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”】

此数据集共包括146个人；数据集中有21个特征；总共有18个POI，占总人数的12%；

有一个特征缺省值非常多，一下即是超过100人都为空的特征值：

deferral\_payments : 107

loan\_advances : 142

restricted\_stock\_deferred : 128

director\_fees : 129

缺省值太多的特征值不建议用作机器学习算法；对于缺省值，使用了项目中自带的featureFormat函数中的remove\_NaN将所有缺省值标为0，以及remove\_all\_zeros去除所有特征值都为0的数据点；

通过利用散点图的形式，可视化一对特征，观察是否有异常值，最终发现total异常值，去除掉即可；

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”】  
   最终的POI标识使用了如下变量：['poi', 'total\_payments', 'expenses', 'exercised\_stock\_options', 'other', 'shared\_receipt\_with\_poi', 'percent\_to\_poi']；  
   通过利用SelectFromModel以及DecisionTreeClassifier进行特征挑选，依据的是决策树中的feature\_importances\_来进行的，selectfrommodel的阈值是根据尝试得出的，设置为0.05后最终会挑选出5个左右的特征值；  
   因为使用的决策树算法，所以并不需要进行特征缩放，特征进行缩放，决策树的分割顺序不会发生变化；  
   根据课程上的提醒以及观察散点图，创建了percent\_to\_poi, percent\_from\_poi两个特征值，用来判断一个人发给或者接收POI邮件的比例是否可以用来判断此人为POI；
2. 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”】  
   最终使用了决策树算法；还尝试了支持向量机(svm); SVM算法对于判断POI性能十分的差，个人猜测是因为没有进行特征缩放的原因，导致SVM非常困难地找到决策面。
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”】

调整算法的参数就是控制算法使用的各项值，以让机器学习的性能得到可控的提高，如果不这样做，有可能不能获得最优的性能，甚至有可能造成过拟合等；

通过利用GridSearchCV进行自动调参，首先先去sklearn上了解算法的参数含义，然后根据常识拟定参数的范围，最终使用gridsearchcv从中挑选中最优的参数。

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”】  
   验证就是测试训练好的机器学习算法在测试集中的表现，为正确执行的典型错误是泛化十分不好，造成了过拟合；  
   使用了kfold交叉验证算法的性能；
2. 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”】  
   使用了三个评估值：accuracy，recall，precision；  
   accuracy是预测正确的人占所有人数的比例；  
   precision是预测中真正为嫌疑人的人占所有预测为嫌疑人中所占的比例；  
   recall是预测中真正为嫌疑人的人占所有真正为嫌疑人的比例；

优达学城

2016年9月