**项目**

Predicting Boston Housing Prices

此部分属于 Machine Learning Engineer Nanodegree Program

* [**项目审阅**](https://review.udacity.com/)
* [**代码审阅**](https://review.udacity.com/)
* [**注释**](https://review.udacity.com/)

**SHARE YOUR ACCOMPLISHMENT**

Requires Changes

**还需满足 3 个要求 变化**

可选部分，手动实现网格搜索没有成功，希望多给出一些建议

可以自己再多尝试一下，你的思路还是有点乱，先解决简单的问题。再把各个功能组合起来。

def get\_r2\_score(train\_one, train\_two, train\_three, test\_one, test\_two, test\_three, regressor, scoring = performance\_metric):

""" 返回每组数据 """

print train\_one

feature = train\_one + train\_two

print feature

步骤中多加几个print 也有助于你寻找问题在哪。

**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。**

分析的有理有据！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。**

回答的很好。机器学习的核心目的就是构建具有泛化能力的模型，如果没有测试集，或者用模型已经见过的数据来测试，都无法确切知道模型对预测未知数据的表现如何，也就无法衡量模型表现。

**模型衡量标准**

**性能指标在代码中正确实现。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](https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/assessing-fit-least-squares-regression/v/r-squared-or-coefficient-of-determination" \t "_blank)。

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。不过还要补充下随着训练数据数量的增加训练集评分曲线的跟验证集评分曲线是怎样变化的？谁上升，谁下降。变化速率又如何？

**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](https://book.douban.com/subject/26708119/" \t "_blank)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。**

**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：

* 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？

**学生在代码中正确实现 fit\_model 函数。**

完美地实现了GridSearchCV。

**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。**

分析的不错！除了房屋本身的特征，你还需要用你在分析数据部分计算出来的统计信息（最大值，最小值，平均值，中位数等）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。**

**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。**

 重新提交项目

[**下载项目**](https://udacity-reviews-uploads.s3.amazonaws.com/_submissions/zipfile/641238/boston_housing.zip)









