Kenytt Avery, Computer Science Department, California State University, Fullerton

CPSC 483 - Introduction to Machine Learning

Project 2, Spring 2021

due March 11 (Section 01) / March 12 (Section 02)

Last updated Tuesday, March 9, 12:00 am PST

In this project you will use NumPy to implement vectorized linear and polynomial regression models and compare their performance using separate training and test sets.

The project may be completed individually, or in a group of no more than three (3) students. All students on the team must be enrolled in the same section of the course.

Platforms

The platform requirements for this project are the same as for Project 1.

Libraries and Code

Vector and matrix operations for this project must be implemented in NumPy and results visualized with Matplotlib's pyplot framework. You may not use any other library except the Python Standard Library.

Code from *A Whirlwind Tour of Python*, the Jupyter notebooks accompanying the textbook, and from the library documentation may be reused. All other code and the results of experiments must be your own original work or the original work of other members of your team.

Dataset

The file `boston.npz` contains a version of the Boston house-price dataset in NumPy `.npz` format. See http://lib.stat.cmu.edu/datasets/boston at the CMU StatLib Datasets Archive for a description of the data.

Experiments

Run the following experiments in a Jupyter notebook, performing actions in code cells and reporting results in Markdown cells.

1. Use the NumPy `load()` method to read the dataset. The data contains two arrays: `'features'`, which contains the variables CRIM through LSTAT, and `'target'`, which contains the variable MEDV.

2. Set aside the first 102 items (20% of the total) as a test set, and the remaining 404 items for training.

3. Create a scatterplot of the training data showing the relationship between the number of rooms and the median value of a home. Does the relationship appear to be linear?

4. With RM as ***X*** and MEDV as ***t***, use `np.linalg.inv()` to compute ***w*** for the training set. What is the equation for MEDV as a linear function of RM?

5. Use ***w*** to add a line representing the least squares fit to your scatter plot from experiment *(3)*. How well does the model appear to fit the training set?

6. Use ***w*** to find the predicted response for each value of the RM attribute in the training set, then compute the average loss $\mathcal{L}$ for the model.

7. Repeat experiment *(6)* for the test set. How do the training and test MSE values compare? What accounts for the difference?

8. Repeat experiments *(4)*, *(6)*, and *(7)* using all 13 input features as ***X***. How do the training and test MSEs for this model compare to the values you found for experiment *(7)*? What accounts for the difference?

9. Using the value that you found for ***w*** for this new model, determine for each feature how much a one unit increase in that feature would change the median value of a home. Based on the description of the dataset provided by StatLib, convert your answer to dollars.

10. Based on the amount of change in the value of a home, which features are most important?

Submission

A Markdown cell at the top of the notebook should include project summary information as described in the Syllabus for README files.

Since you may be actively editing and making changes to the code cells in your notebook, be certain that each of your code cells still runs correctly before submission. You may wish to do this by selecting *Run All* from the drop-down menu bar.

Submit your Jupyter `.ipynb` notebook file through Canvas before class on the due date.

If the assignment is completed by a team, only one submission is required. Be certain to identify the names of both students at the top of the notebook. See the following sections of the Canvas documentation for instructions on group submission:

- How do I join a group as a student?
- How do I submit an assignment on behalf of a group?