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

说明：你可以在这里下载此文档的英文版本。

机器学习的一个重要部分就是明确你的分析过程，并有效地传达给他人。下面的问题将帮助我们理解你的决策过程及为你的项目提供反馈。请回答每个问题；每个问题的答案长度应为大概 1 到 2 段文字。如果你发现自己的答案过长，请看看是否可加以精简！

当评估员审查你的回答时，他或她将使用特定标准项清单来评估你的答案。下面是该标准的链接：评估准则。每个问题有一或多个关联的特定标准项，因此在提交答案前，请先查阅标准的相应部分。如果你的回答未满足所有标准点的期望，你将需要修改和重新提交项目。确保你的回答有足够的详细信息，使评估员能够理解你在进行数据分析时采取的每个步骤和思考过程。

提交回答后，你的导师将查看并对你的一个或多个答案提出几个更有针对性的后续问题。

我们期待看到你的项目成果！

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”】

a) 该项目的目标是运用机器学习构建一个算法，通过公开的安然财务和邮件数据集，找出有欺诈嫌疑的安然雇员.因为该数据集已经给出了财务邮件特征和嫌疑的标签,因此这是一个监督学习算法.

b) 该数据集中包含 146 个数据点,其中有 18 个是嫌疑人.使用的特征数量为 20 个,具有 1358 个缺失值.

```
以下特征含有缺失值:
('loan_advances', 142)
('director_fees', 129)
('restricted_stock_deferred', 128)
('deferral_payments', 107)
('deferred_income', 97)
('long_term_incentive', 80)
('bonus', 64)
('to_messages', 60)
('from_poi_to_this_person', 60)
('from_messages', 60)
('from_this_person_to_poi', 60)
('shared_receipt_with_poi', 60)
('other', 53)
('salary', 51)
('expenses', 51)
('exercised_stock_options', 44)
('restricted stock', 36)
```

c) 异常值有 3 个,分别为总计,一个非人名,一个全部特征为空. 使用 pop 删除.

```
#确定财务数据中的异常值，并解释如何消除或以其他方式处理它们。
data_dict.pop("TOTAL",0)
data_dict.pop("THE TRAVEL AGENCY IN THE PARK",0)
data_dict.pop('LOCKHART EUGENE E',0)
print u'删除以下异常值:\nTOTAL\nTHE TRAVEL AGENCY IN THE PARK\nLOCKHART EUGENE E'
print '*'*100
```

```
删除以下异常值:
TOTAL
THE TRAVEL AGENCY IN THE PARK
LOCKHART EUGENE E
```

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”】

a) 最终在 POI 标识符中使用了以下 6 个特征:

```
('bonus', 20.792252047181535)
('salary', 18.289684043404513)
('total_net_worth', 11.458476579280369)
('exercised_stock_options', 6.0941733106389453)
('deferred_income', 1.6463411294420076)
('total_stock_value', 0.22461127473600989)
```

从这里可以看到, 最佳特征组合不一定都是得分高的特征.

b) 首先利用 SelectBest 选择 10 个得分最高的特征, 带入算法求精确度和召回率, 然后逐步减少特征数量, 得出 k=6 时效果最好.

使用SelectKBest尝试不同的特征组合(k=10~1)，并记录了每种组合的性能

```
#测试选择6个最佳特征,各分类器得分
features_selected=[]
clf = SelectKBest(f_classif, k=6)
selected_features = clf.fit_transform(features, labels)
for i in clf.get_support(indices=True):
    features_selected.append(features_list[i+1])
features_list_6 = ['poi']+features_selected

print u'默认参数的朴素贝叶斯分类器最高得分为使用6个最佳特征:'
print u'朴素贝叶斯:'
test_classifier(NB,my_dataset,features_list_6,folds = 1000)

print '*'*100
```

```
选择6个最佳特征,朴素贝叶斯分类器得分最高:
朴素贝叶斯:
GaussianNB(priors=None)
        Accuracy: 0.85250       Precision: 0.47866      Recall: 0.36450 F1: 0.41385
```

c) 没有进行特征缩放,因为数量级差别不大.

d) 创建新特征 total_net_worth,该特征是 total_payments 与 total_stock_values 之和.理由是该特征是所有收入的总和.

e) 测试新特征对最终算法性能的影响

```
朴素贝叶斯:
GaussianNB(priors=None)
        Accuracy: 0.83700       Precision: 0.64159      Recall: 0.23450 F1: 0.34346
```

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”】

a) 最终算法为 k 近邻分类器:

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=4, n_neighbors=2, p=2,
           weights='distance')
        Accuracy: 0.84546       Precision: 0.49715      Recall: 0.39250 F1: 0.43867
```

b) 还尝试使用了朴素贝叶斯和决策树:

```
默认参数的朴素贝叶斯分类器最高得分为使用6个最佳特征:
朴素贝叶斯:
GaussianNB(priors=None)
        Accuracy: 0.85250       Precision: 0.47866      Recall: 0.36450 F1: 0.41385
        Total predictions: 14000        True positives:  729    False positives:  794
s: 11206

********************************************************************************
默认参数的决策树分类器最高得分为使用3个最佳特征:
决策树:
DecisionTreeClassifier(class_weight=None, criterion='gini', max_depth=None,
            max_features=None, max_leaf_nodes=None,
            min_impurity_split=1e-07, min_samples_leaf=1,
            min_samples_split=2, min_weight_fraction_leaf=0.0,
            presort=False, random_state=None, splitter='best')
        Accuracy: 0.80200       Precision: 0.36538      Recall: 0.38950 F1: 0.37706
        Total predictions: 13000        True positives:  779    False positives: 1353
s: 9647
```

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”】

a) 机器学习的模型都是参数化的，以便于其针对特定的问题进行调整。一个模型有很多参数，寻找这些参数的最佳组合其实是一个搜索问题。两个简单易行的搜索策略是网格搜索和随机搜索.

b) 不调整参数,算法性能会显著下降.比如最终算法 k 近邻,使用默认参数的性能:

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=1, n_neighbors=5, p=2,
           weights='uniform')
        Accuracy: 0.86315       Precision: 0.63795      Recall: 0.25550 F1: 0.36487
        Total predictions: 13000        True positives:  511    False positives:  290
s: 10710
```

调参后的性能:

```
KNeighborsClassifier(algorithm='auto', leaf_size=30, metric='minkowski',
           metric_params=None, n_jobs=4, n_neighbors=2, p=2,
           weights='distance')
        Accuracy: 0.84546       Precision: 0.49715      Recall: 0.39250 F1: 0.43867
```

c) 对于 k 近邻进行手动调参.

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”】

a) 在监督学习中,将数据分为训练集和测试集,以测试算法真正的性能.
b) 典型错误是在训练集上得分较高，测试集上得分较低，也就是我们常说的“过拟合”.
c) 我直接使用了测试文件的 test_classifier 函数,该函数使用 StratifiedShuffleSplit 进行交叉验证. 其中的大部分样本进行模型训练，生成模型，留小部分样本用刚建立的模型进行预测，并求这小部分样本的预测误差，记录它们的平方加和. 这个过程迭代 1000 次并且可以重复采样，然后比较每组的预测误差，选取误差最小的那一组作为训练模型。

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”】

a) 精确度:
precision:0.49 的含义是正确预测为嫌疑人的数量占所有预测数量的 0.49.
b) 召回率:
Recall: 0.39 的含义是正确预测为嫌疑人的数量占所有嫌疑人数量的 0.39.

优达学城
2016 年 9 月