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”】

此项目目的在于找出在安然事件中有欺诈嫌疑的安然雇员，机器学习可以构建一个算法，利用公开的安然财务数据和邮件数据集找出嫌疑人。
数据集背景信息如下

- 数据点总数 146 个
- POI 类有数据点 18 个，非 POI 类有数据点 128 个
- 使用的特征数量共 21 个，分别是 salary, to_messages, deferral_payments, total_payments, exercised_stock_options, bonus, restricted_stock, shared_receipt_with_poi, restricted_stock_deferred, total_stock_value, expenses, loan_advances, from_messages, other, from_this_person_to_poi, poi, director_fees, deferred_income, long_term_incentive, email_address, from_poi_to_this_person
- 有很多缺失值的特征值共 6 个，分别是 deferral_payments: 107, loan_advances: 142, restricted_stock_deferred: 128, long_term_incentive: 80, deferred_income: 97, director_fees: 129

选取 salary 和 bonus 两个特征，3 个数据点奖金大于 5 百万，工资大于 1 百万，分别是‘TOTAL’, ‘SKILLING JEFFREY K’ 和 ‘LAY KENNETH L’，其中’TOTAL’表示合计值应该删除，使用列表的 pop 方法，其余两个数据点应保留。观察 PDF 文件，’TRAVEL AGENCY IN THE PARK’ 不是人名，该数据点不表示一个人的数据，也应该删除。

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”】

创建新特征： poi_index，由发出邮件中发往 poi 比例、收到邮件中来自 poi 比例和收到邮件中与 poi 共享邮件的比例三个特征值进行 PCA 得到。
注：发往 poi 比例由 from_this_person_to_poi/from_messages 计算得出，来自 poi 比例由 from_poi_to_this_person/to_messages 计算得出，poi 共享邮件比例由 shared_receipt_with_poi/to_messages 得出。三个特征都是创建的新特征，poi_index 由三个新特征进行 PCA 得出，并未与以前的特征有线性的关系。
理由：邮件特征中，由于每个人收发邮件数量差别较大，需要把与 poi 有关的 3 个特征由绝对数量转换成比例，同时，由于三个特征都是与邮件收发有关，可以使用 PCA 压缩成 1 个特征使后续识别算法效果更好。为了在 tester.py 文件中使用这一特征，需要将计算出的新特征写回 my_dataset 列表进而写入 my_dataset.pkl 文件

选择财务特征：对财务特征使用 SelectKBest 函数，可以发现财务特征的特征得分最大的 4 个特征得分均在 20 左右，明显大于其它特征得分，因此选取这 4 个财务特征，分别是 exercised_stock_options, total_stock_value, bonus, salary，特征得分依次为 25.1，24.5，21.1，18.6

特征缩放：选取的不同特征的取值范围差异较大，因此需要进行特征缩放，同样将缩放结果写回 my_dataset 列表。

新特征对最终算法性能影响：最终算法采用 GaussianNB 分类器。当使用新特征时，test.py 运行得出算法的精确度为 0.478，召回率为 0.336；不使用新特征时，精确度为 0.432，召回率为 0.336，新特征对算法精确度有提升，因此将其包含在最终特征集内。

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”】

最终使用了 GaussianNB 算法，尝试的其它算法及运行 test.py 测试性能结果如下

算法	精确度	召回率
GaussianNB()	0.478	0.336
DecisionTreeClassifier()	0.297	0.301
AdaBoostClassifier()	0.378	0.288
SVC(kernel=’linear’)	1	0.020

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”】

选择决策树进行参数调整。DecisionTreeClassifier 中 min_samples_split 参数表示一个节点如果想继续分割，其中必须包含的最少样本数，默认值为 2。

这个参数可以控制何时停止样本分类，防止过拟合，提升算法效果。

测试过程中使用 GridSearchCV 进行了参数调整， min_samples_split 取值依次为 2, 5, 10, 15：

```
from sklearn.model_selection import GridSearchCV
param_grid = {'min_samples_split': [2, 5, 10, 15]}
clf = GridSearchCV(DecisionTreeClassifier(), param_grid)
```

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”】

验证是将已知数据集分成训练集和测试集两部分，将训练集用于训练算法模型，用测试集进行测试。防止出现典型错误，即测试算法模型性能时只采用用于训练的数据进行测试，无法测试模型用于新数据的性能。导致无法检测出模型对训练数据性能好，而对新数据性能差，这种典型问题。

我使用 train_test_split 将数据分成训练集和测试集进行性能验证，并且运行 test.py 进行交叉验证。其中，test.py 中的 VSratifiedShuffleSplit 是将数据集按照 poi 标记的值分层随机划分，得到训练集和测试集。因为这个数据集中标记为 1 的数据点较少，随机划分训练集和测试集可能会导致集合中标记为 1 的数据点没有或太少，影响模型的训练和检测效果。

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”】

精确度：算法标记 poi 为真的数据点中，poi 实际为真的数据点的比例。即算法识别出来的有欺诈嫌疑的安然雇员中，有多大比例是真正有欺诈嫌疑的雇员。

召回率：poi 实际为真的数据点中被算法标记为 poi 为真的比例。即实际的有欺诈嫌疑的安然雇员中，有多大比例被算法识别为有欺诈嫌疑。