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”】

此项目的目的是想从安然公司的邮件中找出欺诈的线索及证据，通过机器学习能有效提高查找线索的证据和准确性。此数据集包含了146个数据点，每个数据点代表一个人的邮件，每个数据点有20个特征。其中感兴趣的有18个人（POI），POI包括已经认罪的人、交了罚款尚未被刑事控告的人和换得豁免权的人。在这146个数据点中，由于存在缺失值，只有95人有工资记录，111人有邮件地址记录，不同的特征缺失值的数量也不相同。

通过对数据中的工资和奖金两个特征做回归分析，发现数据中存在异常值。其中最大的数据点的关键字是‘Total’，应该代表了总和，所以要删除此异常值，方法为*dictionary.pop( key, 0 )*，'THE TRAVEL AGENCY IN THE PARK'并不是一个人名，删除此异常值，'LOCKHART EUGENE E'虽然是人名，但是所有特征都是NaN，删除此异常值。另外还有4个数据点由于过高的奖金（大于5 百万美元）及过高的工资（大于100万美元）也属于异常值，但由于这些是有效的数据点，而且对于甄别欺诈也有帮助，因此不能删除。

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”】

首先将所有特征都选中，即，features\_list = ['poi','salary', 'deferral\_payments', 'total\_payments', 'loan\_advances', 'bonus', 'restricted\_stock\_deferred', 'deferred\_income', 'total\_stock\_value', 'expenses', 'exercised\_stock\_options', 'other', 'long\_term\_incentive', 'restricted\_stock', 'director\_fees','to\_messages', 'from\_poi\_to\_this\_person', 'from\_messages', 'from\_this\_person\_to\_poi', 'shared\_receipt\_with\_poi','fraction\_from\_poi','fraction\_to\_poi']。运行决策树算法，得出每个特征的重要性为：array([ 0. , 0. , 0. , 0. , 0. ,

0. , 0. , 0.08330184, 0.17014983, 0.40804035,

0. , 0. , 0.04565957, 0. , 0. ,

0. , 0.11832326, 0. , 0. , 0. ,

0.17452516])

选择其中重要性不为0的6个特征，即'total\_stock\_value', 'expenses', 'exercised\_stock\_options', 'restricted\_stock', 'from\_messages','fraction\_to\_poi'。再次运行决策树算法，得出每个特征的重要性依次为[ 0.07851991, 0.41964445, 0., 0., 0., 0.50183564]。

其中'fraction\_to\_poi'为新创建的特征， 'fraction\_to\_poi'表示发给POI的邮件占所发邮件的比例。可以看到新建特征‘fraction to poi’重要性超过了50%。

这些特征是通过决策树来筛选的，由于各个特征之间数量差异很大，数据不平衡，如果不进行缩放那么较大的数量就会占主导。因此进行了MinMax缩放。

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”】

最终使用了决策树算法。还尝试了朴素贝叶斯算法，朴素贝叶斯算法得出的结果为： Precision: 0.44472 Recall: 0.26950

，决策树算法得出的结论为Precision: 0.44444 Recall: 0.53200。

可以看出，朴素贝叶斯算法虽然在精确度上性能不错，但是召回率却很低，而决策树算法在保证精确度不错的情况下，召回率也很高。。

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”】

决策树分类器，调整min\_samples\_split参数，默认的min\_samples\_split=2，意思是当决策树的节点叶子数小于2时，不可再分。当样本数量比较大时，过小的min\_samples\_split易引起过拟合，对训练集的模拟效果好，但是对测试集的预测效果却不好。本次的数据集不足200，相对较小，使用决策树算法，并用 GridSearchCV 进行参数调整，它可以自动调整参数来匹配最优结果。如下：

DecisionTreeClassifier(class\_weight=None, criterion='gini', max\_depth=2,

max\_features=None, max\_leaf\_nodes=None, min\_samples\_leaf=8,

min\_samples\_split=2, min\_weight\_fraction\_leaf=0.0,

presort=False, random\_state=None, splitter='best')

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”】

验证是通过把数据集分为训练集和测试集来验证模型的效果，如果不进行验证容易出现过拟合。由于此项目数据不平衡，因此项目中使用了交叉验证，用Stratified Shuffle Split的方式将数据分为验证集和测试集。。

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”】

Precision: 0.44444 Recall: 0.53200

精确度为0.44444，召回率为0.53200。也就是当测试集出现POI时，模型有0.44444的概率识别此POI，但同时模型有0.53200的概率将非POI标记为POI。

优达学城

2016年9月