优达学城数据分析师纳米学位项目 P5

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”】

   该项目的目标是根据公开的安然财务和电子邮件数据集，找到有欺诈嫌疑的安然员工。通过探索数据集可知：
   - 初始数据集有 146 数据点，样本量较少
   - 对于每个人有 20 个可用特征，1 个 POI 标签
   - E+F 数据集中有 18 个 POI，数据集很不平衡
   - poi_names.txt 文件中有 35 个 names
   - 缺失率超过 50% 的特征有 'deferral_payments','restricted_stock_deferred', 'loan_advances', 'director_fees', 'deferred_income', 'long_term_incentive'
   - 根据财务数据找到异常值 TOTAL,
   - 所有值都是 NAN 的异常值 LOCKHART EUGENE E,
   - 错误的数据点 THE TRAVEL AGENCY IN THE PARK

   对于缺失率超过 50%的特征以及出现异常值的数据均采取丢弃策略

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”】

   - 通过单变量特征选择 SelectKBest，基于 f_classif 得分，选出了得分前四的特征：'exercised_stock_options', 'total_stock_value', 'bonus', 'salary'，其得分分别是：25.098, 24.468, 21.060, 18.576。其余特征分值均在 11 分以下且多数特征缺失率过高，所以只选择了得分排名前四的特征。
   - KNN、SVM 使用了特征缩放
   - 尝试建立了’bonus_salary_ratio’特征，观察奖金与薪水的比例与 POI 之间的关系。在单变量特征选择过程中，由于其得分较低，所以并没有在最终模型中使用。

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”】

   - 分别尝试了默认参数情况下的：朴素贝叶斯、决策树、支持向量机，KNN，使用 precision、recall 评价模型性能
   - 默认参数下，朴素贝叶斯的效果最好，precision=0.425, recall=0.31

- 最终决定使用朴素贝叶斯

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”】

- 最终选取的算法朴素贝叶斯没有参数调整，因此调整参数部分的练习是在 KNN 算法上完成的，该部分代码紧随朴素贝叶斯之后。为了直接运行能得到结果，该部分代码已被注释
- 分类器不仅是算法，而是算法加参数，调整算法的参数以获得分类器或者回归的最佳性能
- 使用 GridSearchCV 调整了 KNN 的 weights 和 n_neightbors 参数，其中 weights 值包括 distiance 和 uniform,n_neightbors 值包括 5,10,15
- GridSearchCV 得出的最佳参数设置为 weights=distance, n_neightbors=5

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”】

- 验证是使用其他数据集去测试模型，未正确执行情况下的典型错误是过拟合。
- 由于数据集很不平衡，因此采用分层随机分割交叉验证的方式来验证模型的性能

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”】
   - Precision，精确度，0.425
   - Recall, 召回率， 0.31
   - 在该项目任务背景下，测量对象为所有员工。当 POI=1，代表该员工是嫌疑人，记为正例；当 POI=0，代表该员工不是嫌疑人，记为负例；基于此有以下关于选定性能指标的解释。
   - Precision 指模型判定的所有正例中，正确的正例的占比
   - Recall 指所有存在的正例中，被模型判定为正例的占比

优达学城
2018 年 3 月