Predicting Boston Housing Prices

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

只有一点小问题了，加油:)

分析数据

✓

请求的所有 Boston Housing 数据集统计数据均已得到精确计算。学生可恰当利用 NumPy 功能获得这些结果。

✓

学生正确解释各项属性与目标变量增加或减少之间的关联。

✓

学生合理解释为何要为某个模型将数据集分解为训练子集和测试子集。训练和测试分解会在代码中正确实现。

当模型出现过拟合时，模型可能在训练集上表现很好，但却不能泛化到新的数据。测试集能够评估模型对未知数据的泛化能力。

模型衡量标准

✓

性能指标在代码中正确实现。（如果做了可选题，但不正确，算过不通过）

✓

学生正确判断假设模型是否能根据其 R^2 分数成功捕捉目标变量的方差。

分析模型的表现

✓

随着训练点的不断增加，学生正确判断图表中训练集和验证集曲线的走向并讨论该模型是否会得益于更多的训练点。

✓

学生提供最大深度为 1 和 10 的分析。如果模型偏差或方差较高，请针对每个图形给出合理的理由。

很好的判断，高偏差通常是由于模型太简单（即模型欠拟合），不能很好的拟合测试集，训练分数、验证分数、测试分数通常都比较低；高方差通常是由于模型过于复杂（即模型过拟合），模型在训练集上表现得很好，在验证集和测试集上得分确比较低，泛化能力差。

✓

学生根据合理的理由使用模型复杂度图形猜测最优模型的参数。

评估模型性能

✓

学生准确说明网格搜索算法，并简要探讨该算法的用途。

其实通常我们不会尝试每个参数，而是只尝试重要的参数，比如决策树中我们只调节了最大深度这个参数。

在进行网格搜索的时候，我们已经指定了算法，并给出一个参数列表和参数值，给出的值要尽可能的包含潜在的最优参数值，然后再搜索。

🔄

学生准确说明如何对模型进行交叉验证，以及它对网格搜索的作用。

“网格搜索时如果不使用交叉验证只能估计出模型对训练集合的评分……”

即使不使用交叉验证，得到的评分也是对验证集而不是训练集的，仍然是对没有参与训练的数据的评分。

"……但这样做只用到了训练数据的80%，并没有充分利用所有训练数据，会有些偏差。"

是的，K折交叉验证避免了因为数据集划分的偶然性造成的评分偏高或偏低的问题。 对于每一组参数对应的模型，通过使用不同的训练集和验证集训练然后取K次评分的平均来得到最终成绩来保证评分的客观和准确，从而准确定位到给出参数中的最优参数。

你可以花几分钟时间阅读一下这篇文章“网格搜索算法与K折交叉验证”来回顾一下这部分知识，还可以自己尝试运行文章后附的代码来帮助理解。

✓

学生在代码中正确实现 `fit_model` 函数。（如果做了可选题，但不正确，算过不通过）

很好的实现，你还可以设定KFold的默认参数。比如`n_splits=10, random_state=1, shuffle=True`。

✓

学生根据参数调整确定最佳模型，并将此模型的参数与他们猜测的最佳参数进行对比

非常好，最优模型的最大深度是4，与你之前的判断相同。

进行预测

✓

学生报告表格所列三位客户的预测出售价格，根据已知数据和先前计算出的描述性统计，讨论这些价格是否合理。

✓

学生计算了最优模型在测试集上的决定系数，并给出了合理的分析。

✓

学生可以合理分析最优模型是否具有健壮性。

✓

学生深入讨论支持或反对使用他们的模型预测房屋售价的理由。

（可选）预测北京房价

✓

学生用代码实现了数据分割与重排、训练模型、对测试集进行测试并返回分数。使用交叉验证对参数进行调优并选出最佳参数，比较两者的差别，最终得出最佳模型对测试集的预测分数。

✓

学生的回答与其实现的代码相吻合。并表达了自己的观点。

给这次审阅打分