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”】

这个项目的目标是，借助机器学习的力量，通过分析 enron mail dataset 的内容来判断出谁参与了欺诈（即谁被标记为 poi）。

#added region

本次的数据集中一共有 146 个数据点，其中 poi18 人，非 poi128 人。最初使用的特征数量以及他们各自的缺省值数量和使用与否，如下表所示

Feature	Sum of NaN	Use or not
salary	51	Yes
to_messages	60	No
deferral_payments	107	No
total_payments	21	Yes
loan_advances	142	No
bonus	64	Yes
email_address	35	No
restricted_stock_deferred	128	No
total_stock_value	20	Yes
shared_receipt_with_poi	60	Yes
long_term_incentive	80	Yes
exercised_stock_options	44	Yes
from_messages	60	No
other	53	Yes
from_poi_to_this_person	60	Yes
from_this_person_to_poi	60	Yes
poi	0	Yes
deferred_income	97	Yes
expenses	51	Yes
restricted_stock	36	Yes
director_fees	129	No

我最初选取了所有的非字符串的特征作为研究对象，同时排出了缺省值较多的‘loan_advances’，‘restricted_stock_deferred’，‘director_fees’，‘deferral_payments’。

#end region

我在获得数据的时候，通过视频和习题得知，由于电子表格的问题，“TOTAL”被误添加进来，所以将其删除。此外，通过对缺省值的分析，我将缺失 90%以上数据的用户的数据进行了丢弃。因为这类数据，不论放在 train 或者 test 里都不会有太大价值。

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”】

我最终使用了 SelectKBest 和 f_classif 来筛选了财务特征和与 poi 有关的邮件特征，来选取 Score 最高的 3 个特征。 在测试 pca+svm 和 svm 算法的时候，我对特征进行了缩放，因为这会影响算法最后的结果。我曾经测试了通过计算每封邮件等值的 total_stock_value，salary 来判断是否是 poi，但是由于缺省值过多。最终没能成功测试。

#添加 reigon
我创建了一个新的 feature，这个 feature 是通过
将 total_stock_value 除以 from_this_person_to_poi 得到的。想法是，股份价值是否和他练习 poi 的次数有关系。然后我将这个 feature 放入了数据集中，进行 SelectedKBest, 从 selector 的 score 中得到了以下结果。从下表可知，新建的 feature 在整体的中等水平。

FeatureName	Score	Ranking
exercised_stock_options	10.75989913	1
total_stock_value	10.26836392	2
bonus	9.499521308	3
salary	9.119320798	4
shared_receipt_with_poi	5.883381149	5
expenses	4.508595915	6
new_feature	4.398545103	7
deferred_income	4.390494169	8
from_poi_to_this_person	3.75568075	9
long_term_incentive	2.086122777	10
total_payments	2.068272592	11
restricted_stock	1.845204796	12
from_this_person_to_poi	1.703818174	13
other	7.88E-08	14

关于为何选择 3 个特征，我对选择不同特征个数时的 GaussianNB 的得分进行了如下的汇总。我们可以发现在使用 3 个特征的时候，最高的召回率和第二高的精确度。

Feature used	Accuracy	Precision	Recall
exercised_stock_options	0.90409	0.46055	0.32100
exercised_stock_options, total_stock_value	0.84069	0.46889	0.26750
exercised_stock_options, total_stock_value, bonus	0.84300	0.48581	0.35100
exercised_stock_options, total_stock_value, bonus , salary	0.84677	0.50312	0.32300
exercised_stock_options, total_stock_value, bonus, salary, shared_receipt_with_poi	0.84214	0.42391	0.29250
exercised_stock_options,total_stock_value, bonus , salary, shared_receipt_with_poi, expenses	0.84771	0.45359	0.32250

对于 svm 为什么需要特征缩放的问题
由于特征缩放会使得数据点在 plot 的时候的位置发生改变，而正巧这种改变会影响数据点到分界线之间的由距离影响的数值(比如说，距离的平方和)，所以特征缩放会影响 svm 模型同样的，决策树这类分类器的判断依据是是否达到某一条件从而产生分支，所以距离对其的影响就不是很大。

#end reigon

3. 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”】

我最终使用了 Gaussian Naive Bayes 算法来处理 enron mail dataset。同时我也尝试了 DecisionTree，RandomTree，PCA+SVM，SVM。
这其中，GaussianNB 具有最短的预测时间，RandomTree 具有最高的精确度（0.64）但是召回率只有可怜的 0.169，DecisionTree 虽然精确度超过了 0.3，但是召回率只有 0.103。PCA+SVM 和 SVM 的算法的 fit 时间都远远超过了预期。

4. 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”】

调整算法得参数就是指在初始化分类器的时候，给予分类器合适的参数，使之更有效的对 data 进行预测。如果不进行参数调整，所有的分类器都将以 default 参数预测，会导致分类器发挥不了应该有的功能。主要表现为精确度低下等等。
对于特定的算法，我使用 GridSearchCV 来自动分析最合适的参数，由于我这次选择的 GaussianNB 并没有特殊的初始参数，所以没对其进行参数调整。但是我在尝试 DecisionTree，和 RandomTree 的时候，都通过 GridSearchCV 对他们的 max_depth 进行了优化。

5. 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”】

验证就是对分类器的性能进行评估的一个过程。未能正确执行验证的情况下的典型错误是，~~无法客观的判断出当前算法对于新数据得有效性和性能。~~

#添加 reigon

在本次项目中 True Positive(TP)就是指被认为是 poi 嫌疑人的真的 poi 嫌疑人，而 False Positive(FP)就是指被判定为 poi 嫌疑人的“好人”。Accuracy 在本次项目中不是一个合理的指标是因为，本身就是少数的 poi 嫌疑人，在一个很高的 FP 的情况下，也能得到很高的 accuracy，但这并不代表着分类器就没问题。

谢谢审核的指出，了解了未进行验证的典型错误是过拟合。

特别是在一些特征过多的复杂环境下，计算机可能会给出一些很”完美”的结果，虽然很符合当前的数据集，但是并不能泛用于其他的数据，所以就需要进行验证。

Tester.py 中使用的 stratified，Shufflesplit 算法，结合了 KFold 和 ShuffleSplit，更大程度上制造出了更”随机”的样本，来对分类器进行分析。

#end reigon

我采用了 KFlod 的策略，将数据分割成了 4 份来交叉验证我的算法以及分析。

6. 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 tester.py 得到的精确度和召回率来评估算法的性能。通过查找资料，我知道了。精确度代表了正确被检测到的（TP）占实际被检测到的比例（TP+FP），而召回率代表了正确被检测到的（TP）占应该检测到的（TP+FN）比例。这 2 个值都是越高代表分类器越好。

优达学城

2016 年 9 月