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！

Meets Specifications

顺利完成project2～

在监督学习这个板块，你学到了很多很多知识，包括pandas，numpy，sklearn三大库的使用，常用的机器学习模型，以及各种评估指标等等，记住如此大量的知识非常头疼，包括reviewer在内，也会有健忘的时候>_< 这时备忘录是如此重要，你可以打印出来方便查看复习：

- pandas cheat sheet
- numpy cheat sheet
- scikit-learn cheat sheet

这个link有很多哦很棒的学习资源，你可以尽情地挖掘

一些名词很容易忘记，这个字典是个很好的汇总，帮助复习理解。

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**

准备数据

✓

学生正确地对特征和目标实现了独热编码。

很棒，完成了所有的独热编码操作。

评估模型表现

✓

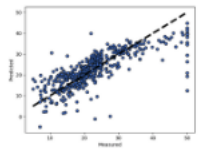
学生正确的计算了简单预测的准确率和F1分数。

非常好，正确计算了`fscore`的数值。

✓

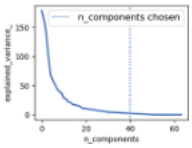
学生解释了选择这几个模型的原因，并说明了每一个模型的优缺点。

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：

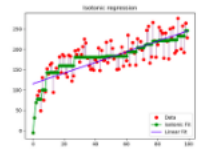


Plotting Cross-Validated Predictions

Concatenating multiple feature extraction methods



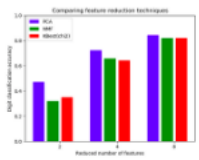
Pipelining: chaining a PCA and a logistic regression



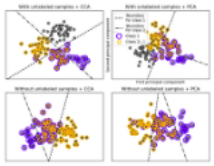
Isotonic Regression

Imputing missing values before building an estimator

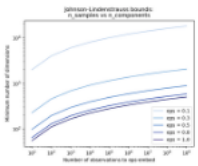
Face completion with a multi-output estimators



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。

你可以使用随机抽样(避免label没有完全打乱，出现有偏的采样)的方式来优化你的算法流程，以下展现其中的一个代码块

```
sampled_row = pd.Series(X_train.index).sample(n = sample_size)
temp_x = X_train.loc[sampled_row, :]
temp_y = y_train.loc[sampled_row, :]
start = time()
learner = learner.fit(temp_x, temp_y)
end = time()
```

✓

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。

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。

✓

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。

顺利完成了整个调参过程的代码编写。

✓

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

✓

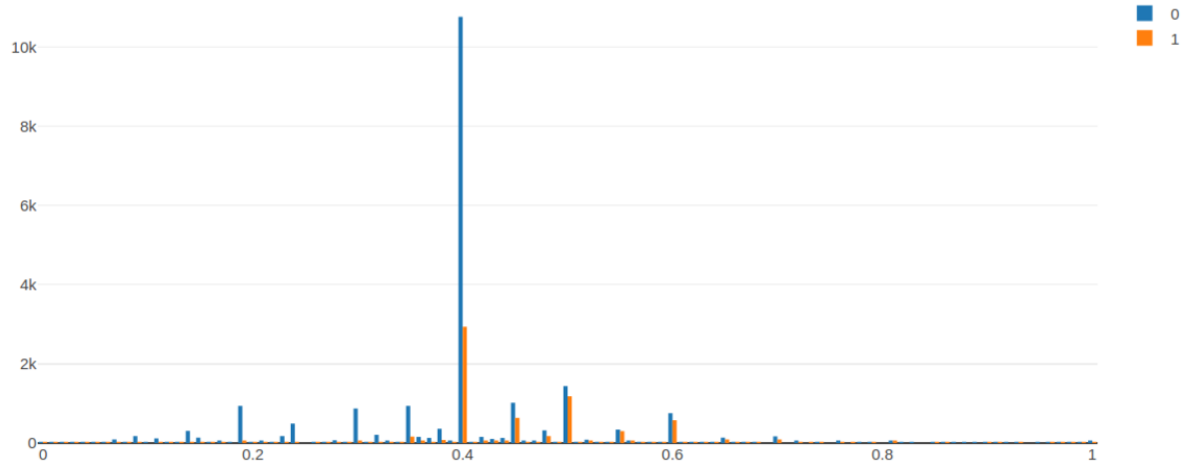
学生列出了他们认为对预测个人收入最重要的5个特征，同时给出了选择这些特征的理由。

✓

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。

尽管我们得到了一个很不错的模型，但是如果我们把模型完全当成一个黑箱，就会失去挖掘知识的机会，所以一定的数据探索是很有必要的，下面举个用`plotly`画图的例子(注:在导入`plotly`的时候，你需要在命令行执行`pip install plotly`下载相关的package):

```
import plotly.graph_objs as go
from plotly.offline import init_notebook_mode, iplot
init_notebook_mode(connected=True)
trace1 = go.Histogram(x = X_train[y_train==0]['education-num'].values, name='0') # 这里选择画出柱状图
trace2 = go.Histogram(x = X_train[y_train==1]['education-num'].values, name='1')
data = [trace1, trace2]
iplot(data)
```



我们发现`education-num`的数值在大于0.4时，收入`>50K`表现很突出，所以这个排在top5的特征具有很不错的重要性。

✓

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。

下载项目

返回 PATH

给这次审阅打分