UDACITY

项目

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 4 个要求 变化

作为第一次提交，做的很棒。有些细节需要再完善一下，期待你下次提交～

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，但这里你用了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。

模型衡量标准

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

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。

测试集只有一个（一次）用途，就是构建完**最终模型后**，看一下表现如何。它只能发现模型是否过拟合，但是因为测试集不能参与选择模型或者调参的过程，因此也就也无法避免过拟合。用测试集的结果来调整模型会造成data leakage。我们通常用**验证集**来调优。详见不能更简单通俗的机器学习基础名词解释。

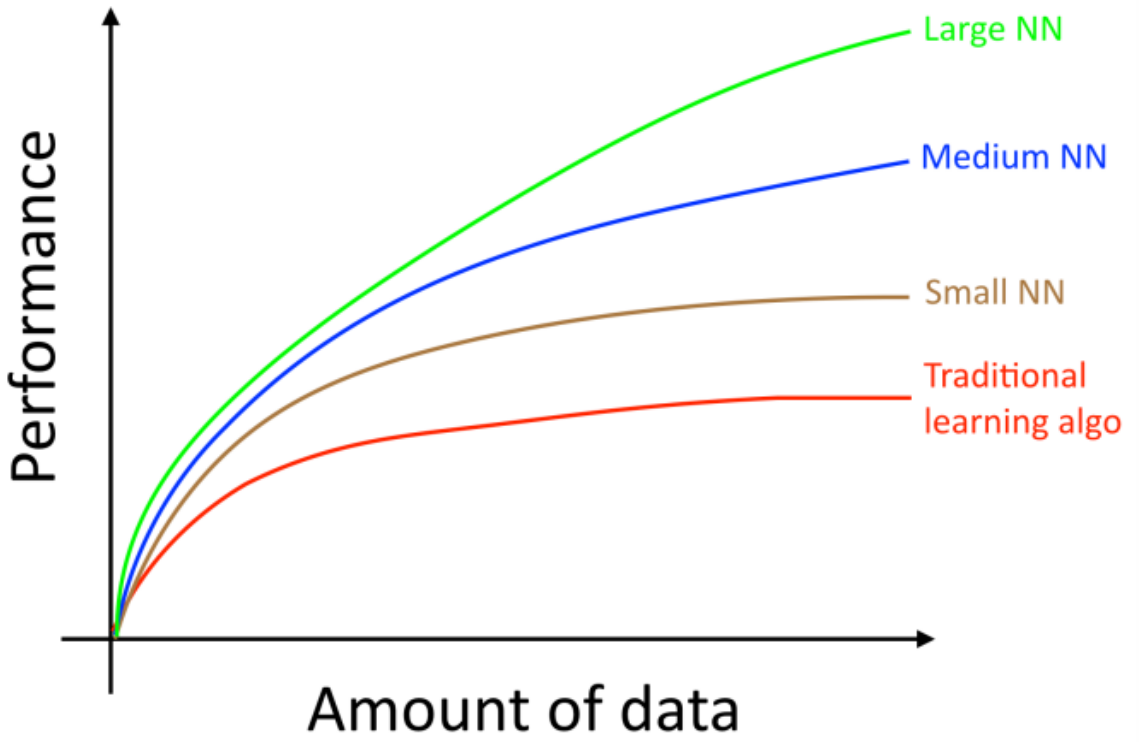
https://en.wikipedia.org/wiki/Test_set#Validation_set 也提到了这三者之间的关系，可以看一下。

"用训练集做测试，会出现过度拟合的问题" 不正确。前者不是造成后者发生的原因。过拟合是模型复杂度高造成的，与拿何种数据来测试无关。没有未知数据，我们无法知道模型是否过拟合。

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。

对训练曲线和测试曲线趋势和意义解释的很好。这里随着数据的增多，max_depth不变的情况下，模型提升的幅度也越来越小。



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。

这里还有更多关于学习曲线的介绍：
https://www.coursera.org/learn/machine-learning/lecture/Kont7/learning-curves
http://scikit-learn.org/stable/auto_examples/model_selection/plot_learning_curve.html

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

sklearn 也有对 validation curve的介绍：http://scikit-learn.org/stable/modules/learning_curve.html

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就是把给定超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超参数范围，给定 Kfold（如果使用）的K下的最优。超参数空间变化和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。

回答的很好，还需要明确一下：

- 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)`）
- “每一份数据成为测试集”不正确，这里用来打分的数据集是验证集，测试集不能参与模型调优。

学生在代码中正确实现 `fit_model` 函数。

`scoring_fnc = make_scorer(r2_score)` 不正确。

TODO: 需要使用你之前定义的 'performance_metric'

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。

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。

对价格的合理性分析地不错！这里我们可以根据房子的特征来做横向比较；也可以与数据集中特征相近的房子做纵向比较。最终还可以跟统计数据中的最大值，最小值，均值等做比较。

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。

关于敏感性分析更多的知识可以参考 https://en.wikipedia.org/wiki/Sensitivity_analysis

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。

重新提交

下载项目

重新提交项目的最佳做法

Ben 与你分享修改和重新提交的 5 个有益的小贴士。

观看视频 (3:01)

返回 PATH

学员 FAQ