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

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”】

本次项目的目标是建立一个预测模型，通过对安然的邮件数据集中的各个职员的多个变量进行分析来确认哪些人是我们在寻找的 POI 人物（Person of Interest）也就是参与到了安然欺诈丑闻中的嫌疑人。安然数据集是一个字典类型数据，包含了 146 条记录，'salary', 'poi', 'total_payments' 等 20 个特征值以及'poi'标签，通过对这些特征值的分析发现了 18 个 POI。

通过对数据集的信息整理发现了 1323 个的'NaN'数据，并通过自定义 countNa()函数，计算出每一个 feature 值中的‘NaN’值的数量，观察发现特征'loan_advances'的'NaN'值达到了 142 个，'restricted_stock_deferred'的空缺值达到 128 个，'director_fees'的空缺值达到 129 个，由于总的数据记录只有 146 个，可以看出以上的三个特征包含信息量较少，可能并不是一个好的特征，后期的特征选择应予以注意。在获得数据后进一步观察发现了三个异常值，分别是'THE TRAVEL AGENCY IN THE PARK'，'LOCKHART EUGENE E'，这两条数据包含的'NaN'值过多，不属于有效的数据点，其中'THE TRAVEL AGENCY IN THE PARK' 数据明显不能表示一个人，'LOCKHART EUGENE E'中包含的‘NaN’值过多予以删除处理。同时对于数值明显异常的 TOTAL 数据点进行了删除处理。'TOTAL'数据点是各条数据的汇总值，数值过大，属于异常值予以删除。

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”】

我通过自定义了一个 get_email_information 函数来自行创建了 fraction_poi_emails 特征作 POI 标识符，在 text learning 的课程中提到了 poi 成员之间会有比较密集的邮件通信，根据这一信息添加一个 fraction_poi_emails 变量来间接判断一个成员是否为 poi 值。

通过决策树算法来检验新特性 fraction_poi_emails 对于算法特性的影响，发现新特性的使用提高了算法的表现，其中 precision 值由 0.333 提高到 0.667, recall 值由 0.125 提高到 0.25。说明新特征有利于准确预测 poi 值。

```
Accuracy: 0.879310344828
Precision:  0.666666666667
Recall:  0.25
Decision Tree algorithm run time:  0.004 s
Feature Ranking:
1 feature salary (0.337437907714)
2 feature fraction_poi_emails (0.223706186287)
3 feature from_poi_to_this_person (0.161410018553)
4 feature from_this_person_to_poi (0.101688311688)
5 feature to_messages (0.100432900433)
6 feature deferral_payments (0.0753246753247)
7 feature total_payments (0.0)
8 feature exercised_stock_options (0.0)
9 feature bonus (0.0)
10 feature restricted_stock (0.0)
```

```
Accuracy: 0.844827586207
Precision:  0.333333333333
Recall:  0.125
Decision Tree algorithm run time:  0.003 s
Feature Ranking:
1 feature salary (0.337437907714)
2 feature from_poi_to_this_person (0.161410018553)
3 feature from_this_person_to_poi (0.109563164109)
4 feature to_messages (0.101688311688)
5 feature deferral_payments (0.0753246753247)
6 feature total_payments (0.0753246753247)
7 feature exercised_stock_options (0.0564935064935)
8 feature bonus (0.048961038961)
9 feature restricted_stock (0.0337967018319)
10 feature shared_receipt_with_poi (0.0)
```

在逻辑回归算法中，新特征的使用降低了算法的表现，accuracy 值由 0.724 减少到 0.690，precision 值由 0.1 减少到 0.0833。

```
Accuracy: 0.724137931034
Precision:  0.1
Recall:  0.125
F1 score: 0.111111111111
Logistic regression algorithm run time:  0.014 s
```

```
Accuracy: 0.689655172414
Precision:  0.0833333333333
Recall:  0.125
F1 score: 0.1
Logistic regression algorithm run time:  0.015 s
```

在特征选择步骤中，我使用了 SelectKBest 函数筛选出了得分最高的前六位特征值来作为我最后的标识符： ['exercised_stock_options'（得分：33400），'bonus'（得分：503），'expenses'（得分：inf），'from_messages'（592），'director_fees'（4445），'deferred_income（得分：1846）']等特征作为我最后的标识符。

对于支持向量机算法，进行了特征缩放处理 得到了 rescaled_value 数据列表，并应用特征缩放后的值发现提高了 accuracy 值，特征缩放对于算法性能起到了优化的效果。

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”】

我最终使用了决策树算法，因为决策树算法的 precision 值和 recall 值的表现更好。我还尝试了逻辑回归算法和支持向量机算法。
各个算法的性能如下

算法名称	决策树	逻辑回归	支持向量机
Accuracy	0.7413	0.793	0.655
Precision	0.2308	0.25	0.167
Recall	0.375	0.25	0.375

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”】

调整算法的参数就是为了提高算法的准确率，如果不进行调试会使得算法的准确率很低并且运行速度更慢。我没有选择对算法的特定参数进行调整，而是选择了 SelectKBest 算法来筛选对于最终结果影响更大的数据集特征来提高算法的准确率。如果选择调整算法参数，对于 SVM 可以选择优化 C 参数，对于决策树算法可以选择调整 min_samples_split, max_depth 等参数来优化。

同时我通过使用 GridSearchCV 进行参数调整，通过多次运行之后，确定的参数为 class_weight=None, criterion='gini', max_depth=7,max_features=3, max_leaf_nodes=None, min_samples_leaf=1,min_samples_split=2,min_weight_fraction_leaf=0.0,presort=False, random_state=None, splitter='best'将最后的 DT 算法中的参数调整为 GridSearchCV 所得到的结果。并经过多次运行检验，得到的 Precision, Recall, 以及 F1 值均大于 0.3， 参数调节达到要求。

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”】

验证就是评估机器学习预测模型的可靠性，未正确执行的情况下的典型错误是模型的 overfit， 对同一数据集进行了训练和检测会导致得到的准确率过高，不能够准确反映模型的可靠性。我的验证方法是将数据集分为训练数据集（60%的数据点）和测试数据集（40%的数据点）两部分分别用于模型的训练和测试。
StratifiedShuffleSplit 函数结合了 StratifiedKFold 和 ShuffleSplit 两个函数的功能，按照固定的 train/test 数据集比例随机对数据集进行抽样和分集，并且保证 poi/非 poi 的数据比例与整体数据集中 poi/非 poi 数据比例一致，避免出现将所有的 poi 都划分到训练集或是测试集中的情况，从而保证在执行 fit/validate 过程中的准确性。

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”】

准确率评分：表示被机器学习模型预测为 POI 的数据中实际值也为 POI 的比率，例如评分为 0.5，则表示所有被预测为 POI 的数据点有 50%的点实际就是 POI 值。

召回率评分：所有值中被正确的预测出来的点所占的比例，例如召回率为 0.5 表示，所有的 POI 值中有 50%被准确的预测。召回率不考虑预测值的多少，也就是多余的预测，而只考虑了应当被预测的值是否被预测出来，我对于召回率的理解是其只关注了收益量，而没有看到成本。

作为验证我们应该结合召回率和准确率两个变量共同判断最后的算法性能。

优达学城
2016 年 9 月