Project Report Questions

A description of the Boston Housing dataset can be found **here (https://archive.ics.uci.edu/ml/datasets/Housing)**. Use this slide as reference to the project questions you will encounter in the notebook. These questions (and your answers) must be present in your submitted report.

Data Exploration

This question is integrated into the project notebook output.

Using the NumPy library, calculate a few meaningful statistics about the dataset:

- How many data points (houses) were collected?
- How many features are present for each house?
- What is the minimum housing price? The maximum?
- What is the mean housing price? The median?
- What is the standard deviation of all housing prices?

1) Using your intuition, for each of the three features present in the dataset, do you think that an increase in the value of that feature would lead to an increase in the value of `'MDEV'` or a decrease in the value of `'MDEV'`? Justify your answer for each.

Developing a Model

2) For the hypothetical model presented in the project, would you consider this model to have successfully captured the variation of the target variable? Why or why not?

3) What is the benefit to splitting a dataset into some ratio of training and testing subsets for a learning algorithm?

Analyzing Model Performance

4) Choose one of the learning curves graphs presented in the project and state the maximum depth for the model. What happens to the score of the training curve as more training points are added? What about the testing curve? Would having more training points benefit the model?

5) When the model is trained with a maximum depth of 1, does the model suffer from high bias or from high variance? How about when the model is trained with a maximum depth of 10? What visual cues in the graph justify your conclusions?

6) Which maximum depth do you think results in a model that best generalizes to unseen data? What intuition lead you do this answer?

Evaluating Model Performance

7) What is the grid search technique and how can it be applied to optimize a learning algorithm?

8) What is the k-fold cross-validation training technique and how is it performed on a learning algorithm?

9) What maximum depth does the optimal model that you have implemented have? How does this result compare to your guess in Question 6?

10) What price would you recommend each client sell his/her home at, given the data presented in the project? Do these prices seem reasonable given the values for the respective features?

11) In a few sentences, discuss whether the constructed model should or should not be used in a real-world setting.

Starting on **May 30th**, the *Predicting Boston Housing Prices* project will undergo a transition

Submission

into its final revision to improve the student experience. The transition will complete on **June 20th**, leaving students three weeks to complete the old version of the project, if they so choose. Please see **this forum broadcast (https://discussions.udacity.com/t/final-revisions-for-predicting-boston-housing-prices-project/170451)** for additional details.