CS412 - Machine Learning
Homework 2

Due: April 8, 2023 **Late Accepted Until:** April 10, 2023

Starter Notebook Link

https://colab.research.google.com/drive/1C2B6xbYp4wScS3ZPbcC3pKOcQzDbGCqI?usp=sharing

Goals

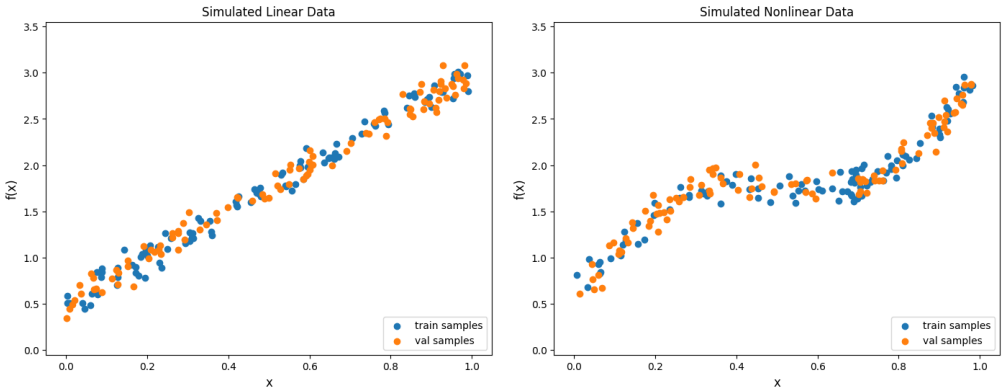
The goal of this homework is four-fold:

- 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

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 both 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 each dataset and each regression method, compute the mean squared error (MSE) on the test set. Use the following formula to compute the MSE:

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

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.

Submission Guideline

You will be supplied a **starter notebook** that you will need to fill. Further guidelines for the report are included at the end of this notebook.

- Your report should include visualizations of the generated datasets, as well as the results of the linear regression and polynomial regression models. Provide a table of your results.
- Put the report part of your notebook (see the Report part of the notebook) separately in a PDF document and include a link to your notebook at the top of your PDF (make sure to include the link obtained from the Share button link on the top right), as a PDF file.
- Submit your PDF report to SUCourse - with the name: **HW2-CS412-Yourname.pdf**. Also, download your .ipynb file and submit it to SUCourse - with the name: **HW2-CS412-Yourname.ipynb**. Submit these two files without zipping them.

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 shd. have higher performance, what are your scores, etc.