从安然公司邮件中发现欺诈证据

-by WuXuhua-

项目简介

安然曾是 2000 年美国最大的公司之一。2002 年，由于其存在大量的企业欺诈行为，这个昔日的大集团土崩瓦解。 在随后联邦进行的调查过程中，大量有代表性的保密信息进入了公众的视线，包括成千上万涉及高管的邮件和详细的财务数据。在这个项目中，我将根据安然丑闻中公开的财务和邮件数据来构建嫌疑人识别符。

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”】

此项目的目的是利用安然数据建立一个模型，用它来将某个个人归类成嫌疑人或非嫌疑人。以下是数据概述：

- 数据点总数： 146
- 嫌疑人总数：18
- 非嫌疑人总数：128

数据集中可用的特征数为：21，使用的特征数量为：20

每个特征的缺失值：

特征	缺失值数量
salary	51
to_messages	60
deferral_payments	107
total_payments	21
long_term_incentive	80
loan_advances	142
bonus	64
restricted_stock	36
restricted_stock_deferred	128
total_stock_value	20
shared_receipt_with_poi	60
from_poi_to_this_person	60

exercised_stock_options	44
from_messages	60
other	53
from_this_person_to_poi	60
deferred_income	97
expenses	51
email_address	35
director_fees	129

查看散点图，我们会发现图中包含一个异常值，进一步查找，异常值归属于名字叫"TOTAL"的个体，这应该是输入时的错误导致的，错误的将总数也包括在数据内了，所以我将这个异常值删除。然后，我发现" THE TRAVEL AGENCY IN THE PARK"不是人名，" LOCKHART EUGENE E"的所有属性值都是"NaN", 将他们一起删除。此后，没有发现其他异常值。

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"】

根据直觉，我认为嫌疑人有较多的可能会互发邮件，因此我建立了两个新的特征，分别叫做 ratio_of_msg_from_poi 和 ratio_of_msg_to_poi，分别表示从嫌疑人收到邮件占总收到邮件的比例和发送给嫌疑人占总发送邮件的比例。

算上我添加的两个数据集中总共有 22 个特征，我不打算全部都使用这些特征，因此我使用 SelectKBest 类来获取每个特征的得分，并保留得分最高的 8 个特征，来构建我的最终预测模型。下表是这些模型和得分：

特征	SelectKBest 得分
exercised_stock_options	25.097541528735491
total_stock_value	24.467654047526398
bonus	21.060001707536571
salary	18.575703268041785
ratio_of_msg_to_poi	16.641707070468989
deferred_income	11.595547659730601
long_term_incentive	10.072454529369441
restricted_stock	9.3467007910514877

从得分来看，ratio_of_msg_to_poi 的得分较高，ratio_of_msg_from_poi 的得分并没有到达我的期望。因此我打算在把 ratio_of_msg_to_poi 特征包含在模型中，并测试添加和不添加 ratio_of_msg_from_poi 特征时对模型的影响。

同时，在这些特征的值的范围变化很大，如果不对他们进行特征缩放，在某些机器学习

算法(如 SVM)中，对结果会有很大影响。因此我使用 MinMaxScaler 来对特征进行缩放。

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”】

一个理想的机器学习算法应该具有较高的 accuracy、precision 和 recall 值，这是我选择算法时使用的标准。我测试了 3 个算法：朴素贝叶斯、支持向量机和决策树。从 accuracy 和 precision 得分来看，支持向量机比其他算法表现略好，但是考虑到 recall 得分低于 0.3，我还是决定使用朴素贝叶斯算法。同时，我发现添加自定义的 ratio_of_msg_to_poi 特征后，算法得分略有下降，因此我决定不将 ratio_of_msg_to_poi 特征包含在最终的模型中。

以下是各算法的得分：

算法	accuracy	precision	recall	最优参数
朴素贝叶斯	0.847	0.409	0.374	
支持向量机	0.855	0.441	0.291	C = 1000.0, gamma = 0.0001
决策树	0.838	0.255	0.22	min_samples_split = 20, max_leaf_nodes = None, criterion = 'gini', max_depth = None, min_samples_leaf = 5

添加 ratio_of_msg_to_poi 特征后的得分：

算法	accuracy	precision	recall	最优参数
朴素贝叶斯	0.843	0.396	0.373	
支持向量机	0.841	0.376	0.295	C = 1000.0, gamma = 0.0001
决策树	0.837	0.250	0.207	min_samples_split = 20, max_leaf_nodes = None, criterion = 'gini', max_depth = None, min_samples_leaf = 5

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”】

调整参数的意思就是对于某种机器学习算法尝试使用不同的参数，以期望找出最佳参数组合，使机器算法达到最佳性能。我使用 sklearn 提供的 GridSearchCV 类来找到各种算法的最佳参数组合。

- 朴素贝叶斯是最简单的算法，不需要任何参数。
- 支持向量机：C 设置成 1000，gamma 设置成 0.0001
- 决策树：min_samples_split 设置成 20, max_leaf_nodes 设置成 None, criterion 设置成 'gini', max_depth 设置成 None, min_samples_leaf 设置成 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”】

验证是一种模型验证技术，用来评估模型对于数据的泛化能力。机器学习中最容易犯的错误就是过拟合，一个过拟合的模型在训练数据集上往往会有非常好的表现，但是到了测试数据集上，得分会有很大的下降。考虑到这些问题，我使用了 sklearn 提供的交叉验证技术，对三种算法进行循环测试，并取每次得分的平均分来进行观察。

我采用的是 K 折交叉验证技术，将我们的数据平均分成 k 个子集，使用 k-1 个子集来训练数据，留下最后一个子集作为测试数据。对每个子集都做相同的事情，最后平均计算每个子集的得分作为最后得分。

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”】

我选择了 3 个评估度量，acurracy、 precision 和 recall 值，平均得分请参见算法得分表。是最直观的性能测量指标，就是正确预测的观察值和总的观察值的比例。对于结果比较成比例的数据集，acurracy 是很好的预测指标。测试结果，我的模型 accuracy 值是 0.846，意味着我的模型有 84.6%的准确率。

但是，安然数据集中的嫌疑人比例占很小的一部分，所以我们还要看其他两个指标。precision 表示的是正确预测的正面的观察值和总的预测正面观察值的比例，也就是说被预测的嫌疑人中，有多少真的是嫌疑人，模型的值是 0.4537，也就是 45.37%的精确率。

最后的指标是 recall 值。recall 的意思是正确预测的正面的观察值和实际所有正面观察值，就是说所有嫌疑人中，模型预测到了几个。模型的值是 0.365。

参考资料

http://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures

https://en.wikipedia.org/wiki/Hyperparameter_(machine_learning)

https://machinelearningmastery.com/how-to-tune-algorithm-parameters-with-scikit-learn/

https://datascience.stackexchange.com/questions/10773/how-does-selectkbest-work

https://towardsdatascience.com/train-test-split-and-cross-validation-in-python-80b61beca4b6