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”】

   此项目的目标是找到嫌疑人，机器学习可以通过特征学习，选择和判断来完成嫌疑人和非嫌疑人的识别，在数据集中，包括 146 个安然员工的数据，其中有 18 个的“POI”值为 1，即有 18 个确定的嫌疑人。每个数据点（人）具有 20 个特征，标签值为 poi，其中“deferred_income”，“ director_fees”和“loan_advances”等特征具有很多缺失（NaN）值。同时，在 salary 和 bonus 的数据中，发现了异常值 ‘TOTAL’(是所有财务数据的汇总)，‘TRAVEL AGENCY IN THE PARK’（不是一个人）和 'LOCKHART EUGENE E'（所有特征全部为 NaN，没有有用信息），移除，还有一些值得注意但不应该移除的数据点，如 SKILLING JEFFREY K 和 LAY KENNETH L 的财务数据远远高于正常值。

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”】

   初步预测，选择了“poi”,“bonus”，“salary”，“from_this_person_to_poi”和“from_poi_to_this_person”等特征来进行分析。

   设计了新特征，“from_poi_ratio”和“to_poi_ratio”即数据点来自 poi 的邮件在全部收到的邮件里面的占比和数据点发给 poi 的邮件在全部发出邮件中的占比，由于每个数据点收发的邮件数量可能有差别，所以相比绝对值，占比更能说明问题。加入新特征之前的算法得分如下：
   Accuracy: 0.70691 Precision: 0.19091 Recall: 0.18900 F1: 0.18995 F2: 0.18938
   加入新特征之后的算法得分如下：
   Accuracy: 0.74918 Precision: 0.30667 Recall: 0.30100 F1: 0.30381 F2: 0.30212
   由此可知，加入新特征后能明显改善算法得分。
   在分析过程中进行了特征缩放，原因是选择的特征之中，“bonus”，“salary”与其他特征明显不是个数量级，在进行算法分析时，不进行特征缩放会影响结果。

   利用决策树进行特征选择，经过多次测试，用各特征重要性和 SelectFromModel 选定了“bonus”，“salary”和“to_poi_ratio”三个特征，各特征重要性得分和算法得分如下：

[0.3234133 0.21382685 0. 0.13626437 0.22738835 0.09910714]
Accuracy: 0.76000 Precision: 0.33740 Recall: 0.33200 F1: 0.33468 F2: 0.33307

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 DecisionTree 算法，还尝试了 GaussianNB，KMeans 和 SVC 算法，各个算法得分如下：
DecisionTree:
Accuracy: 0.76045 Precision: 0.33506 Recall: 0.32250
F1: 0.32866 F2: 0.32494

GaussianNB (priors=None)
Accuracy: 0.79364 Precision: 0.36253 Recall: 0.17800
F1: 0.23877 F2: 0.19817

KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,n_clusters=2, n_init=10, n_jobs=1, precompute_distances='auto',random_state=None, tol=0.0001, verbose=0)
Accuracy: 0.58218 Precision: 0.22512 Recall: 0.53150
F1: 0.31627 F2: 0.41778

SVC 算法速度过慢。

分析得出，对于本次分析目标，DecisiongTree 算法的拟合性相对较强。

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”】

参数是指算法运行计算时使用的相关系数和方法，根据分析对象选择系数可以让算法按照分析者的目标进行运算，减少运行时间，提高算法效率，更快得出结果。如果不调整系数，则影响算法效率，有时会得出差别非常大的结果。通过不同系数的组合，多次测试结果，以结果来选择最有系数。在本次完成项目的过程中，使用了 GridSearchCV 调整了算法的参数。

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”】

在机器学习中，我们经常会将数据集分为训练集（training set）跟测试集（testing set）这两个子集，前者用以建立模型（model），后者则用来评估该模型对未知样本进行预测

时的精确度，正规的说法是泛化能力（generalization ability）。我们需要在测试集上进行验证，来确定训练集是否“过拟合”。

验证就是通过测试集反复测试训练集得出的模式来验证算法的正确性，若未正确执行，则可能由于训练集和测试集分类不当，造成无法相互验证，不能防止训练集的“过拟合”。该项目是通过 train_test_split(特征选择时)和 StratifiedShuffleSplit(算法验证时)来验证分析的。

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”】

首先来看算法最终的评估数据：

Precision: 0.33506 Recall: 0.32250

Total predictions: 11000

True positives(正确结果): 645 False positives(意外结果): 1280

False negatives(缺少结果): 1355 True negatives(正确的没有结果): 7720

精确率是针对我们预测结果而言的，它表示的是预测为正的样本中有多少是真正的正样本。即 645/（645+1280）=0.33506，在本项目中指，被识别为 poi 的样本（645+1280）中有 645 个是真正的 poi。

召回率是针对我们原来的样本而言的，它表示的是样本中的正例有多少被预测正确了。即 645/(645+1355)=0.32250，在本项目中指在所有的 poi 样本（645+1355）中，有 645 个被识别出来了。