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！

• 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”】

1.2000年，安然公司是美国最大的公司之一,主营美国能源、大宗商品和服务。2002年,安然公司因欺诈申请破产,。在联邦调查中, 发现了在安然的158名员工有超过600000封电子邮件。该数据集被广泛用于机器学习的研究。在这个项目中,我做了一个机器学习模型使用“scikit-learn”算法预测嫌疑人poi,在安然可能参与欺诈的人。

学习这个项目可以很清楚从处理数据到训练模型到预测分类的准确度整个机器学习算法的流程都可以掌握，数据集有3066数据点,其中1708数据点是有效的。共146个人其中18人是poi，有19个特征加POI特征加email\_address特征

异常值1)无效的一个人的名字: THE TRAVEL AGENCY IN THE PARK', 'TOTAL”这两个是无效的名字。;2)无效特征，restricted\_stock\_deferred，director\_fees，存在大量缺失值，为无效特征，从特征中删除这两项

一共有3066个数据，其中1708个有效，1358个为NaN无效， 所以这个数据集有很多缺失值，也就说明accuracy并不是很好的评估指标，选择precision和recall更好一些。

2.数据样本比较少，因此我们可以使用GridSearchCV来进行参数调整，如果较大的数据则会花费较长的时间，可以考虑使用RandomizedSearchCV

在验证方式中我选择了**train\_test\_split** 来随机划分测试集和训练集，他是交叉验证中常用的函数，是样本中随机的按比例选取测试数据和训练数据。

我使用了grid\_search来调整参数

• 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”】

我最终选择的特征是'poi', 'exercised\_stock\_options', 'total\_stock\_value', 'bonus', 'salary', 'total\_asset', 'deferred\_income', 'long\_term\_incentive', 'restricted\_stock', 'total\_payments', 'shared\_receipt\_with\_poi'

首先有两个特征的大部分值都为，NaN我认为这属于无效特征予以删除，

然后，通过对数据的观察以及思考创建了新的特征，是有关于钱和邮件来往两方面。之后，我对特征进行了缩放，因为不同特征权重不同，之后用k-best进行特征选择，经过反复迭代，最终选择十个特征

我使用了除了目标标签poi和无关信息邮件地址的其他所有特征，进行了特征缩放

• 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”】

最终使用了朴素贝叶斯算法， svm等，总体还是朴素贝叶斯的召回率更高，性能也就更好一点

• 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”】

算法调整的意义是对分类器的参数进行调节，来更好地对数据进行拟合，从而优化分类器的性能。

• 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”】

如果没有执行验证，那可能该算法在当前数据上表现好，但预测表现不好，泛华能力差，或者出现过拟合的情况。而验证过程中，通过对测试集的测试更全面的说明算法的误区，不容易忽略这种问题

通过分训练集合测试集，来验证数据预测的准确性

• 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”】

我使用了三个评估度量来说明算法的性能：Accuracy， Precision，Recall

准确率Accuracy是指在分类中，使用测试集对模型进行分类，分类正确的记录个数占总记录个数的比例

精确率是指分类器分类正确的正样本的个数占该分类器所有分类为正样本个数的比例针对预测结果而言，预测为poi的就是poi来衡量他的准确率Accuracy。

召回率是指分类器分类正确的正样本个数占所有的正样本个数的比例。

召回率是针对原来的样本而言，原本是poi的被正确预测了Recall

##### 在不同的数据中，所代表的含义有所不同，一般情况下准确率越高，性能越好，但是本数据中数据很不平衡，所以应该更多关注于精确率-召回率

。

。

优达学城

2016年9月