UDACITY

登出

PROJECT

Predicting Boston Housing Prices

A part of the Machine Learning Engineer Nanodegree Program

PROJECT REVIEW | NOTES

Meets Specifications

SHARE YOUR ACCOMPLISHMENT

Hello Student,
恭喜你！！！ 你做到了。
我可以看到你仔细审查了以前的建议和意见，并遵循必要的指引，调整你的答案。不错！
我亦十分欣賞您給了很多的耐性去閱讀建議並加以修正答案。學習態度十分出色！
希望讲座和项目已经教会了你机器学习的基础知识。
继续你的良好的工作！ 加油

数据研究

✓ 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。

做得好! 简洁地了解资料组(Housing dataset) 以及计算出主要的统计数据。而且，你能成功地运用NumPy去计算所需的数据。

Pro Tips:

- 习惯去了解你的 dataset 的数据会对于你运行和应用预测模型(predictive model)时事半功倍，这是因为:
  - 我们能从中去得知这 dataset 是否符合我们模型(model)的基本假设，从而选出最合适的模型(model)
  - 在预测后，这些数据可以帮我们去进行事后的研究和比较，去了解我们我预测是否合理和没有离题

✓ 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。

真是十分的优秀，你可以用简单直接的答案如此明确地解释的价格与特征的相关性。做得好！

开发模型

✓ 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。性能指标在代码中正确实施。

很好! 你能好好运用 `sklearm R^2` function了

Suggestions and Comments:

- 這裡有一網頁，老幫助我了解更多不同的performance metric

✓

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施。

你给了一个不错的原因去解释为何我们要把dataset分割，而且你也成功地应用sklearn的 `train_test_split`！

Suggestions and Comments:

- 你提出了一个恰当的解释去说明我们要把dataset分割为training和testing subsets，这里有一些有用的数据可供你参考:
  - 请看Wikipedia page了解更多，这网站提供了一些背景资料去了解test sets
  - 在ML中Test sets的是被创造作测试被训练好的模型(trained model)的归纳能力(ability to generalize)，用一些这训练好的模型未遇过的data去测试其能力。
  - 这Wikipedia page也提到test sets是作了解这model的预测力强度和实用性的工具。
- 温馨一提，即使把dataset分割，模型(model)也有可能overfitting
- Amazon提供了充足的数据去了解我们要把dataset分割为training和testing subsets的原因。

分析模型性能

✓ 随着训练点的不断增加，学生正确判断图表中训练及测试曲线的走向并讨论该模型是否会得益于更多的训练点。

非常正确的描述。正确指出training set size 增加时 training and testing errors的改变

✓ 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。

极好! 能指出 max_dept = 1 时 模型(model) 受到 high bias 影响 以及说出 max_dept = 10 时模型(model) 会 overfitting ， 亦即是受到high variance影响

Suggestions and Comments

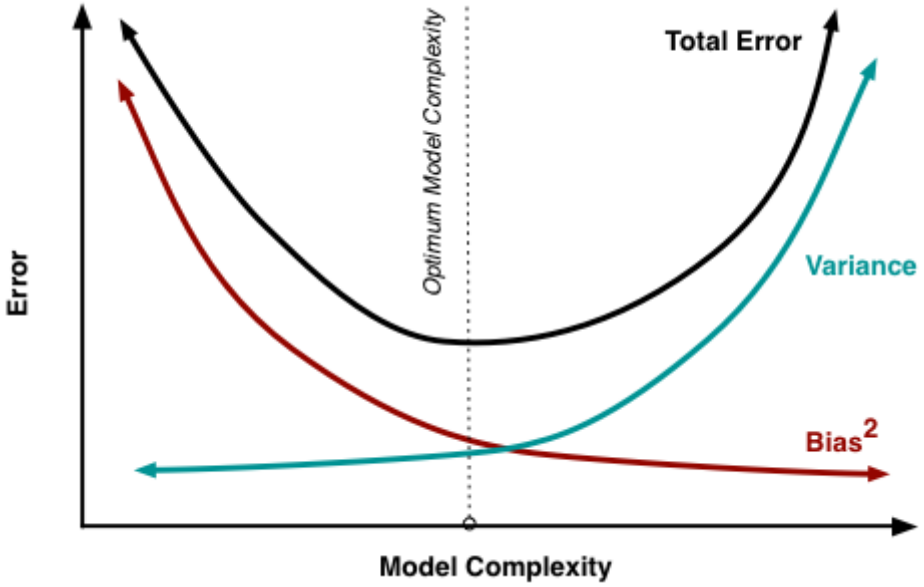
- 请查看这 去帮助你了解更多有关model bias-variance tradeoff

✓ 学生根据合理的理由使用模型复杂度图形选择最佳猜测最佳模型。

好!一个不错的预测。基于一个对于overfitting 和 uderfitting的合理和恰当的解释去处理这问题。让我们看看你这预想与model计算的optimal max depth是否相同。

Pro Tip:

请看这图去了更多有关 bias-variance tradeoff :



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。**GridSearchCV** 会在代码中正确实施。

非常好。能解释和应用 grid search algorithm。

Pro Tips:

另一个很有效的模型参数优化工具(parameter tuning algorithm)是RandomizedSearchCV. 与GridSearchCV对比，它是一个从指定样本中抽样的方法，以特定数量去设定参数(parameters)，而非设定所有参数(parameters)。

- 使用RandomizedSearchCV的优点是它比GridSearchCV快，以及在理论上证明它甚至比grid search更好。

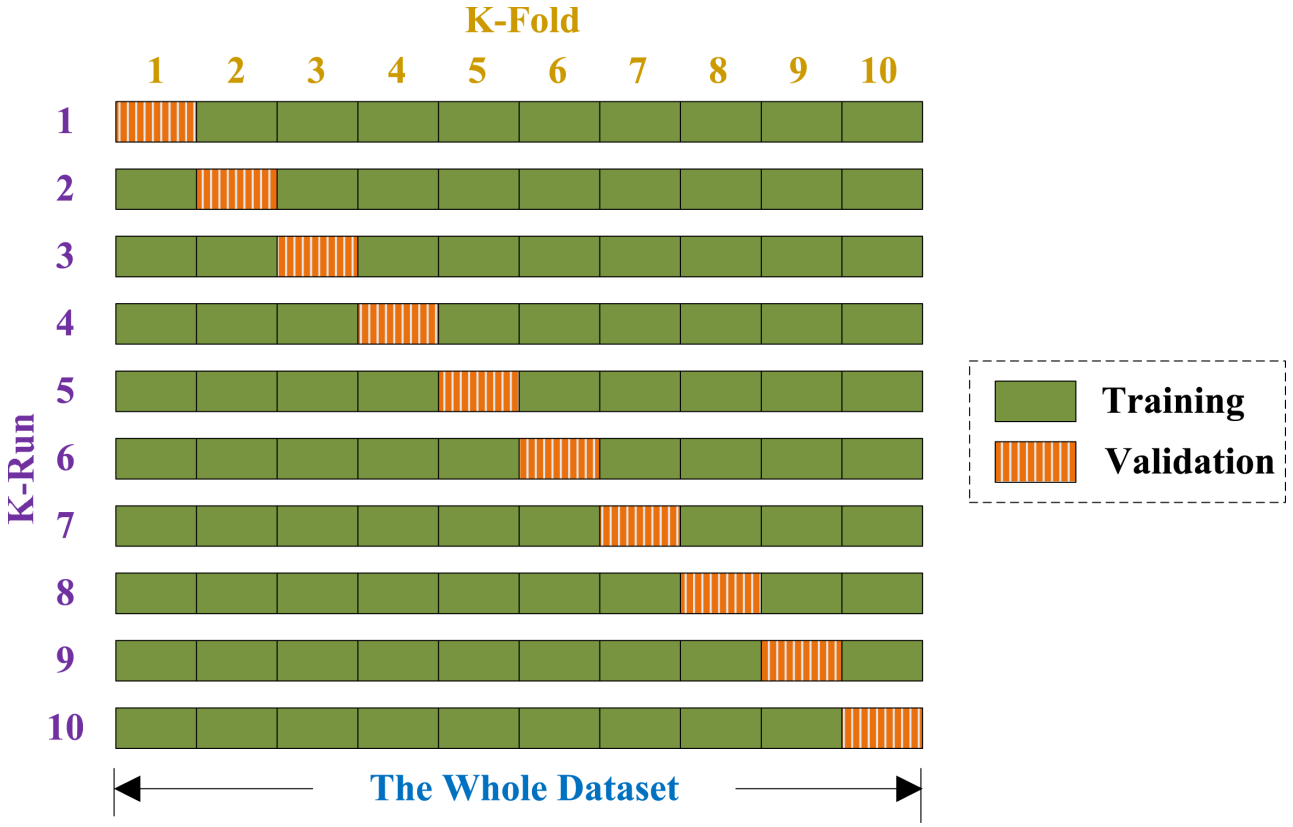
✓ 学生准确说明如何对模型进行交叉验证，以及它对网格搜索的作用。除非有合理的理由，否则不得对默认的 **3** 折交叉验证以外的 **GridSearchCV** 部分进行修改。

很正确的描述! 能正确指出 cross validation如何在模型中运行(model)

Suggestions and Comments:

- 请查看scikit-learn page去透彻了解 cross-validation 的定义和运作。同时请浏览 cross-validation on Wikipedia) 去得出更多相关信息

- 简单地说一下我对这部分的见解，请看下图



- 以下是我由这图中对于cross-validation运作描述:
  从图可见，我们把整个 数据(dataset)分成了 k-folds，或是k个分档(subsets)，这图所示的是10十个分檔(subsets)
  然后这model会被训练和验证K次，这例子中便是10次
  在每次运算时，便会有一个分档(subsets)当作验证而其余K-1使用作训练
  最终，验证的得分会被记录和平均化，以得出最终最佳得训导模型(model)
- 当中的优点是，它能完美地极限化使用这数据(data)，即便数据是十分有限也能帮助避免overfitting
- 如sklearn page for cross-validation所示，当只有一个subset被用作参数优化(parameter tuning)，只有这subset才知道当中的信息，而其他不会，从而避免了overfitting。
- 希望这会对你学习ML和什么是cross-validation有所帮助
- 因为Cross validation把training 和testing data的可用性极大化，所以它对于我们要符出最佳学习结论时是十分有用，这对于我们在数据(data)十分有限时是极其重要和具帮助性。

如果我们限制grid search着于一个 testing set，当数据(data)是不平衡(imbalanced)时，这可能偶然地出现overfitting。利用cross validation能帮我们把参数(parameters)优化以及消除一些随机性的异常

✓ 学生在代码中正确实施 `fit_model` 函数。

好!非常正确的运用。

✓ 学生根据参数调整确定最佳模型，并将此模型与他们选择的模型进行对比。

不错! 你前早所选的与模型(model)中所需的max_depth很相似，如果能解释当中的差别会更好。

Pro Tips:

为了得出更强大的 `max_depth` parameter,你可用运行grid search algorithm多次
以下是我建议用的Code

```
max_depths = []
for i in range(1000):
reg = fit_model(housing_features, housing_prices)
max_depths.append(reg.get_params()['max_depth'])
best_max_depth = np.mean(max_depths)
print 'The Best model, on average, has a max depth of:', best_max_depth
```

基本上，如果你能正确理解和解读complexity curves的图，你的答案应于以上的code所得的十分相近。

Reply your question:

用肉眼看 `max_depth 3` 好像是最佳，但其實經過代碼程式驗證後得出的結果才是最佳，因為我們不能用肉眼比較訓練和測試分數的相對水平是怎樣才能得出最佳的預測如果。因此，作為一個機械學習工程師，我們需要學會客觀地相信數據反映出的結果。

✓ 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。

很好! 基于client 的情况，一个合理的价格预算。而且，能把你的答应与 median 和 mean 比较是一个不错的尝试。

✓ 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。

在这里，你提出了一个非常彻底的和逻辑的讨论！ 做得好！

Suggestions and Comments:

可以从以下这几个问题作出反思:

- 把grid search用到整个dataset会否影响到你对这模型(model)的信心?
- 若把这模型(model)用到Boston 以外的地区，你觉得还能发挥它的功用吗?
- 你觉得一个不同的运算方法(或是不同的判定树(Decision Tree))会使其更健全更强大吗?

DOWNLOAD PROJECT

Have a question about your review? Email us at review-support@udacity.com and include the link to this review.

RETURN TO PATH

Student FAQ