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。

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。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。

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”】

数据探索：该数据集中共有 146 个数据点，每个数据点有 21 个变量，其中 20 个特征（14 个财务特征，6 个邮件特征），1 个嫌疑人标签。有 POI 标识符的 18 个，特征 loan _advances 缺失值 144 个，为最大值缺失特征。

异常值调查：该数据集中存在一个 TOTAL 为键的异常值，LOCKHART, EUGENE E 的值全为空值，THE TRAVEL AGENCY IN THE PARK 不是人名，利用 pop()方法进行了删除。

2. 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”】

创建了新特征为 fraction_from_poi 和 fraction_to_poi，主要计算收到嫌疑人邮件数据在总收件数量中的占比、写给嫌疑人邮件数量在总发件数量中的占比。

因为特征中的值之间相差较大，因此采用了 MinMaxScaler 进行特征缩放。

利用 SelectKBest 方法选择了前五个值最大的特征分别为：

exercised_stock_options 25.10，total_stock_value 24.47，bonus 21.06，salary 18.58 fraction_to_poi 16.64

最终使用的特征为 exercised_stock_options，total_stock_value，bonus，salary，

fraction_to_poi。

3. 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”】

在算法选择方面共尝试了 5 中算法，分别是朴素贝叶斯、支持向量机、决策树、随机森林、KNN，其中支持向量机、随机森林效果不理想，代码中未体现。最终使用了朴树贝叶斯算法，因为它的 precision 和 recall 较高，且运行速度最快。

4. 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”】

调整算法参数主要是对各个算法的参数进行调整，以实现最优参数。本项目中利用了 GridSearchCV 方法对决策树、KNN 进行了参数调整，其中决策树最优参数是{'min_samples_split': 2,'max_leaf_nodes': None, 'max_depth': None, 'min_samples_leaf': 10}，KNN 最优参数是{'n_neighbors': 5}。

5. 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”】

验证就是将数据集划分为训练集和测试集，用训练集进行拟合，利用测试集进行验证，避免过拟合，确保评估分类器或者回归的性能。典型错误是训练集和测试集的折中问题，划分不合理，导致最终的测试集精确度低。

本项目采用了 StratifiedShuffleSplit 和 train_test_split 方法，根据 test_size 参数设置进行了训练集和测试集比例划分。两者的主要区别是前者可以保证样例在两种集中的比例一样，后者无法实现，但是通过设置参数 random_state=42 可以实现样例在两种集中的比例一致。

6. 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”】

本项目中数据集很不平衡，精确度（accuracy）不是一个好的评估指标，因此对测试集使用了准确率和召回率评估指标，准确率是预测为真且实际为真的个数与预测为真的个数的比值，召回率是预测为真实际为真的个数与实际为真的个数的比值。

朴树贝叶斯算法中被识别 POI 身份 4 个中有 2 个是真的，实际为 POI 身份的 5 个中，识别出来 2 个。

决策树算法中被识别 POI 身份 7 个中有 2 个是真的，实际为 POI 身份的 5 个中，识别出来 2 个。

KNN 算法中被识别 POI 身份 3 个中有 2 个是真的，实际为 POI 身份的 5 个中，识别出来 2 个。

利用 tester.py 测试性能，结果如下：

```
GaussianNB(priors=None)
        Accuracy: 0.85629       Precision: 0.49545      Recall: 0.32650 F1: 0.39
361     F2: 0.35040
        Total predictions: 14000        True positives:  653    False positives:
  665   False negatives: 1347   True negatives: 11335
```

优达学城

2017 年 8 月