Introduction to Machine Learning
Lab 2: Dataset Preparation and Regression

Danna Gurari

January 24, 2018

Generate Dataset

1. Create a linear function with noise (example is on page 110 of Hands-on Machine Learning with Scikit-Learn & TensorFlow by Aurelien Geron).

2. Display the data you created. The following tutorial offers greater detail about plotting using matplotlib in Python.

Split Data into a Train Dataset and Test Dataset

3. Use this Python function to automatically split your data into the following four parts: (1) training features, (2) training labels, (3), testing features, and (4) testing labels. Importantly, this function randomizes the order of samples before partitioning them.

Apply Regression Models

4. An excellent overview of regression models supported in scikit-learn is located here.

5. The following links provide greater detail about the models you will need to use in your lab assignment:

   - Linear model
   - Ridge regression model
   - Lasso regression model

6. Visualize predicted versus actual values for the above three models in the same plot.

Evaluate Regression Models

7. An excellent overview of evaluation metrics supported in scikit-learn is located here.

8. Below is a list of metrics you will use in your lab assignment with extra links about metrics that are not discussed in the aforementioned tutorial. Gain experience using them with the linear model you created:

   - Correlation Coefficient
   - Coefficient of determination
   - Mean Absolute Error
   - Mean Squared Error

Load Real-World Data

9. scikit-learn datasets: load the Boston Housing dataset.

10. Note that many more datasets can be found at the links listed on pages 33-34 of Hands-on Machine Learning with Scikit-Learn & TensorFlow by Aurelien Geron.