Linear Regression

Abstract

Linear Regression is the oldest statistical model and a simple Machine learning algorithm to predict data between independent and dependent variables. We'll be using Ordinary Least Square estimator to calculate the coefficient and predict the value of new test data. Finally, we'll also train the model using N-Cross Validation method and see which N gives us the best accuracy score.

Concept

Simple Linear regression is done between two vectors of n_samples, where the best fit line is given by formula:

$$Y = w0 + w1*X$$

Which is similar to algorithm to find the slope

$$Y = mX + c$$

Where
w1 = mean(y) – (w1*mean(x))
&
w0 = covariance(x,y)/variance(x)

This is the concept to find predicted values Y after calculating the coefficient values w0 and w1 for one independent variable and a dependent variable.

Multivariate Linear Regression

This type of Linear Regression is nothing but Regression with two or more independent feature variables.

X (Feature Matrix) => size of n X p

$$\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$$

Y (Response Vector) => size of n

$$y = \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix}$$

B0 , B1, B2 ……Bp are coefficient and we use this to assert our Hypothesis (**h(Xi))** which in our case is whether the set of give features belong to either of Iris-Setosa, Iris-Versicolor, Iris-Virginica.

$$h(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

We can write the above equation as :

$$\hat{y} = X\hat{\beta}$$

The equation above is used to derive the predicted values based on given input X

Where B vector is calculated using a simple Learning Rule called Ordinary Least Square (OLS)

$$\hat{\beta} = (X'X)^{-1}X'y$$

We then calculate the Residual Error and using the Residual Error we calculate the Root Mean Square,

Minimize (W0, W1)

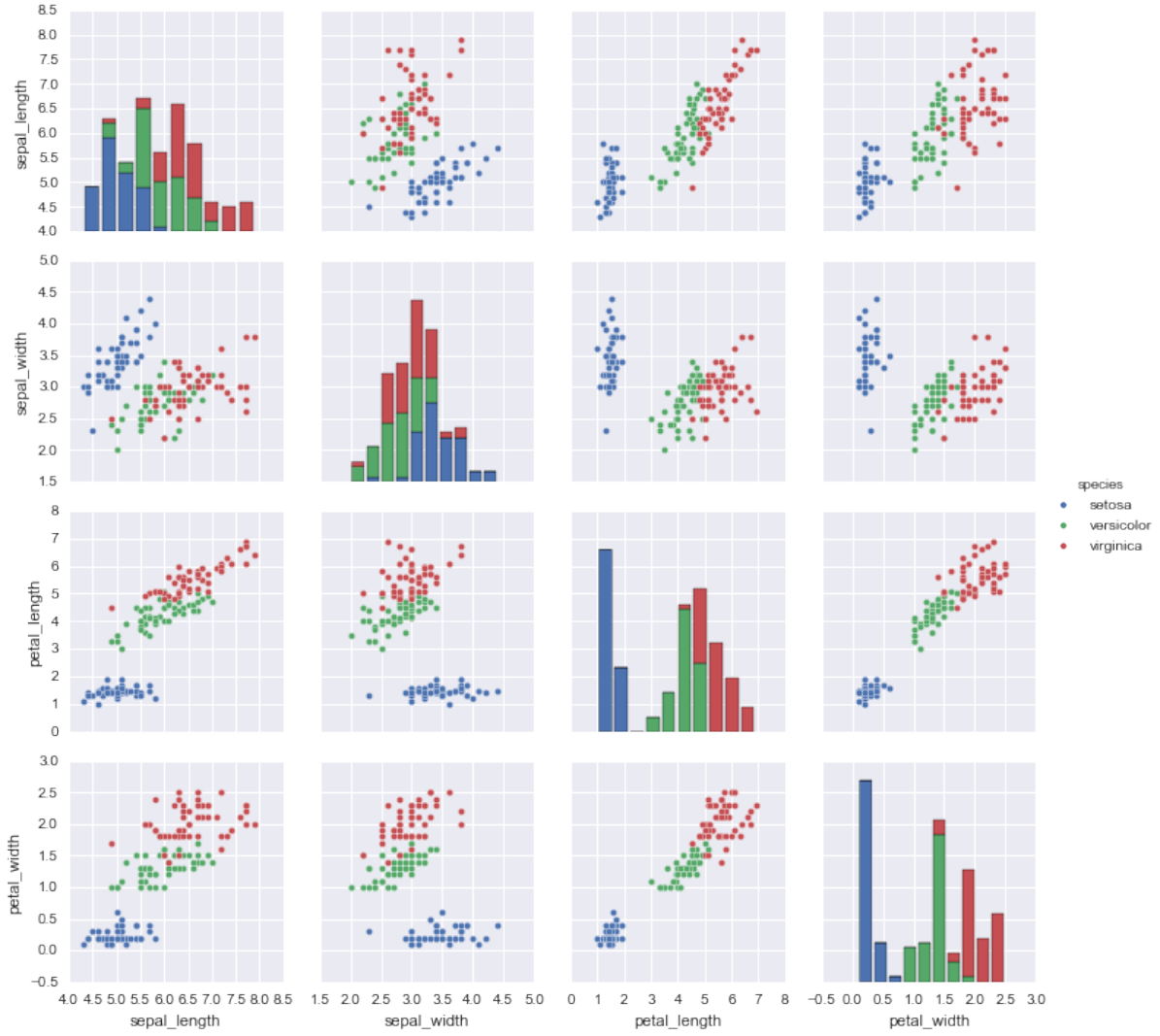
1/2m [Sum from i=1 to i=m (The predicted value (Denoted as h(xi)) – The target)?

where m is the number of Samples.

We Use K-Cross Validation to Train and Test the given data. K-Fold Cross Validation is technique to split the data in K parts where K-1 Matrix is used as Training Data and the rest is used as Test data.

OBERVATION & DATA INTERPRETATION

The feature matrix of Iris dataset was plotted against each other as seen below.



We can notice from here that petal_width and petal_length have some correlation compared to other feature vectors.

We used the least square method to find the Coefficient matrix which can then in turn be used to predict the Y values or Class Labels based on the learning rule algorithm.
Another famous learning rule is the Gradient Decent rule, we won't discuss much about it in this assignment.

Fit() ->
This function takes in a Matrix with Xip value is the feature matrix and the actual class value Y and trains the Linear Regression using Kate

Predict() ->
Predicts what the class label would be for given set of new data using the coefficient calculated earlier.

Norm_class value() ->
This function normalizes the predicted Class Value to its nearest neighbor.

Check_accuracy() ->
This function compares the Actual_Y_value with the Predicted_Y_value and returns the accuracy score in percentage.

Residual_Error() ->
Is a function that computes the Residual_Error by subtracting from predicted value of Y the actual value.

K – Fold Validation ->
Breaking up of data into K components or K parts where K-1 part of data is used as training and the rest is used as testing dataset.

NOTE: I chose K value as **3** as it gave the best accuracy result of **96%**, while K value greater than 3 reduces the accuracy percentage by very little amount.

- Had to shuffle the data before breaking it into parts as same class prediction leads to reduced accuracy.
- 4 value gives 95.333% and as we increase the K value the accuracy percentage decreases.

I used complete feature set to calculate the coefficient values which can be seen in the program. I used random data [Filename: iris_dataset_test.csv] and fed the feature set to predict the class value and compared the Actual and Predicted value. As you can notice the algorithm works really well on full feature set as it gives 100% accuracy rate in predicting the class values.