rubrics/310/view)。每个问题有一或多个关联的特定标准项，因此在提交答案前，请先查阅标准的相应部分。如果你的回答未满足所有标准点的期望，你将需要修改和重新提交项目。确保你的回答有足够的详细信息，使评估员能够理解你在进行数据分析时采取的每个步骤和思考过程。

提交回答后，你的导师将查看并对你的一个或多个答案提出几个更有针对性的后续问题。

注意；很详细的思路在 poi\_id.ipynb 中

我们期待看到你的项目成果！

1. 向我们总结此项目的目标以及机器学习对于实现此目标有何帮助。作为答案的部分，提供一些数据集背景信息以及这些信息如何用于回答项目问题。你在获得数据时它们是否包含任何异常值，你是如何进行处理的？【相关标准项：“数据探索”，“异常值调查”】

数据集中共有人数 **146** 人，其中涉嫌欺诈的有 **18** 人, 特征个数是 **20** 个。基于数据集具有以下特点，大致思路如下：

1. 这个数据集很不平衡（imbalance）, 也就说明accuracy并不是很好的评估指标，选择precision和recall更好一些。
2. 在交叉验证的时候，因为数据的不平衡性，我们会选用Stratified Shuffle Split的方式将数据分为验证集和测试集。
3. 数据样本比较少，因此我们可以使用GridSearchCV来进行参数调整。如果较大的数据则会花费较长的时间，可以考虑使用RandomizedSearchCV。

他们包含异常值，其中有个 TOTAL，其实是其他数据的总和，这个数据被我拿掉了。其中还有人叫 THE TRAVEL AGENCY IN THE PARK, 这显然不是个人名，应该删掉。

还有一个人 LOCKHART EUGENE E，所有特征都是NaN，这个数据完全没有意义，也应该删掉。

另外 数据量一共只有100多条，非常小，很多数据都是NaN，但经过分析，我最后还是没有拿掉，因为本来数据就少，拿掉部分数据之后，就更没什么数据了。所以还是在特征选择上多下了功夫。

1. 你最终在你的 POI 标识符中使用了什么特征，你使用了什么筛选过程来挑选它们？你是否需要进行任何缩放？为什么？作为任务的一部分，你应该尝试设计自己的特征，而非使用数据集中现成的——解释你尝试创建的特征及其基本原理。（你不一定要在最后的分析中使用它，而只设计并测试它）。在你的特征选择步骤，如果你使用了算法（如决策树），请也给出所使用特征的特征重要性；如果你使用了自动特征选择函数（如 SelectBest），请报告特征得分及你所选的参数值的原因。【相关标准项：“创建新特征”、“适当缩放特征”、“智能选择功能”】

我一开始筛选的特征：

['poi', 'bonus', 'total\_payments', 'restricted\_stock', 'long\_term\_incentive', 'from\_poi\_to\_this\_person', 'from\_this\_person\_to\_poi', 'other', 'expenses', 'exercised\_stock\_options', 'deferred\_income']

我使用决策树看每个特征重要性，然后筛选出了影响 > 0 的特征

然后我新增了特征，

'from\_this\_to\_poi\_ratio'：从此邮箱发出的邮件中，发到poi邮箱的占比

'from\_poi\_to\_this\_ratio'：从此邮箱收到的邮件中，来自poi邮箱的占比

'bonus\_times'：奖金/工资。相当于国内常说的奖金系数

然后，再次用决策筛选除了，有用的特征, 作为最终的模型

['poi', 'exercised\_stock\_options', 'expenses', 'from\_this\_to\_poi\_ratio', 'deferred\_income', 'long\_term\_incentive', 'bonus', 'restricted\_stock'] **其中 'from\_this\_to\_poi\_ratio' 是新增的特征**

在ipynb中，我分别使用新老特征，采用朴素贝叶斯的方式做了对比。使用老特征的f1分数是 0.321380952381 使用新添加的特征后的f1分数是 0.413857142857 可以看出，使用新添加的特征后，结果是好于老特征的。

关于缩放，我在Pipeline尝试过 MinMaxScalar，但是结果不如没有缩放的，于是最终模型时没有使用智能缩放的

1. 你最终使用了什么算法？你还尝试了其他什么算法？不同算法之间的模型性能有何差异？【相关标准项：“选择算法”】

我最终使用了 朴素贝叶斯算法。原因是因为他的结果最好。

我还分别尝试了 KNN, 决策树，SVM，AdaBoost，RandomForest

不过其中SVM因为特征多行，性能不够所以没有跑出结果来。

他们的模型大致区别如下：

朴素贝叶斯算法 是基于概率的模型

KNN 是每个点找相邻的N个点来分来

SVM是支持向量机，用一个内核算法去拟合

决策树，相关于自动根据每个特征写 if-else 来分类

AdaBoost和 RandomForest 是集成算法，集成各种普通的机器学习算法，他们的区别只是集成策略的不同。

1. 调整算法的参数是什么意思，如果你不这样做会发生什么？你是如何调整特定算法的参数的？（一些算法没有需要调整的参数 – 如果你选择的算法是这种情况，指明并简要解释对于你最终未选择的模型或需要参数调整的不同模型，例如决策树分类器，你会怎么做）。【相关标准项：“调整算法”】

我用GridSearchCV 对所有尝试的算法做了不同参数寻找最佳结果，具体尝试过程见poi\_id.ipynb

我在寻找最佳结果的指标是用f1 score, 因为

f1 = 2 precisionrecall / (precision + recall)

这样f1 score 同时考虑了 查准率precision 和 召回率recall

不过我最终选择是 朴素贝叶斯 本身没什么参数可调

1. 什么是验证，未正确执行情况下的典型错误是什么？你是如何验证你的分析的？【相关标准项：“验证策略”】

验证就是，我们经常会将数据集分为训练集（training set）跟测试集（testing set）这两个子集，前者用以建立模型（model），后者则用来评估该模型对未知样本进行预测时的精确度，正规的说法是泛化能力（generalization ability）。我们需要在测试集上进行验证来确定训练集是否“过拟合”。

未正确执行情况下的典型错误 是 过拟合，当训练集上的得分和测试集上的得分差距很大的时候，一般来说就可能发生了过拟合。

我在程序中使用的是 StratifiedShuffleSplit。这个的原理是，把数据集中，随机的拿出一部分，默认50%做训练集，随机拿出另外一部分，默认50%做测试集。然后评估结果。按照这个结果迭代N次，每次求出评估指标，最后再对每次的评估指标求平均。

1. 给出至少 2 个评估度量并说明每个的平均性能。解释对用简单的语言表明算法性能的度量的解读。【相关标准项：“评估度量的使用”】

本项目要求的3个评估度量是 查准率 precision 和 召回率 recall 和 accuracy

他们分别是

precision = TP / (TP + FP) = 真poi且判对 / 预测认为poi的人

也就是 **被识别为POI的数据中有多少是真正的POI**。

recall = TP / (TP + FN) = 真poi且判对 / 真实poi的人

也就是 **真实的pos数据中，有多少被识别出来了。**

accuracy = (TP+TN) / All = 判断正确 / 所有

其中

|  |  |  |
| --- | --- | --- |
| 真实情况 | 预测结果 | |
| 正例 | 反例 |
| 正例 | TP 真正poi且判对 | FN 真正poi但错判 |
| 反例 | FP 不是poi但误判是 | TN 不是poi且没误判 |

王闻宇

2017年12月