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 5 个要求 变化

作为第一次提交，做的很棒！有几个答案需要进一步完善一下，期待下一次提交～

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。

做的不错，不少人在这里用了pandas来计算，并非我们要求的NumPy。并不是我们对NumPy有偏好，我们希望通过这里让你了解：虽然他们在大部分时候得出的结果相同，但是在某些情况下，例如这里的求标准差的计算上，是不一样的。具体区别可以查看他们的文档：

https://docs.scipy.org/doc/numpy/reference/generated/numpy.std.html
http://pandas.pydata.org/pandas-docs/stable/generated/pandas.DataFrame.std.html?highlight=std#pandas.DataFrame.std

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。

分析的不错。有时候在美国，如果学生/老师比例越高（学生多），证明该区域教育质量不高，因此很可能房价也不高。这也是中美的不同之处之一。这里我们想让学生知道，我们可以用先验知识（domain knowledge）做一些推断，机器学习算法可以帮我们验证我们的推断是否正确。

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。

总体回答的不错。不过"为避免过拟合，将数据集分割为训练集合测试集，使用测试集检查模型的泛化能力。" 不准确。

数据集划分不是为了避免过拟合，因为它不是产生过拟合的原因。同时单纯划分训练集测试集也无法避免过拟合，因为测试集不能参与模型优化的过程。见不能更简单通俗的机器学习基础名词解释。

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）

很棒！自己实现可以让我们打开黑盒，了解函数到底做了什么。

✓

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。

做的很好，R^2是评价模型表现的方法之一，每个机器学习模型的建立都要有相对应的评价指标，后面我们会学到更多的评价指标。不过R^2其实也有很多局限性需要注意

https://en.wikipedia.org/wiki/Coefficient_of_determination#Caveats

可汗学院对此也有很精彩的讲解。

skearn对于常见的模型表现衡量方法也有详细的介绍。
http://scikit-learn.org/stable/modules/model_evaluation.html

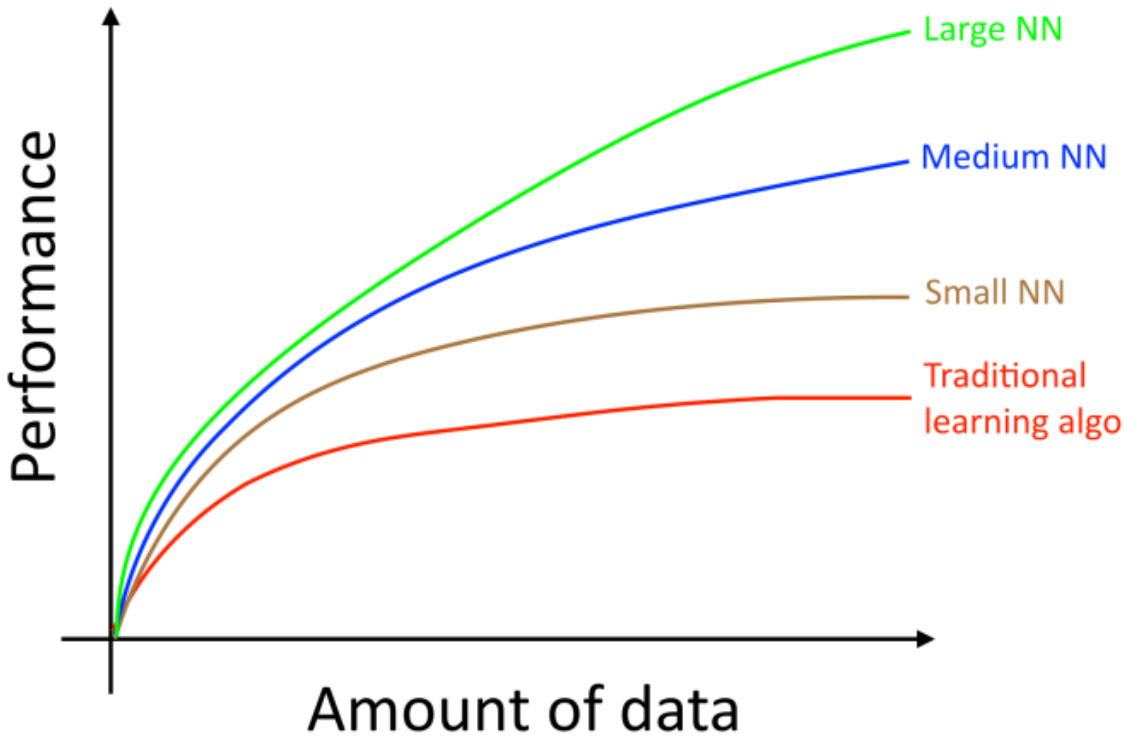
分析模型的表现

✓

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。

回答的不错。

注意：如果出现过拟合，单纯增加训练数据，也无法一致提升模型表现。



传统的机器学习算法（又被称为基于统计的机器学习）在数据量达到一定程度后，更多的数据无法提升模型的表现。深度学习的一个优势就是它可以把大量的数据利用起来，提升学习表现。

✓

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。

对偏差和方差的理解很不错！借用西瓜书上的比喻，用机器学习来判断一个物体是不是树叶，underfitting是以为所有绿色的都是树叶（没学会该学的）；overfitting是以为树叶都要有锯齿（学过头了，不该学的也学了进去）。这两者都不是我们想要的。

维基百科对此也有详细的解释 https://en.wikipedia.org/wiki/Bias%E2%80%93variance_tradeoff

华盛顿大学机器学习的课程详细讲了这个问题，你可以免费观看。 https://www.coursera.org/learn/ml-regression/home/week/3

✓

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。

回答正确，这里4是最佳选择。这与之前学习曲线图也是一致的。

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。

没错，GridSearch就是把给定参数下所有可能的组合都试一遍，通过指定的评价函数找出最优。

同时还要注意，这里的最优也是我们给定参数下，给定 Kfold（如果使用）的K下的最优。参数空间变化和K取值的变化都会引起结果不同，所以即使是GridSearch，也无法保证是绝对最优。

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。

回答的不错。你还要补充：

- “将原始数据分成K组”不准确。K折交叉验证分割的是训练数据还是全部数据？（提示：`reg = fit_model(X_train, y_train)`）
- 默认情况下 Kfold 是对数据按顺序切分还是随机切分？
- 我们可以把训练数据划分为8:2的训练集和验证集，然后每个参数组合在训练集上训练，验证集上打分。选出表现最好的一组参数。这样的交叉验证没有使用网格搜索。像对数据进行单次分割来进行网格搜索，可能会有什么问题？交叉验证又是如何避免这个问题的？

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）

完美地实现了GridSearchCV。自己的实现也很精彩！

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

此答案与你在问题 6所做的猜测是否相同？

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。

从房屋特征的数值判断，这样的价格合理吗？为什么？
提示：用你在分析数据部分计算出来的统计信息来帮助你证明你的答案。

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。

对于这个结果，你的结论是什么？

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。

做了很好的尝试。这里希望学生熟悉一下学过的代码，决策树本身并不适合这个问题。后面我们会学到更多算法和技巧来解决它。

给这次审阅打分