Introduction

In this project we will create a ML model that can learn from the data set and predict the median housing price in any district, given other metrics.

Design

a. The Model's output will be fed to another ML system along with many other signals. The downstream system will determine whether is worth investing in a certain area or not.
b. Pipelines can make components that makes the system run asynchronously.
c. The census dataset looks like a good source to exploit the median housing price purpose.
d. The dataset contains thousands of districts, and other data.
e. The system should be a supervised learning with labeled training examples. It is also a regression task to find a value from multiple values per district.
f. It is actually a multivariate regression problem that we have.
g. It should use a plain batch learning to do the task since there is no continuous incoming data flow into the system.
h. Consideration about the input data.
i. If the price is not imputed as a value, but as a class (cheap, medium, expensive), than the prediction model used should not be the regression model, but the classification model.

Implementation

Selecting a performance measure

The typical performance measure for regression problems is the Root Mean Square Error (RMSE). This can provide us with the error that the system typically makes in its predictions, with a higher weight for large errors.

The formula is as follows:

Root Mean Square Error (RMSE)

$$\text{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h\left(\mathbf{x}^{(i)}\right) - y^{(i)}\right)^2}$$

Implementation Steps

1. Download the data
2. Data cleaning and fixing
3. Create test dataset
4. Visualize the data to gain insights
5. Look for correlation (closer to 1)
6. Try different attribute combinations
7. Prepare the data for ML
8. Handling Text and Categorical Attributes
9. Feature Scaling (getting attributes scales closer)
10. Training and evaluating
11. Using Cross Validation
12. Grid Search
13. Randomized Search (for many hyperparameters)
14. Ensemble Methods
15. Analyze best model and their errors (which features to drop)
16. Evaluate the system on the test set

Bibliography / References

https://learning.oreilly.com/library/view/hands-on-machine-learning/9781492032632/ch02.html#download_the_data