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”】

数据一共有 146 条， 通过 matplotlib 作图发现有一个点 salary 特别高，查找并删除了该值’TOTAL’。 此外，删除了异常值‘THE TRAVEL AGENCY IN THE PARK’ （不是一个人）和 'LOCKHART EUGENE E' （所有特征全部为 NaN，没有有用信息）

Dataset 中一共有 22 个字段，其中一个是 poi，其余都是自变量特征，Poi 有 18 个，非 poi 有 125 个

有一些字段缺失严重，例如 total_payments 缺失 140 条，director_fees 缺失 127 条，restricted_stock_deferred 缺失 126 条。

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”】

朴素贝叶斯和决策树都不会受特征缩放的影响，所以没有选择对特征进行缩放。

创建新特征 ft_with_poi，含义是和 poi 来往邮件的总和，没有使用该特征时 Precision: 0.31179，Recall: 0.32800，使用该特征后 Precision: 0.31145 Recall: 0.32500，使用特征后得分反而降低了一些，因此不使用该特征。

采用 SelectPercentile 进行特征自动选择，特征得分如下

salary : 15.8060900874

deferral_payments : 0.00981944641905
total_payments : 8.962715501
loan_advances : 7.03793279819
bonus : 30.6522823057
restricted_stock_deferred : 0.679280338952
deferred_income : 8.49349703055
total_stock_value : 10.814634863
expenses : 4.31439557308
exercised_stock_options : 9.95616758208
other : 3.19668450433
long_term_incentive : 7.53452224003
restricted_stock : 8.051101897
director_fees : 1.64109792617
to_messages : 2.60677186644
from_poi_to_this_person : 4.93930363951
from_messages : 0.434625706635
from_this_person_to_poi : 0.105897968337
shared_receipt_with_poi : 10.6697373596
ft_with_poi : 2.64273164064

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”】

最终使用了最近中心法 NearestCentroid，尝试了朴素贝叶斯、支持向量机、决策树算法、最近中心法 NearestCentroid 和神经网络。

其中支持向量机的结果最差，弃用

朴素贝叶斯最终的得分：Precision: 0.22604　Recall: 0.39500（召回率较高，但精确度较低）

决策树最终的得分：Precision: 0.23203　Recall: 0.19050（较差）

最近中心法：Precision: 0.31179，Recall: 0.32800（符合要求）

支持向量机：没有命中

神经网络：Precision: 0.11856　Recall: 0.24950（较差）

3. 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”】

算法调整主要是对分类器的参数进行调节，从而更好地对数据进行拟合，优化分类器的性能，解决过拟合和欠拟合现象。决策树分类器如果不调整 min_samples_split 表现较差，调整该参数为 10，出现比较明显的改善，解决了过拟合的问题；最近中心法中将 shrink_threshold 调整到 1.5 左右性能最好。

4. 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”】

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，即泛化能力（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”。因

为测试集和训练集并没有被随机打乱而且数据中 poi 与非 poi 分配非常不均匀，因此采用了 StratifiedShuffleSplit 来验证，这个方法有 5 个参数 len(y), n_iter, test_size, train_size, random_state，分别表示样本总体数量，迭代次数，测试样本占比，训练样本占比和固定随机状态，固定随机状态如果不设置，那每次验证的结果可能会不一样。

5. 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”】

朴素贝叶斯最终的得分：Precision: 0.23647 Recall: 0.39750
决策树最终的得分：Precision: 0.24353 Recall: 0.20700
最近中心法：Precision: 0.31762 Recall: 0.33350
根据信号检测论的原理，测试结果一共有 4 种，正确找到了 poi，

	Poi	非 poi
找到	正确找到	虚报
没找到	漏报	正确拒绝

Precision 的含义是正确找到/（正确找到+虚报）

Recall 的含义是正确找到/（正确找到+漏报）

优达学城
2016 年 9 月