Lab 4 – Linear Regression

MACHINE LEARNING

SAUMAY AGRAWAL
16BCE1151

EXPERIMENT

- Implement linear regression, try with sklearn
- Try additional regression types given in the book by Sebastian.
- Compare the performance of the various algorithms on the same dataset(yours) and infer from the results.

ALGORITHMS

LINEAR REGRESSION

Linear regression is a basic and commonly used type of predictive analysis. The overall idea of regression is to examine two things: (1) does a set of predictor variables do a good job in predicting an outcome (dependent) variable? (2) Which variables in particular are significant predictors of the outcome variable, and in what way do they–indicated by the magnitude and sign of the beta estimates–impact the outcome variable? These regression estimates are used to explain the relationship between one dependent variable and one or more independent variables. The simplest form of the regression equation with one dependent and one independent variable is defined by the formula y = c + b*x, where y = estimated dependent variable score, c = constant, b = regression coefficient, and x = score on the independent variable. However, it can also be extended to multiple dependent variables also.

LINEAR REGRESSION USING GRADIENT DESCENT

We have used the gradient descent algorithm, to minimize the cost function in our implementation. The cost function is nothing but a Sum of Squared Errors (SSE), which helps us in getting the best fitting regression line. We can specify the learning rate and epochs (apart from defaults) for the gradient descent algorithm, so that the parameters are best suited to our dataset.

RIDGE AND LASSO REGRESSION

Ridge and Lasso are another names for L1-regularisation and L2-regularisation respectively. Regularisation involves minimizing the number of misclassified examples while also minimizing the magnitude of the parameter vector. The main difference lies in the calculation of norm (vector magnitude) for the regularisation procedure. L1 norm is calculated using taxicab norm (manhattan distance) while L2 norm is calculated using Euclidean norm (ie. Euclidean distance).

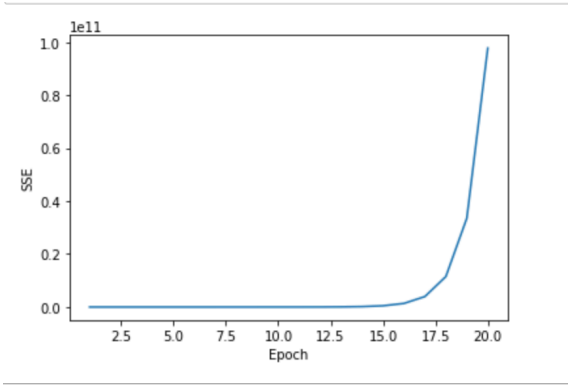
RANSAC REGRESSION

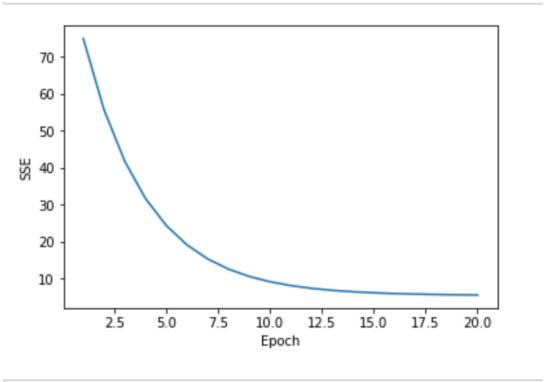
Random sample consensus (RANSAC) is an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers, when outliers are to be accorded no influence on the values of the estimates. Therefore, it also can be interpreted as an outlier detection method. A basic assumption is that the data consists of "inliers", i.e., data whose distribution can be explained by some set of model parameters, though may be subject to noise, and "outliers" which are data that do not fit the model. The outliers can come, for example, from extreme values of the noise or from erroneous measurements or incorrect hypotheses about the interpretation of data.

OBSERVATIONS

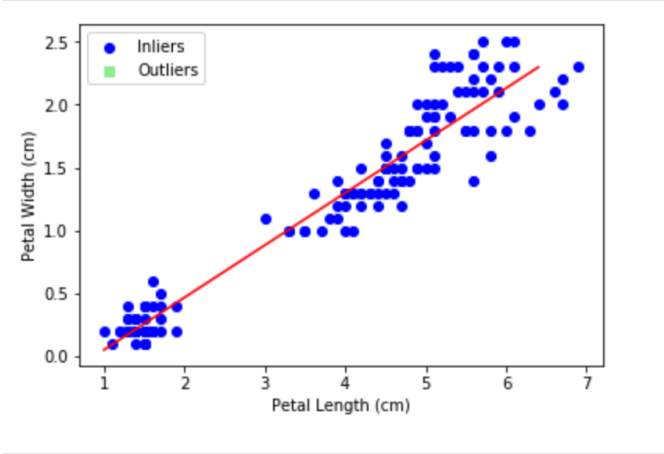
IRIS DATASET

- Upon plotting the correlation matrix for the attributes of iris dataset, it was found that the attributes “Petal Length” and “Petal Width” had the maximum correlation (0.96).
- The SSE (cost function) didn’t converge with the values of the dataset. Instead, it diverged after approx. 15 epochs. (figure on the left below)
- To tackle this, feature scaling was performed on the values of both the attributes. For these values, the cost function diverged appropriately, resulting into a regression line. (figure on the right below)



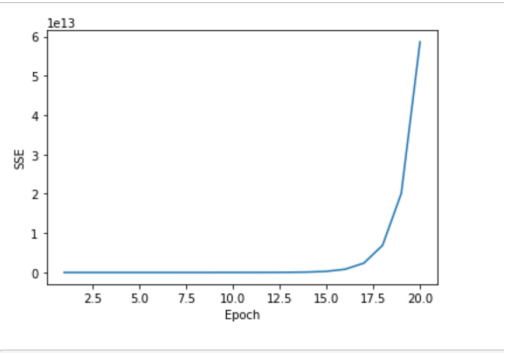


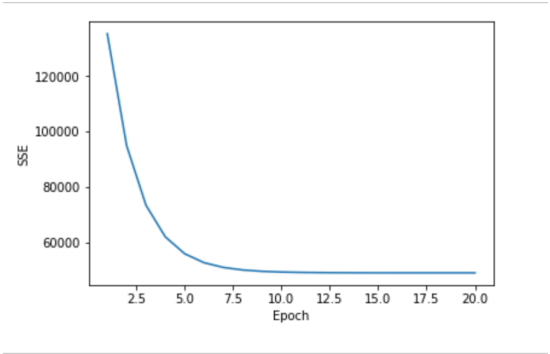
- Feature scaling is done for the purpose of faster convergence and better interpretation of values. Here, it also increased the accuracy in the sklearn implementation of linear regression. (given in the accuracy comparison table)
- No outliers were found with respect upon implementation of the RANSAC algorithm (figure below)



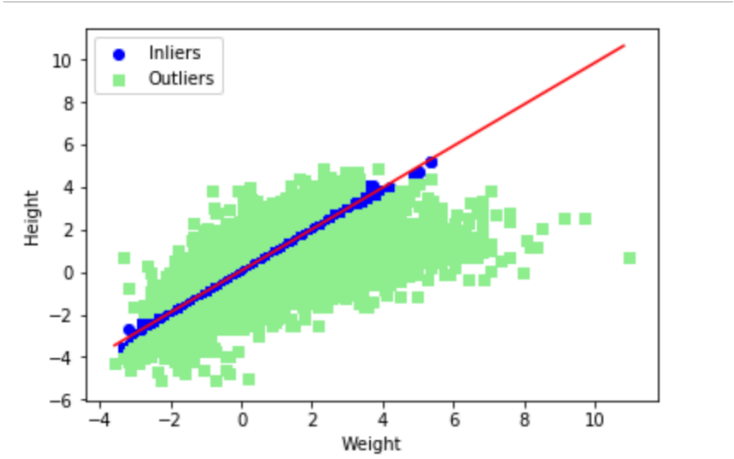
OLYMPICS DATASET

- Upon plotting the correlation matrix, the attributes “Height” and “Weight” were found to have maximum correlation (0.8)
- The gradient descent algorithm didn’t work for actual as well as the scaled values for these attributes. For both, the cost function diverged, as the learning rate was higher as per the dataset.
- The results were visible only after reducing the learning rate. The learning rate was set to 0.00001 for 20 years and 40 years data, and 0.000001 for entire 120 years data, which is 1/100 th and 1/1000 th times the default value of learning rates respectively.





- Feature scaling didn't help with accuracy values for this dataset.
- Almost 40% of the values were classified as outliers by the RANSAC algorithm. (figure below)



ACCURACY COMPARISON TABLE

Regression Algorithm	Iris data	Olympic (20 years)	Olympic (40 years)	Olympic (120 years)
Linear	0.92	0.63	0.64	0.64
Linear (scaled)	0.94	0.62	0.63	0.64
Ridge (scaled)	0.94	0.62	0.63	0.64
Lasso (scaled)	0.91	0.60	0.62	0.64
RANSAC	0.927	0.599	0.597	0.603

INFERENCE

- The linear regression algorithm is a descent algorithm for a continuous scale of values.
- The more the number of outliers in a data set, the lesser the accuracy of linear regression will be.
- So if the attributes have a very high correlation (like >0.95), then only this method will produce usable results.