CPSC 483 - Introduction to Machine Learning

Project 1, Fall 2020

due September 28 (Section 02) / October 1 (Section 01)

Last updated Sunday September 20, 2:20 pm PDT

So far we have seen linear regression, but not yet derived an analytical solution to find the estimated parameters w_0, w_1, w_2, … Therefore we will experiment with the scikit-learn library's implementation of the algorithm.

In the next project we will work with a lower-level NumPy implementation of the algorithm that directly uses the analytical solution derived in Section 1.3 of the textbook.

This project may be completed individually, or in a pair as long as both students are enrolled in the same section of the course.

Platforms

For this project (and, in general, for most machine learning or data science projects) you will need a Jupyter notebook with Python 3. Jupyter allows you to create documents mixing text, equations, code, and visualizations.

Alternatively, you may use the newer JupyterLab IDE, which has additional features but a more complex user interface.

The Jupyter project itself recommends Anaconda if you intend to run notebooks locally on a laptop or desktop computer, but there are several cloud services that offer free Jupyter notebooks, including Microsoft Azure Notebooks and Google Colaboratory.

Libraries

You will need scikit-learn to obtain the data and run linear regression. You may also wish to use pandas DataFrames to examine and work with the data, but this is not a requirement. Use Matplotlib's pyplot framework or pandas to visualize your results.

You may reuse code from the Jupyter notebooks accompanying the textbook and from the documentation for the libraries. All other code and the results of experiments should be your own.

Dataset

The scikit-learn `sklearn.datasets` module includes some small datasets for experimentation. In this project we will use the Boston house prices dataset to try and predict the median value of a home given several features of its neighborhood.

See the section on scikit-learn in Sergiy Kolesnikov's blog article Datasets in Python to see how to load this dataset and examine it using pandas DataFrames.

Experiments

Run the following experiments in a Jupyter notebook, performing each action in a code cell and answering each question in a Markdown cell.

1. Load and examine the Boston dataset's features, target values, and description.
2. Create a scatterplot showing the relationship between the feature LSTAT and the target value MEDV. Does the relationship appear to be linear?
3. Create and `fit()` an `sklearn.linear_model.LinearRegression` model using LSTAT as a predictor of MEDV. Using the `coef_` and `intercept_` attributes of the model, what is the equation for MEDV as a function of LSTAT?
4. Use the `predict()` method of the model to find the response for each value of the LSTAT attribute in the dataset. Using `sklearn.metrics.mean_squared_error()`, find the average loss $\mathcal{L}$ for the model.
5. Add a line to your scatter plot representing the least squares fit to the data. How well does the model fit the data?
6. Now repeat experiments *(3)* and *(4)* using all 13 input features at the same time. How does the average loss change?
7. Based on the `coef_` attributes of the new model, which features are desirable in a home? Which features detract from its value?
8. Given the `coef_` attributes, find the following for each feature: how much does a one unit increase in that feature change the median value of the home? Based on the description of the dataset, convert your answer to dollars.
9. Based on the amount of change in the value of the home, which features don't seem to be important?

Submission

Submit your Jupyter `.ipynb` notebook file through Canvas before class on the due date. Your notebook should include the usual identifying information found in a `README.TXT` file.

If the assignment is completed by a pair, only one submission is required. Be certain to identify the names of both students at the top of the notebook.