CS412 - Machine Learning: Homework 2

Due: March 16, 2025 (23:55)
Late Accepted Until: March 18, 2025 (23:55)

Goals

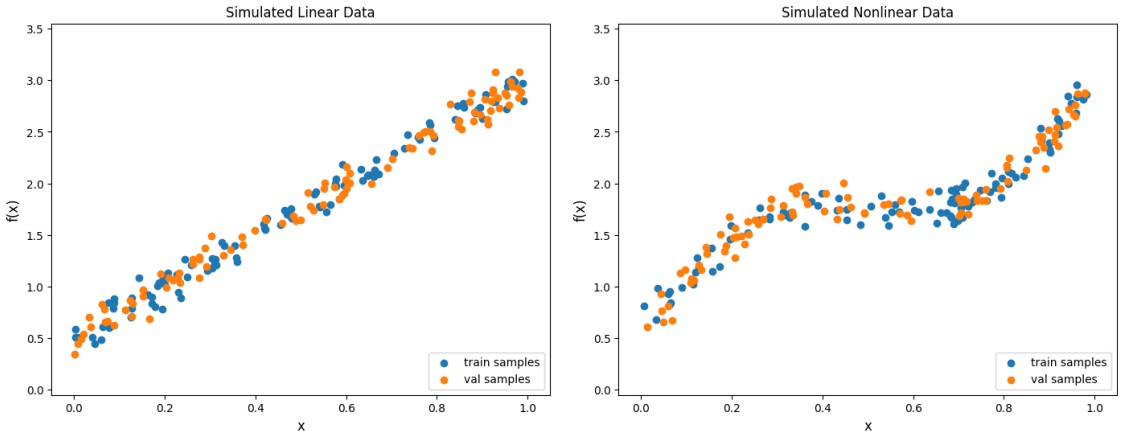
The goal of this homework is four-fold:

- Introduction to linear regression, a fundamental machine learning method for modeling the relationship between variables.
- Gain experience in using/implementing the least squares solution and gradient descent approach using NumPy.
- Learn to apply polynomial regression to model nonlinear relationships between variables.
- Gain experience in constructing the polynomial expansion of the data matrix, and using/implementing the least squares solution.

Datasets

The first dataset is a collection of (x, y) pairs, where the x values are uniformly sampled from a given range, and the y values are generated from a linear function. The linear function is corrupted with random Gaussian noise to make the data more realistic. There is only one input feature in this dataset.

The second dataset is also a collection of (x, y) pairs, but the y values are generated from a nonlinear function. As with the first dataset, the x values are uniformly sampled from a given range, and the y values are corrupted with random Gaussian noise.



(a) Dataset 1 (b) Dataset 2

Figure 1: Scatter plots of Dataset 1 and Dataset 2

Starter Notebook Link

Link for the starter notebook: https://colab.research.google.com/drive/1acd3buFfmcqGCpUJv-0M9Cq7v6D9Zphq?usp=sharing. Please make a copy on your own drive and work on that.

Report

- Write a report, that should read like a project report. So, have a well structured report which includes a cover page (where you should write your name, ID number, the course name, HW No. etc.), an introduction section, a conclusion section, and of course all of the methods and results from your notebook.
- Have well structured headers in your report so that we can easily understand which method's result and discussion you are providing. Also provide which section it corresponds to in the starter notebook, which we are providing in this report. e.g. "Applying Manual Gradient Descent on Dataset 1 (Part 1.c)"
- Make sure to include all visualizations, line fits and loss curves in your report. Each plot should be provided under its corresponding header.
- Discuss your results. How well did the your lines fit? How well do the different methods in Part 1 perform compared to each other? Similarly how well do the different methods in Part 2 perform compared to each other? Any key findings or interesting outcomes?

Submission Guideline

You are provided with a starter notebook link (see above in the *Starter Notebook Link* section). Further guidelines for the report are included at the end of this notebook as well.

- **Jupyter Notebook:** Include all code cells and output. (Ensure that all outputs remain in the notebook in the given order, as it will not be re-run during grading.)
- **Notebook Link:** At the top of your PDF report, include the shareable link to your notebook (accessible via the Share button on the top right). Place this link at the beginning of your report. **The link should be set so that anyone with the link can access it.**
- **Submission Files:** Submit the following:
  - Your Jupyter Notebook as `CS412-HW2-YourName.ipynb`
  - Your PDF report as `CS412-HW2-YourName.pdf`
- **Late Submissions:** **Late submissions will be accepted for up to two days with a penalty of 10 points per day.**

Task

In this homework, you will be given two datasets, which are generated using different underlying functions. Your task is to perform linear and polynomial regression on these datasets and compare the performance of these two methods. More specifically, you will need to complete the following tasks:

Part 1 - Linear Regression on Dataset 1

Using the starter notebook we provide, generate the first dataset given in the Datasets section. You should generate the indicated number (N) data points. Then split the dataset into training and validation sets. Use 50% of the data for training, and the remaining 50% for validation.

This part consists of three linear regression implementations:

- In Part 1.a, you will use the scikit-learn library's linear regression method.
- In Part 1.b, you will implement the ordinary least squares (OLS) algorithm manually, using the pseudoinverse method.
- Finally in Part 1.c, you will implement the gradient descent algorithm to find the regression coefficients.

Part 2 - Polynomial Regression on Dataset 2

In this part, instead of generating the data, you will be reading the data for Dataset 2 from the .npy files provided with the homework. The .npy file format is a file format used in Python to store arrays and matrices of numerical data, optimized for use with the NumPy library. You will then split the dataset into training and validation sets, using the same split ratio as mentioned in Part 1.

- In Part 2.a, you will use the scikit-learn library to perform the polynomial regression method, using polynomial degrees of 1, 3, 5, and 7.
- In Part 2.b, you will implement the polynomial regression algorithm manually only for degree 3.

For all datasets and regression methods: Compute the mean squared error (MSE) on the validation set for each dataset and each regression method using the following formula:

$$MSE = \frac{1}{2N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

where $\hat{y}_i$ is the predicted value, and y_i is the true value of the i-th sample, respectively. The number of samples is N.

Questions?

- You should ask all your Google Colab-related questions to Discussions and feel free to answer/share your answer regarding Colab.
- You can also ask/answer about which functions to use and what libraries...
- However, you should not ask about the core parts, that is what is validation/test, which one should have higher performance, what are your scores, etc.