UDACITY

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 4 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**

请注意最后一行的输出：

```
Percentage of individuals making more than $50,000: 0.25%
```

这里是以**百分数**的形式输出,所以你需要做出相应的数值转换。

准备数据

学生正确地对特征和目标实现了独热编码。

仔细查看题目的注释：# TODO：将'income_raw'编码成数字值

```
income = income_raw == ">50K"
```

稍微修改一下以上的代码

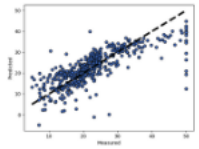
评估模型表现

学生正确的计算了简单预测的准确率和F1分数。

`fscore`的计算结果准确无误

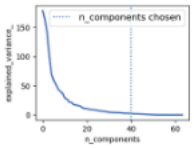
学生解释了选择这几个模型的原因，并说明了每一个模型的优缺点。

Learn from examples 是掌握一个算法模型的关键，这个sklearn gallery是一个非常好的资源：

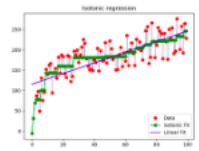


Plotting Cross-Validated Predictions

Concatenating multiple feature extraction methods



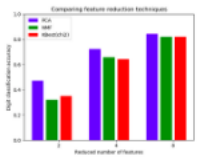
Pipelining: chaining a PCA and a logistic regression



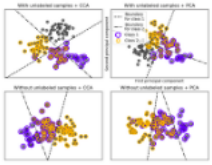
Isotonic Regression

Imputing missing values before building an estimator

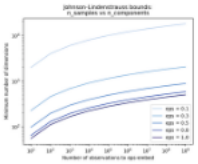
Face completion with a multi-output estimators



Selecting dimensionality reduction with Pipeline and GridSearchCV



Multilabel classification



The Johnson-Lindenstrauss bound for embedding with random projections

你可以通过`ctrl`+`f`快速导航到想要学习的模型。不嫌多，四五个例子你就能对模型有个大概的理解。

- 判断一个模型是否适合该问题，可以从数据规模，问题类型，模型复杂度等等来谈.这个sklearn的算法地图能够帮你快速导航到相关的算法模型。
- 同时，微软这两个机器学习算法模型备忘录也是份很不错的笔记：cheat sheet1,cheat sheet2

学生成功的实现了一个监督学习算法的流程。

注意这一行代码`predictions_train = learner.predict(X_train)`。**先取数据的前300个在进行预测和先对全部数据进行预测再取前300个**两者稍微有点不一样，对于预测时间长的算法，比如说`knn`，后面的做法会消耗大量的时间，所以你需要修改这一行以及下面相关的代码。

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。

注意，你还需要为能设置`random_state`的算法设置`random_state`参数，这样做reviewer可以重现你的结果为你后续调参提供一个可重现的基准模型

- 避免因随机种子的干扰造成结果的变化
- `random_state`的作用可以看看以下帖子:http://discussions.youdaxue.com/t/svr-random-state/30506

注意

修改了算法流程的错误后，别忘了重新运行这里的代码来进行正确的可视化

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。

整个调参过程一气呵成

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。

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。

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))
```

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。

得到了一个泛化能力很不错的模型

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

给这次审阅打分

学员 FAQ