Assignment on Regression

Question 1.

Implement Linear Regression for the Abalone Dataset. The dataset contains 9 variables out of which the last column is the output variable and the other 8 are input attributes. You need to implement gradient descent from scratch. Normalize the data and choose an appropriate learning rate. Divide the dataset into training and testing (80-20)

(a) Include plots for the root mean squared error (RMSE) vs. gradient descent iterations for both training as well as validation set

(b) Also implement the normal equation (closed form) for linear regression and get the optimal parameters directly. Compute the RMSE (after getting the optimal parameters) for both training and testing set and report them.

(c)  Compare the final RMSE obtained from (a) after convergence and the RMSE from (b) and make a note of any observations you might have.

Question 2:

Use 5-fold cross validation with grid search on the training set (from the above dataset) to find the appropriate regularization parameter (hyperparameter). You may use Ridge, Lasso and GridSearchCV routines from the sklearn library to perform the following:

(a) Find the optimal regularization parameter for L2

(b) Find the optimal regularization parameter for L1

Write the above two regularization parameters in the report.

Once the optimal regularization parameters for both L2 and L1 have been found out, modify the gradient descent algorithm implemented in (1) to accommodate for the L2 and the L1 regularization term.

Use the values of the “optimal” regularization parameter found out in (a) and (b) and use it with the modified gradient descent algorithm and plot the RMSE error vs iterations curve. Also report the RMSE on the test set. [Two plots – one for L1, one for L2]

Question 3:

Use the data in file head\_brain.csv that contains only 1 input variable and 1 output variable i.e. the brain-weight to the body-weight proportion for varying species. Consider the dataset as a whole i.e. do not split it into train, val or test. Perform the following tasks:

(a) Plot the data points using a scatter plot along with the best fit line found out using linear regression (without regularization).

(b) Use L2 regularization and plot the data points and the new best fit line.

(c) Use L1 regularization and plot the data points and the new best fit line.

Use the gradient descent algorithm implemented in (1 & 2) to perform regression on the whole dataset in all these 3 parts.

Compare how the best fit line changes visually with adding different kinds of regularization, was it a better fit, worse fit than the regression performed without regularization?