Non-Linear Regression and Regularisation

Note: It is a graded assignment with a duration of 48 hours hard deadline.

Problem 1: [7 marks]

1. Use California Housing Dataset:
   from sklearn.datasets import fetch_california_housing
2. Perform an 80:20 train/test split, followed by an 80:20 train/validation split from the trainset. Perform standard Scaling on all splits.
3. Perform Linear Regression and Non-Linear Regression using an appropriate Polynomial curve to reduce error on the Train and Validation Datasets. To prevent Overfitting, use L2 Regularizer with appropriate Lambda values. (Optimize for 100 Epochs)
4. Plot the loss vs epoch curve on different parameters of Polynomial degree and Lambda values on the Train and validation sets to choose appropriate values.
5. Predict on the Test set for any top-3 fine-tuned models.
6. Print SSE, R2 scores for Train, Validation, and Test Sets for top-3 Models

Problem 2: [3 Marks]

1. Use Our World in Data (OWID) COVID-19 dataset from the link below:
   https://ourworldindata.org/coronavirus/country/india
2. As we know, in the initial months from 1st Mar 2020 to 31st May 2020, COVID cases spiked at an exponential rate. (hint:
3. Assume an exponential model for the growth as below:
   $$y=A \cdot e^{Bx}$$
4. Use the extracted data for the above-mentioned period to fit a Linear Regression model after transforming the data to a log scale. (hint: convert the dates in the given range to days 1,2,,3,...)
5. Plot the Actual Data and predicted data in log scale using a scatter plot.
6. Print the SSE between the actual and predicted.