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。

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”】

此数据集包含146个人，其中POI共18人，非POI有128人，共有2920个数据点，20个特征。异常值有TOTAL（总数量）、THE TRAVEL AGENCY IN THE PARK（不是一个人） 和 LOCKHART EUGENE E（没有有用信息），我将其删去。

此外部分特征的值为NaN，也将其删去。

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”】

我最终使用了'exercised\_stock\_options', 'total\_stock\_value', 'bonus', 'salary', 'fraction\_to\_poi'五个特征。我设计了自己的三个特征，分别为'fraction\_from\_poi'，'fraction\_to\_poi'和'wealth'。通过用SelectKBest算法，得分最高的五个特征分别是'exercised\_stock\_options'（24.82）, 'total\_stock\_value'（24.18）, 'bonus'（20.79）, 'salary'（18.29）, 'fraction\_to\_poi'（16.41），而我构建的另外两个新特征得分是'fraction\_from\_poi'（3.13）和'wealth'（15.37），故选择上述五个特征。

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”】

我最终选择了GaussianNB，在过程中，我还尝试了DecisionTree、RandomForestClassifier、AdaBoost。我用精确率、召回率两个指标评估算法性能，最终GaussianNB均达到了0.3以上。

GaussianNB的精确率为0.42，召回率为0.31。此外，DecisionTree的精确率为0.20，召回率为0.22；RandomForestClassifier的精确率为0.33，召回率为0.17；AdaBoost的精确率为0.40，召回率为0.24。

在 tester.py中，GaussianNB 的各项指标分别为：Accuracy: 0.85629 Precision: 0.49545 Recall: 0.32650 F1: 0.39361 F2: 0.35040 Total predictions: 14000 True positives: 653 False positives: 665 False negatives: 1347 True negatives: 11335

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”】

我所尝试的算法如DecisionTree和AdaBoost有不同的参数，调整参数就是尝试这些参数不同的值，使算法的精确率、召回率等指标更高，更契合该数据集的实际情况。如果不调整参数可能会使决策树太深入导致过拟合等问题。我调整算法的方法是列出我认为可选的参数，用GridSearchCV选择最佳参数。

例如，在DecisionTree中，最终选择的参数是criterion='entropy', max\_depth=None, max\_leaf\_nodes=None, min\_samples\_leaf=10, min\_samples\_split=2；在AdaBoost中，最终选择的参数是algorithm='SAMME', learning\_rate=0.5, n\_estimators=10。

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”】

交叉验证是一种模型选择方法，其将样本的一部分用于训练，另一部分用于验证。典型错误是过拟合。

我用train\_test\_split() 将数据集拆分为训练集和测试集，用GridSearchCV 交叉验证，进行100次数据集的拆分与验证，取精确率和召回率的平均值，得到精确率、召回率和最佳参数。

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”】

我选择的两个评估度量是精确率和召回率。

精确率计算的是所有"正确被检索的"占所有"实际被检索到的"的比例。在此项目中，精确率是被判断为的POI人中，真正的 POI 所占的比例。

召回率计算的是所有"正确被检索的"占所有"应该检索到的"的比例。在此项目中，召回率是被判断为POI的真实POI占总POI的比例。