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！

Meets Specifications

恭喜你，顺利完成了本次project，接下来你将会学到更多算法模型，了解到更多相关的数学细节，所以在这里推荐李航的<<统计学习方法>>，希望你能够保持学习的热度完成接下来的项目，加油:-)

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。

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。

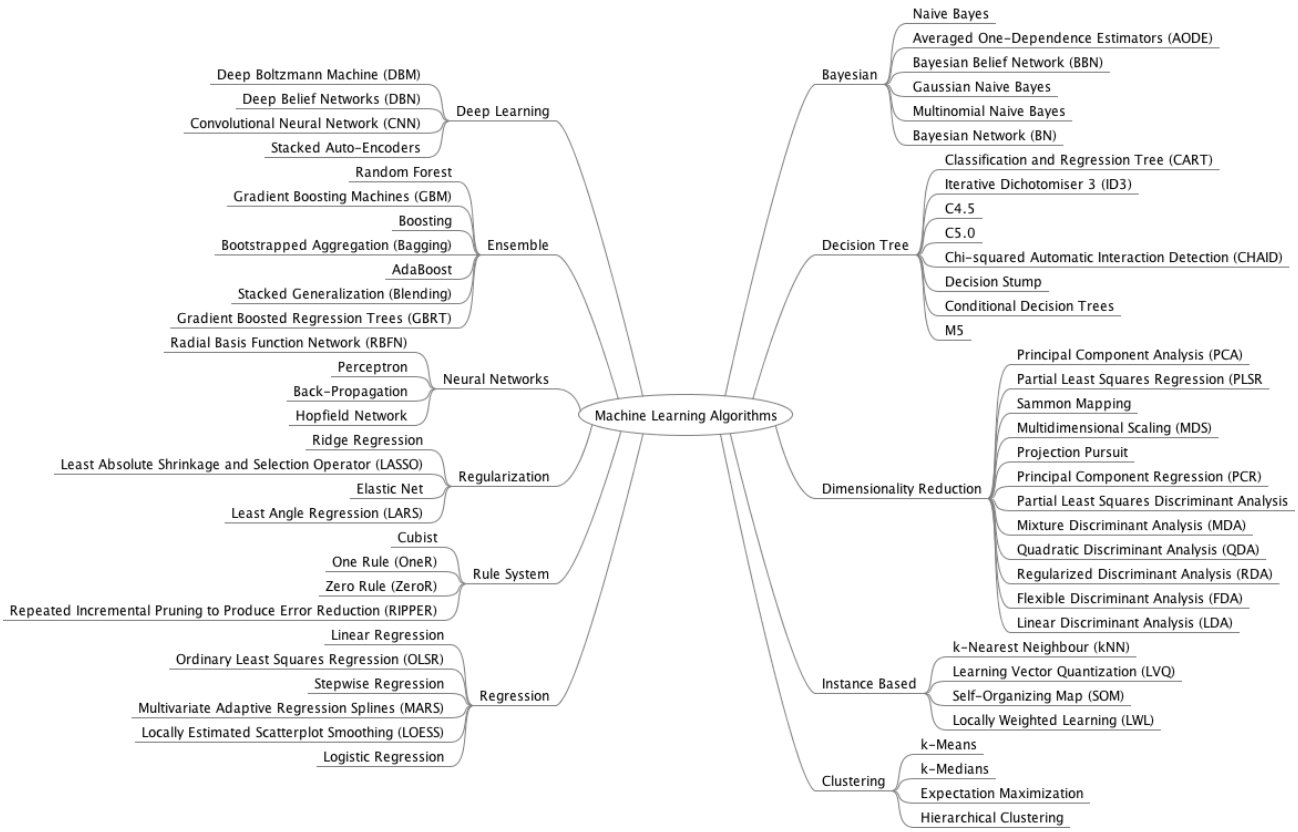
✓

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。

很不错的叙述

提高

- 选择一个好的算法尤其重要，首先你需要对每一个算法有个大概的分类，这里是一个很棒的思维导图：



这个blog对每一个算法都做了对应的说明，尽管是英文的，但是通俗易懂值得一读。

- 不仅要知道每一个模型的优缺点，而且要清楚每一个模型的用法，这个sklearn的算法地图能够帮你快速导航到相关的算法模型。

✓

学生成功的实现了一个监督学习算法的流程。

✓

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。

多个模型进行比较得出非常漂亮的图表。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

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`logistic regression`的训练和预测过程掌握得很不错。

✓

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。

通过调参得到一个很不错的模型。

✓

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。

✓

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。

你还可以从这几个方面入手来思考这个问题：

- 这个模型是不是不可以做online training。如果不能，在实际生产中，训练结束之后把模型保存下来，就可以不停地使用，训练时间就不是考虑的范畴，andrew Ng老师对这个概念做了很详细的描述，你可以从这个coursera视频了解更多。
- 我从特征选择和特征降维的角度入手，把要丢弃的特征利用起来，在保证训练和测试时间(视降维算法而言)的同时在一定程度上也能有较高的精度。
    - 关于特征选择通俗易懂的说明，介绍了什么是特征选择，特征选择可以解决什么问题，以及常用的特征选择的方法。
    - 接下来要介绍特征选择的进阶版，配合特征降维来使用，大大提高模型的性能
    - 配合sklearn来做特征选择，简单明了
    - 本数据很多特征都是one-hot的形式，对于one-hot特征的降维，我建议你读一下这篇blog，配合sklearn库来做特征降维，能得到很不错的结果。
- 并且随着硬件以及算法的发展，模型的训练速度会不断提升，准确率才是我们评价一个模型好坏的最终标准。

下载项目

返回 PATH

学员 FAQ