项目

Finding Donors for CharityML

此部分属于 Machine Learning Engineer Nanodegree Program

项目审阅

代码审阅

注释

与大家分享你取得的成绩！

Requires Changes

还需满足 1 个要求 变化

同学你好，你的首次提交非常优秀，只差一步就能完成本次项目，请认真修改`logistic regression`模型叙述部分，期待你的下次提交。

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。

使用了`apply`函数，非常高效的独热编码。

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。

✓

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。

✓

学生成功的实现了一个监督学习算法的流程。

算法流程的代码编写准确无误。

✓

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。

很棒，正确设置了`random_state`参数。

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。

机器学习模型一般都有这三个要点需要解释:

- 模型的基本流程．(你已经提到逻辑回归需要得到一个概率模型还有逻辑回归中特征的权重这两个概念)
- 优化目标．(你应该提及损失函数)
- 模型的训练是如何进行的 (可以提到数据的作用，机器学习一定要提到数据，最好能解释一个训练方法，如梯度下降).
  有关`logistic regression`更加具体的内容可参考这里

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。

或许你需要设置`cv`在`gridsearchcv`里面的数值。

- cv在帮助文档的解释里面的说明:
  Determines the cross-validation splitting strategy. Possible inputs for cv are:
  - None, to use the default 3-fold cross validation,
  - **integer, to specify the number of folds in a (Stratified)KFold,**
  - An object to be used as a cross-validation generator.
  - An iterable yielding train, test splits

默认参数是3，这样在进行交叉检验时，模型的结果表现得很不稳定，你需要设置更加大的`cv`，使得训练集的分布接近于整个数据的分布，比如说设置`cv = 10`

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。

一般在kaggle比赛上获奖的大多是树型模型和经过`stacking`后的融合模型。

- 你可以从这里了解到`stacking`的基本原理
- 你可以从这里把`stacking`实践应用到真实的数据上，得到很不错的分数
- 或许你需要了解这个强大的机器学习的库mlxtend，像sklearn一样直接调用stacking
- 尝试得到更大的提高吧

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。

很不错的特征选择。

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。

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))
```

这里显示前20个数据:

	education-num	education_level
0	13.0	Bachelors
1	13.0	Bachelors
2	9.0	HS-grad
3	7.0	11th
4	13.0	Bachelors
5	14.0	Masters
6	5.0	9th
7	9.0	HS-grad
8	14.0	Masters
9	13.0	Bachelors
10	10.0	Some-college
11	13.0	Bachelors
12	13.0	Bachelors
13	12.0	Assoc-acdm
14	4.0	7th-8th
15	9.0	HS-grad
16	9.0	HS-grad
17	7.0	11th
18	14.0	Masters
19	16.0	Doctorate

✓

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。

重新提交

下载项目

重新提交项目的最佳做法