项目

Finding Donors for CharityML

此部分属于 Machine Learning Engineer Nanodegree Program

项目审阅

注释

与大家分享你取得的成绩！

Requires Changes

还需满足 3 个要求 变化

探索数据

学生正确地计算了下列数值:

- 记录的数目
- 收入大于50000美金的人数
- 收入小于等于50000美金的人数
- 收入大于50000美金的人数所占百分比

非常好，你的计算完全正确。

准备数据

学生正确地对特征和目标实现了独热编码。

注意`income_raw`的独热编码：

```
income = (income_raw == '>50K')
```

记过计算得出的值是布尔类型的，根据注意的要求：# TODO：将'income_raw'编码成数字值，你需要做出小修改。

评估模型表现

学生正确的计算了简单预测的准确率和F1分数。

注意题目的要求：那么这个模型在**验证集**上的准确率，查准率，查全率和 F-score是多少？而不是使用全体数据(`n_records`)进行计算。

学生解释了选择这几个模型的原因，并说明了每一个模型的优缺点。

对选择的三个特征，给出了很不错的解释

学生成功的实现了一个监督学习算法的流程。

算法流程中有一个粗心的地方需要关注：

```
# TODO：计算在验证上的准确率
    results['acc_val'] = accuracy_score(y_val, predictions_val)
```

```
# TODO: 计算在最前面300个训练数据上的F-score
results['f_train'] = accuracy_score(y_val, predictions_val)
```

两行代码编写是一样的，注意观察注释的要求来修改你的代码。

✓

学生正确的实现了三个监督学习模型，得出了模型表现可视化的图表。

正确设置了`random_state`参数，并通过可视化，进行了模型之间性能的比较

优化结果

✓

在考虑了计算成本、模型表现和数据特点之后，学生选出了最好的模型并给出了充足的理由。

✓

学生能够用清晰简洁的话来向一个没有机器学习或任何其他技术背景的人来解释最优模型的工作原理。

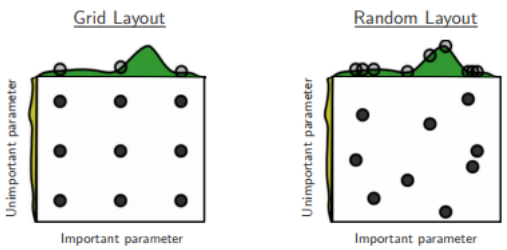
对`Adaboost`的算法流程掌握得很不错。

✓

最终模型利用了网格搜索进行参数调优，至少挑战了一个参数，并且至少有三个可选值。如果模型参数不需要任何调整，学生需要给出明确的理由。

很不错的调参，模型有了不错的提高。

`Adaboost`的调参过程很耗费时间，所以使用随机调参可以节省由遍历搜索所耗费的**大量时间**，同时，最终调参得到的模型性能还是有保证的。



`sklearn`对该算法封装得很好，用起来非常友好:

```
clf = ... # 模型
param_dist = ... # 参数列表
n_iter_search = ... # 你想要搜索的次数
random_search = RandomizedSearchCV(clf, param_distributions=param_dist, n_iter=n_iter_search)
random_search.fit(X, y)
```

你可以从这里阅读到`gridsearch`与`randomsearch`两者的比较。你也可以从这篇通俗易懂的论文了解到更多相关的知识。

✓

学生在表格中正确汇报了调优过后、调优之前以及基准模型的准确率和 F1 分数。学生把最终模型的结果与之前得到的结果进行了对比。

正确填写了表格中的计算结果。

特征重要性

✓

学生列出了他们认为对预测个人收入最重要的5个特征，同时给出了选择这些特征的理由。

对选择的5个特征，给出了很不错的解释。

✓

学生调用了一个监督学习模型的`feature_importances_`属性。此外，学生列出了这些重要的特征并讨论了这些特征的相同点和不同点。

`education_level`，`occupation`和`workclass`都是分类类型的特征，经过独热编码后，对应的值都被打散成新的特征。所以没有进入到前五重要的特征中。

通过合适的编码方式，我们仍然可以知道分类类型变量的重要性。

首先你需要比较one hot 还有labelEncoding这两种编码方式对最终结果的影响。

- 对于决策树或者本质算法是基于信息增益的，选择这两种编码方式对最终结果不会有太大的影响。
- 但是对于迭代优化算法来说(Logistic regression, svm, 神经网络)如何选择这两种编码方式需要从实验中得到结果。你可以参考这里的回答了解更多。

所以对于可以直接调用`feature_importances_`属性的模型，你可以先对数据进行`labelEncoding`然后训练得出重要性的评估。

✓

学生用最重要的5个特征建模并分析了和对比了改模型与问题五中的最优模型的表现。

注意一下：这里的education-num不仅代表教育时长，而且它是education_level的labelEncoding结果，某种程度来说也代表学习水平。

以下代码可以清楚解释这个原因，注意观察以下代码的输出值：

```
zip(list(data.education_level.values), list(data['education-num'].values))
```

重新提交

下载项目

了解 修改和重新提交项目的最佳做法.

返回 PATH

学员 FAQ