Fondations of Machine Learning Assignment 1

Linear Regression

DA24D402, Shuvrajeet Das

September 22, 2024

Contents

List of Tables

List of Figures

Setup:

You are given a data-set in the file *FMLA1Q1Data train.csv* with 10000 points in ($\mathbb{R}^2$, $\mathbb{R}$) (Each row corresponds to a datapoint where the first 2 components are features and the last component is the associated y value).

0.1 Question 1

Write a code to obtain the least squares solution w_{MLE} to the regression problem using the analytical solution.

Answer:

See the referred code

The losses are detailed as follows:

Train Loss	Test Loss
123.4761	66.2564

The w_{MLE} is $[1.7677, 3.5238, 9.8974]$

0.2 Question 3

Code the gradient descent algorithm with suitable step size to solve the least squares algorithms and plot $\|w_t - w_{MLE}\|_2$ as a function of t. What do you observe?

Answer:

See the referred code: parameters, $\alpha = 0.0001$, epochs=20,000

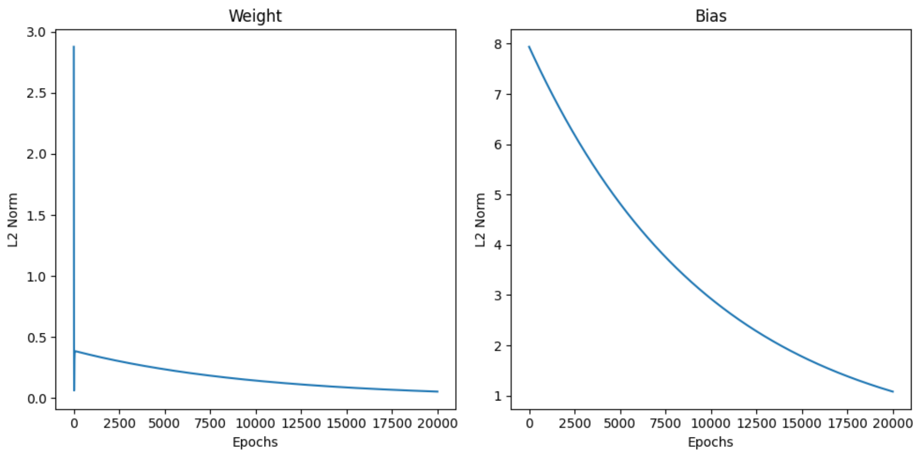

Figure 1: $\|w_t - w_{MLE}\|_2$ vs t plot for Linear Regression with gradient descent

The losses are detailed as follows:

Train Loss	Test Loss
124.6388	65.2987

The w_{GD} is $[1.7322, 3.5627, 8.8178]$

As far as the observation goes, the weight does reduce and the norm difference between w_{MLE} and w_t is converging to 0 also for the bias term too.

0.3 Question 3

Code the stochastic gradient descent algorithm using a batch size of 100 and plot $\|w_t - w_{MLE}\|_2$ as a function of t. What do you observe?

Answer:

See the referred code: $\alpha = 0.0001$, epochs=20,000, batch size=100

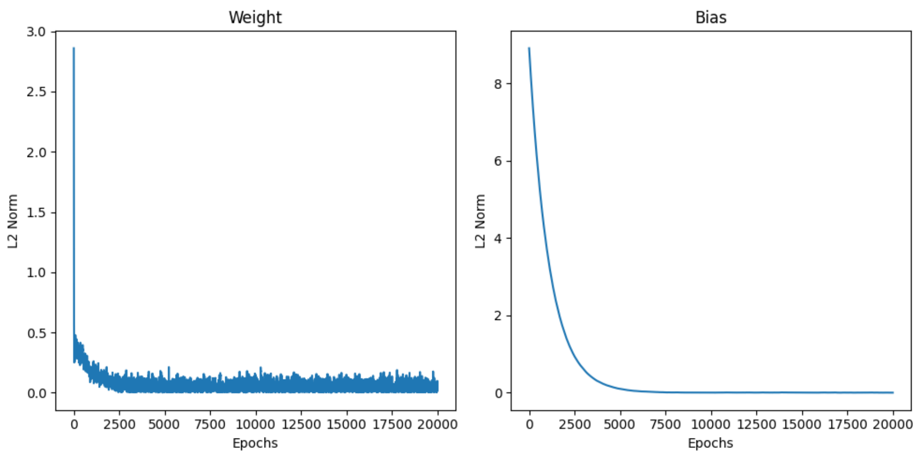

Figure 2: $\|w_t - w_{MLE}\|_2$ vs t plot for Linear Regression with stochastic gradient descent

The losses are detailed as follows:

Train Loss	Test Loss
123.4811	66.4079

The w_{SGD} is $[1.78473.45529.8976]$

Here in this case there is a disturbance in the graph due to the stochasticity like choosing the batch size but it was able to converge much faster as the L2 norm is quite low.

0.4 Question 4

Code the gradient descent algorithm for ridge regression. Cross-validate for various choices of λ and plot the error in the validation set as a function of λ. For the best λ chosen, obtain w_R. Compare the test error (for the test data in the file *FMLA1Q1Data test.csv*) of w_R with w_{MLE}. Which is better and why?

Answer

See the referred code: α= 0.0001, epochs=20,000, batch size=100, $\lambda = [1e-4, 1e-3, 1e-2, 1e-1, 0, 1, 10, 100]$, k=5 (cross validation size)

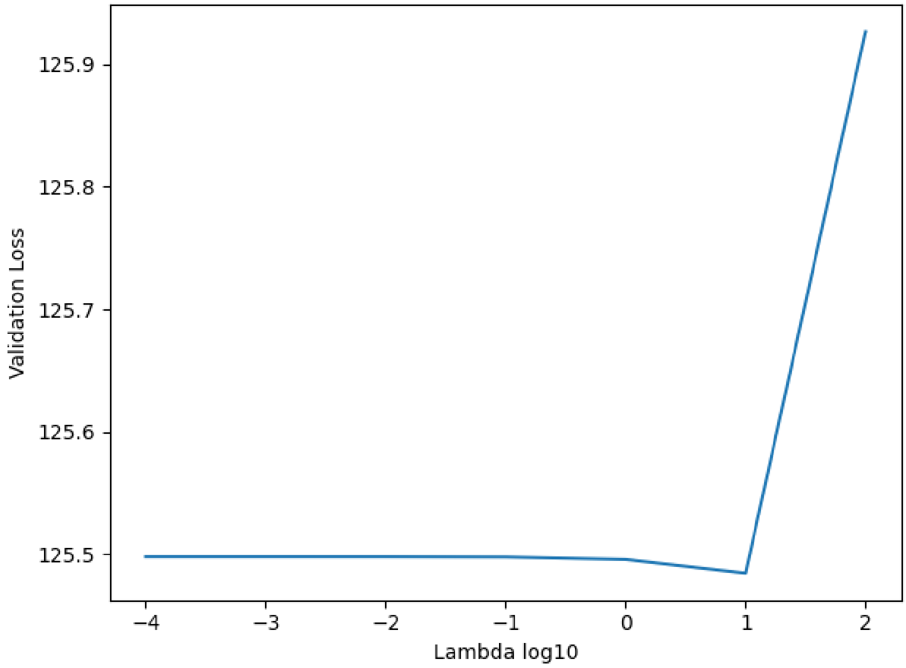

Figure 3: Plotting of validation loss with $log_{10}\lambda$

The losses are detailed as follows:

Train Loss	Test Loss
123.6681	66.7849

The $w_{\lambda=10}$ is $[1.4745, 3.1965, 9.9352]$
The Ridge regularization and the Maximum likelihood estimation differ very little. Still, choosing the proper lambda there may or may not be a circumstance where Regularization performs better than the ML Estimate. But in my case the w_{MLE} is slight better than w_λ.

0.5 Question 5

Assume that you would like to perform kernel regression on this dataset. Which Kernel would you choose and why? Code the Kernel regression algorithm and predict the test data. Argue why/why not the kernel you have chosen is a better kernel than the standard least squares regression.

Answer:

I've used polynomial kernel as a start to work with and had seen it had performed very well, and gave good results, (since the data was 3D plotting the data helped me to find the perfect kernel)

The code can be referred.

The polynomial kernel I've chosen is good because it helped me to get the inner product of the non-linear expansion of the features that further helped the model to become better and allowed it to give better results. In my case, I've chosen a polynomial kernel of degree 2 with coeff 1.

The losses are detailed as follows:

Train Loss	Test Loss
0.01007	0.0098

References