#Machine Learning Engineer Nanodegree

##Model Evaluation & Validation

##Project 1: Predicting Boston Housing Prices

Welcome to the first project of the Machine Learning Engineer Nanodegree! In this notebook, some template code has already been written. You will need to implement additional functionality to successfully answer all of the questions for this project. Unless it is requested, do not modify any of the code that has already been included. In this template code, there are four sections which you must complete to successfully produce a prediction with your model. Each section where you will write code is preceded by a **STEP X** header with comments describing what must be done. Please read the instructions carefully!

In addition to implementing code, there will be questions that you must answer that relate to the project and your implementation. Each section where you will answer a question is preceded by a **QUESTION X** header. Be sure that you have carefully read each question and provide thorough answers in the text boxes that begin with "**Answer:**". Your project submission will be evaluated based on your answers to each of the questions.

A description of the dataset can be found [here](https://archive.ics.uci.edu/ml/datasets/Housing), which is provided by the **UCI Machine Learning Repository**.

#Getting Started To familiarize yourself with an iPython Notebook, **try double clicking on this cell**. You will notice that the text changes so that all the formatting is removed. This allows you to make edits to the block of text you see here. This block of text (and mostly anything that's not code) is written using [Markdown](http://daringfireball.net/projects/markdown/syntax), which is a way to format text using headers, links, italics, and many other options! Whether you're editing a Markdown text block or a code block (like the one below), you can use the keyboard shortcut **Shift + Enter** or **Shift + Return** to execute the code or text block. In this case, it will show the formatted text.

Let's start by setting up some code we will need to get the rest of the project up and running. Use the keyboard shortcut mentioned above on the following code block to execute it. Alternatively, depending on your iPython Notebook program, you can press the **Play** button in the hotbar. You'll know the code block executes successfully if the message "Boston Housing dataset loaded successfully!" is printed.

#Statistical Analysis and Data Exploration In this first section of the project, you will quickly investigate a few basic statistics about the dataset you are working with. In addition, you'll look at the client's feature set in CLIENT\_FEATURES and see how this particular sample relates to the features of the dataset. Familiarizing yourself with the data through an explorative process is a fundamental practice to help you better understand your results.

##Step 1 In the code block below, use the imported numpy library to calculate the requested statistics. You will need to replace each None you find with the appropriate numpy coding for the proper statistic to be printed. Be sure to execute the code block each time to test if your implementation is working successfully. The print statements will show the statistics you calculate!

##Question 1 As a reminder, you can view a description of the Boston Housing dataset [here](https://archive.ics.uci.edu/ml/datasets/Housing), where you can find the different features under **Attribute Information**. The MEDV attribute relates to the values stored in our housing\_prices variable, so we do not consider that a feature of the data.

Of the features available for each data point, choose three that you feel are significant and give a brief description for each of what they measure.

Answer: I feel like the variables CRIM, DIS and RAD are the most significant in this dataset. They measure the per capita crime rate by town, the weighted distances to five Boston employment centres and the index of accessibility to radial highways.

##Question 2 Using your client's feature set *CLIENT\_FEATURES*, which values correspond with the features you've chosen above?

Answer: In this client's feature set, 11.95 corresponds to the per capita crime rate, 1.385 is the weighted distances to five Boston employment centres and 24 is the index of accessibility to radial highways.

##Question 4 Why do we split the data into training and testing subsets for our model?

Answer: We do split the data into training and testing because evaluating the performance of a classifier or a regression on the same data used to train an algorithm usually leads to an optimistically biased assessment. The simplest strategy for correcting the optimism bias is to hold out a portion of the development data for assessment.

##Question 4 *Which performance metric below did you find was most appropriate for predicting housing prices and analyzing the total error. Why?*

* *Accuracy*
* *Precision*
* *Recall*
* *F1 Score*
* *Mean Squared Error (MSE)*
* *Mean Absolute Error (MAE)*

Answer I find that the Mean Squared Error is the most appropriate performance metric for predicting housing prices because the response is quantitative. Accuracy, Precision, recall and F1 score are metrics used when predicting a qualitative response variable. MSE has nice mathematical properties which makes it easier to compute the gradient. MAE requires more complicated tools such as linear programming to compute the gradient. Because of the square, large errors have relatively greater influence on MSE than do the smaller error. Therefore, MAE is more robust to outliers since it does not make use of square. On the other hand, MSE is more useful if we are concerned about large errors whose consequences are much bigger than equivalent smaller ones.

##Question 5 What is the grid search algorithm and when is it applicable?

Answer: The Grid search algorithm is a way of systematically working through multiple combinations of parameter tunes, cross-validating as it goes to determine which tune gives the best performance. The Grid search algorithm picks out a grid of hyperparameter values, evaluates every one of them, and returns the winner. Grid search is most appropriate when function evaluations are cheap or the number of parameters is small.

##Question 6 What is cross-validation, and how is it performed on a model? Why would cross-validation be helpful when using grid search?

Answer Cross-Validation is a model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. It is performed by removing some data if data are not scarce, before training begins. Then when training is done, the data that was removed can be used to test the performance of the learned model on new data. When combined with grid search, cross-validation is used to evaluate every possible combination of a set of parameter values that we want to try on a given model.

##Question 7 Choose one of the learning curve graphs that are created above. What is the max depth for the chosen model? As the size of the training set increases, what happens to the training error? What happens to the testing error?

Answer: The max depth for the chosen model is 10. As the size of the training set increases, the training error rate increases slightly while the testing error decreases suddenly from around 130 to 30 and remains at this level whatever dimensions are added to the model. The training error increases slightly because as we get more training observations, it becomes harder for the model to fit the training data, so the training error goes up. On the other hand, if we have more data the model becomes better at generalizing, so the testing error goes down.

##Question 8 Look at the learning curve graphs for the model with a max depth of 1 and a max depth of 10. When the model is using the full training set, does it suffer from high bias or high variance when the max depth is 1? What about when the max depth is 10?

Answer: The training and testing errors in the model with max depth 1 converge and are quite high when the model is using the full training set. This model has systematic high errors and that means it is high biased. When the model is using the full training set and the max depth is 10, there is a large variance gap between the training and the test errors. It has a high variance.

##Question 9 From the model complexity graph above, describe the training and testing errors as the max depth increases. Based on your interpretation of the graph, which max depth results in a model that best generalizes the dataset? Why?

Answer: The complexity graph above indicates that the bias is reduced to almost zero for maximum depth greater than 12, but the variance term and, consequently, the generalization performance is not reduced further, staying approximately constant after a maximum depth of 6 or 7 is reached. A maximum depth between 5 and 7 generalises the dataset the best. It seems the model reaches its limit at that maximum depth because the training error and the testing error seem to be very close.

##Question 10 Using grid search on the entire dataset, what is the optimal *max\_depth* parameter for your model? How does this result compare to your intial intuition?

Answer: The optimal max depth for the model is 7. This exactly matches my initial intuition...

##Question 11 With your parameter-tuned model, what is the best selling price for your client's home? How does this selling price compare to the basic statistics you calculated on the dataset?

Answer: The best selling price for the client's home is 19.997, below the mean house price and far above the minimum house price.

##Question 12 (Final Question): In a few sentences, discuss whether you would use this model or not to predict the selling price of future clients' homes in the Greater Boston area.

Answer: I find that the selling price is below the average... For a real estate Agent it's definitely not a good deal! Technically, I find that we do not realize data cleaning, variable screening or even variables clustering. My feeling is that we skip some steps necessary for the model building procedure!

## Does Not Meet Specifications

## Quality of Code

#### Specification

Student’s code runs successfully and produces results similar to those in the PDF. No modifications are made to the template code beyond what is requested without justification.

Meets Specification

## Statistical Analysis & Data Exploration

#### Specification

All requested statistics for the Boston Housing dataset are accurately calculated. Student correctly leverages NumPy functionality to obtain these results.

Does Not Meet Specification

**Reviewer Comments**

The statistics that you calculated for the Boston housing data set are accurate. Except for the "Total number of features" that include 2 values. Please include the statistics also in the pdf version of the report.

#### Specification

Student adequately describes three separate features of the dataset. The corresponding values in the client’s feature set are correctly identified for the chosen features.

Does Not Meet Specification

**Reviewer Comments**

For the client house the DIS is not equal to 24 and RAD is not equal to 680. Please double check the data.

## Evaluating Model Performance

#### Specification

An appropriate performance metric is chosen with thorough justification. The metric is correctly implemented in code.

Does Not Meet Specification

**Reviewer Comments**

The performance metric that you choose is appropriate. However the justification is not accurate or specific enough. Please note that house price in most cases will take only a positive values. Do you mean that price is a continuous value? Please explain why not using Accuracy, F1 Score, Precision or recall in this case?

Considering the other metrics, MAE might also be appropriate for continuous features. Please explain the differance between MSE and MAE. Why you choose the MSE? For example the MSE weights are proportional to the squared size of the error where MAE weights are proportional to the size of the error. Can you see any benefit for the "Predicting Boston Housing Prices" in this feature?

#### Specification

Student provides a valid reason for why a dataset is split into training and testing subsets for a model. Training and testing split is correctly implemented in code.

Meets Specification

**Reviewer Comments**

You are correct that the estimation of a model using the same data use to train the model might lead to over optimistic or overfitting the model to the data set. This is especially important when we want to generalize the model to unseen data.

#### Specification

Student correctly describes the grid search algorithm and briefly discusses its application. GridSearchCV is properly implemented in code.

Meets Specification

**Reviewer Comments**

You are correct, Grid search perform hyper-parameter optimization or model selection simply by exhaustive searching through a manually specified subset of the hyperparameter space.

#### Specification

Student correctly describes how cross-validation is performed on a model, and why it is helpful when using grid search.  
Modifications beyond the default 3-fold cross-validation for GridSearchCV are reasonably justified.

Does Not Meet Specification

**Reviewer Comments**

The provided answer is not specifically referring to cross validation but seems the same one that you would have provided for justifying a train test split. This question requires you to describe what cross validation is and why it is useful when combined with grid search. You might find these links useful,  
[https://en.wikipedia.org/wiki/Cross-validation\_(statistics](https://en.wikipedia.org/wiki/Cross-validation_%28statistics))  
<http://www.anc.ed.ac.uk/rbf/intro/node16.html>

## Analyzing Model Performance

#### Specification

Student correctly identifies significant qualities of the training and testing errors as the training set size increases.

Meets Specification

**Reviewer Comments**

It might not be obvious from the figure but the training error is not constant but increase slightly, can you explain why?

#### Specification

Student provides analysis for both a max depth of 1 and a max depth of 10. Reasonable justification is given for each graph if the model suffers from high bias or high variance.

Meets Specification

**Reviewer Comments**

You are correct, the model high bias (underfit) for 'max depth' 1 and high variance (overfit) for 'max depth' 10.

For the 'max depth' 10 it is also important to indicate that the training error is close to 0, what that indicate about the model?

#### Specification

Student identifies how the training and error curves relate to increasing the model’s complexity.

Does Not Meet Specification

**Reviewer Comments**

For this question you are expected to describe the training error and separately the testing error as the model complexity increase. Please note that the answer describe the gap between the training and the testing error, but this is not what being asked.

#### Specification

Student picks a best-guess optimal model with reasonable justification using the model complexity graph.

Does Not Meet Specification

**Reviewer Comments**

You are correct that the model is optimal for max depth ~ 4. Please justify your insight from the chart. It is important to note here that it is important to measure the performance of the model using unseen data (testing data set). What in the testing error might indicate that the model is optimal?

## Model Prediction

#### Specification

Student determines the optimal model from parameter tuning and compares this model to the one they chose.

Meets Specification

**Reviewer Comments**

Please note that you can also run the model few times (~10) and choose the median value.

#### Specification

Student’s model produces a valid result. The predicted selling price is adequately justified by the calculated descriptive statistics.

Meets Specification

#### Specification

Student thoroughly discusses justification for or against using their model for predicting future selling prices.

Does Not Meet Specification

**Reviewer Comments**

The fact that the predicted price is below the average is not a good reason to conclude that the model is not appropriate. It might be the case that the features of the house indicated that the house is cheaper than the average price in the data set.

Do you think a different form of algorithm (other than a decision tree) would be better?