CSE574 Coding Assignment One

Mohamed Ehab salah Eldin Hassan Lasheen,50542195

mhassanl@buffalo.edu

1. Linear Regression on the wine quality-red Dataset

1. Linear regression problem

It is a simple prediction problem of finding linear relationship between the features of a given data. This relationship can be represented by equation (1)

$$\mathbf{y} = \mathbf{w}^T \mathbf{X} \quad \textbf{(1)}$$

Where:

y is the predicted (target) vector.

X is the input (features) data matrix.

$\mathbf{w}^T$ is the weight vector.

x_i are the features for each sample.

y_i is the true value for each sample.

n is number of samples.

MSE is mean square error.

In linear regression problem, Weight vector is chosen to get the least mean square error (MSE). Such that weight vector is calculated directly for the training data using equation (2). The least square error is calculated using equation (3).

$$\mathbf{w} = (\mathbf{X}^T \mathbf{X})^{-1} \ \mathbf{X}^T \ \mathbf{y} \quad \textbf{(2)}$$

$$\mathbf{MSE} = \frac{1}{n} * \sum_i^n (y_i - \boldsymbol{w}^T x_i) \quad \textbf{(3)}$$

2. Dataset description

The dataset consists of 12 columns of 1599 sample. Eleven columns are used as the input data which are (fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, quality, sulphates, alcohol) to predict the target for the linear regression problem which is the pH of the wine. The true value for pH of wine is in the eighth column. (note I have many trails to choose the target value which gives the minimum error and finally decide to choose pH value as the target value).

3. Data preprocessing

The data is preprocessed by determining the missing cells in the data and calculating their number in each column. Then these rows which contain missing values are dropped. All the data are numerical data type (No string data type). The input data are normalized using the standardization method using equation (4).

$$x' = \frac{x - \mu}{\sigma} \quad \textbf{(4)}$$

Where:

x' *is normalized input data by standardization*

x *is input data*

σ *is standard deviation*

4. Constructing training and testing dataset

The dataset is shuffled randomly to ensure that the training and testing samples are chosen randomly. Then the dataset is divided into two sets training set and testing sets 80 % for training and 20% for testing.

5. Steps of Linear regression

First the weight is calculated by closed form solution to solve linear regression in equation 2 and then the predicted value is calculated using equation 1.

6. Results

The calculated weight vector is shown in equation (5).

$\mathbf{w}^T$ = [-60.83, -0.0996, 0.0156, -0.0133, -0.0268, -0.5218, 0.0019, -0.0008, 64.64, -0.0796, 0.0724, -0.0071] **(4)**

The calculated mean square error between the predicted and target value is shown in equation 5.

MSE = 0.00671 **(5)**

the predicted value is plotted versus the target value as shown in Figure 1. The red line is 45 degrees line which represents perfect prediction (predicted = actual value)

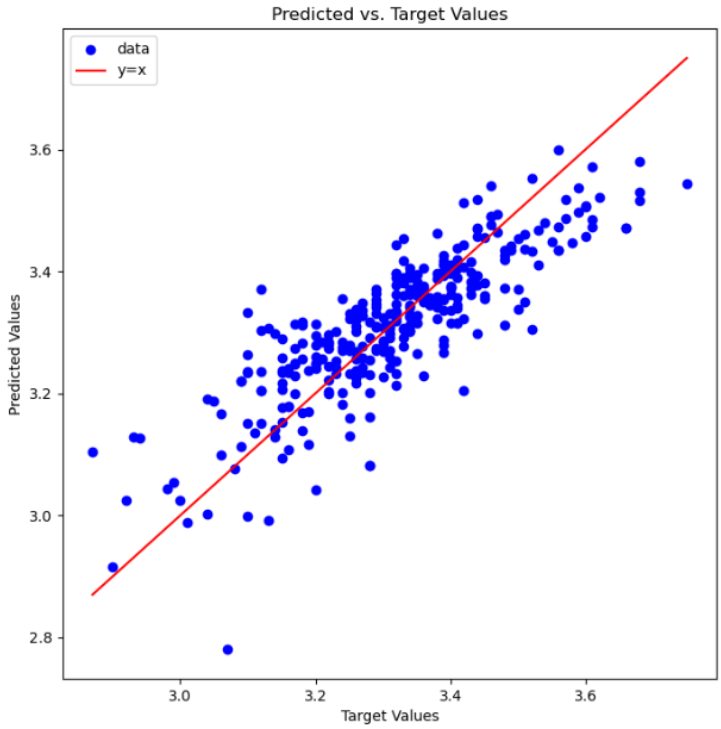

Figure 1 Show the relation between target and predicted values.

7. Discussion

Dataset in "wine quality-red" is fitted using Linear regression model. Linear regression model has closed form solution which is used in calculating the weight vector. Closed form solution gives the weight vector directly without any iteration however this process would be computationally expensive because if the number of samples and number of features increase then the size of input matrix increase, and inverse calculation would be hard and expensive.

The link for python coded script is in the following link:

https://github.com/MohamedHassanLasheen/coding-assigment-1-CES574/upload/main

2. Logistic regression on the Penguins Dataset

8. Logistic regression problem

It is a simple binary classification problem of finding the class in which the input belongs to.

The used activation function for logistic regression is the sigmoid function shown in equation 8. The output of the sigmoid function is always between $\boldsymbol{y} \in (0,1)$ so

$$z = \boldsymbol{w}^T \mathbf{x} + b \quad \textbf{(7)}$$

$$\boldsymbol{y}_{predicted} = \frac{\mathbf{1}}{\mathbf{1} + \boldsymbol{e}^{-z}} \quad \textbf{(8)}$$

Where:

$\boldsymbol{\sigma(z)}$ is the sigmoid function.

e is the exponential

b is the bias

Maximum likelihood estimation is used to determine the optimum weights (w, b) which correspond to the minimum loss. By Taking log for both sides the likelihood estimation is calculated then the likelihood is maximized by using gradient descent and finally the optimum weights are calculated (w,b).

9. Dataset description

The dataset consists of 7 columns of 344 samples. The first six columns are used as the input data (species, island, bill_length_mm, bill_depth_mm, flipper_length_mm, body_mass_g) to predict the target for the logistic regression problem which is the **penguin sex**.

10. Data preprocessing

The data is preprocessed by dropping the rows which contain NaN values in the dataset.
The features with string data type are converted into categorical data type then the categorical data type is converted into binary data type. The non categorical data type (numerical) is normalized using equation (9) so that the values range between zero and one.

$$\text{Normalized } x = \frac{x - min}{max - min} \quad \textbf{(9)}$$

11. Constructing training and testing dataset

The dataset is divided into two sets training set and testing sets (80 % of the dataset for training and 20% of the dataset for testing).

12. Steps of the Logistic regression

The optimum weight (w) is calculated using equation (12) using gradient descent in equation (11) and the average loss function is calculated using equation (10). Three scenarios of different hyperparameters values (learning rate and number of iteration) are considered and the accuracy is determined in each case to determine the optimum hyperparameters.

$$l = \sum_1^m \frac{(y*(log(y_{predicted})+(1-y)*(log(y_{predicted}))}{m} \quad (10)$$

$$\frac{dl}{dw} = (y_{predicted} - y)*(X^T) \quad (11)$$

$$w = w - learning\ rate * \frac{dl}{dw} \quad (12)$$

Where:

m: number of samples

f: number of features

$y_{predicted}$*: predicted value vector*

y: label (true value) can be either 0 or 1

X^T*: feature array size (m, f)*

l: average loss for all samples

$\frac{dl}{dw}$*: is loss derivative with respect to weight.*

W: weight vector

13. Results

For the first scenario the learning rate equals $1e^{-6}$ and number of iterations equal 100,000.

The accuracy of the first scenario = 68.7%. the loss value plotted with each iteration is shown in Figure 2.

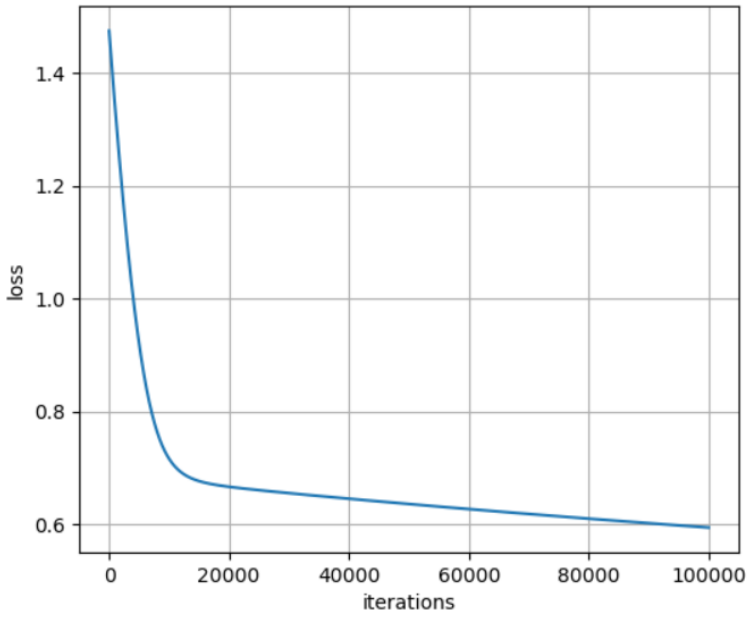

Figure 2 Show the loss value for training logistic regression for scenario I

For the second scenario the learning rate equals $1e^{-3}$ and number of iterations equal 10,000.

The accuracy of the second scenario =89.5%. the loss value plotted with each iteration is shown in Figure 3.

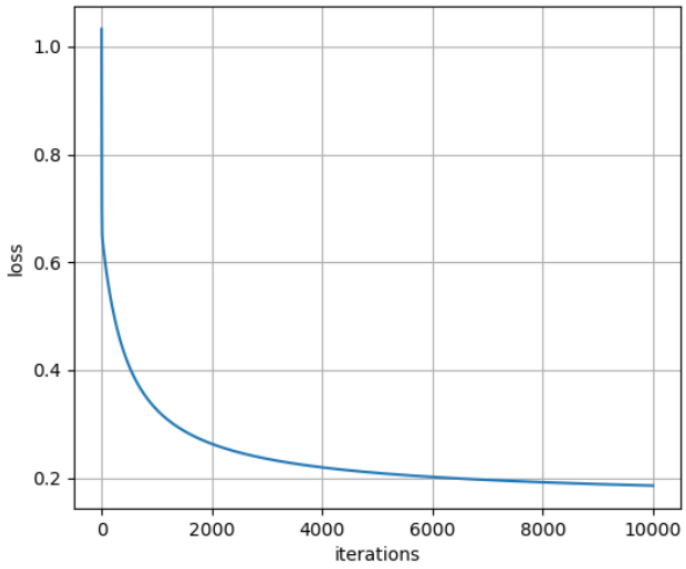

Figure 3 Show the loss value for training logistic regression for scenario II

For the third scenario the learning rate equals $1e^{-2}$ and number of iterations equals 1,000.

The accuracy of the third scenario = 89.5%. the loss value plotted with each iteration is shown in Figure 4.

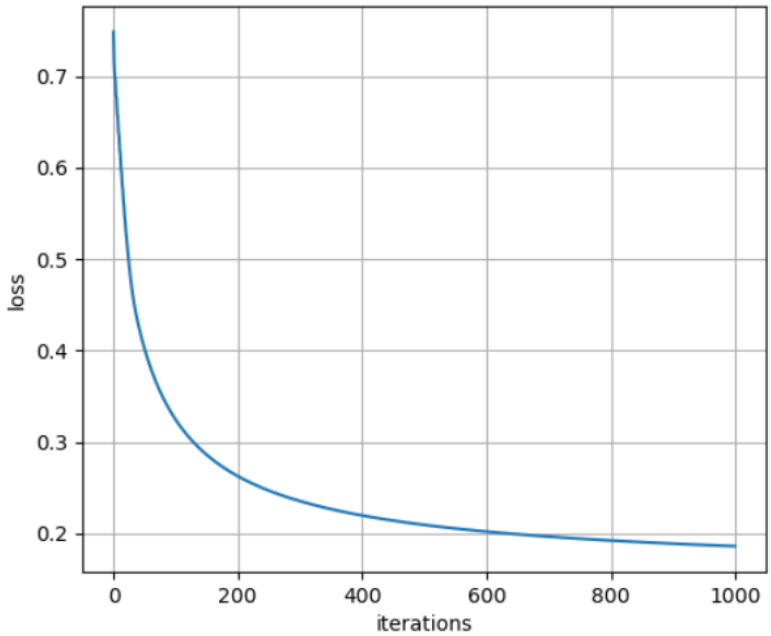

Figure 4 Show the loss value for training logistic regression for scenario III

The accuracy of the three cases is summarized in the table.

	Accuracy
Case I	68.7%
Case II	89.5%
Case III	89.5%

From the previous the best hyperparameters is case III (Learning rate = 0.01, Iteration = 1,000) has the highest accuracy with least learning rate and least number of iterations.

	Weight vector
Case III (best accuracy)	[9.7, 8.8, 2, 16.7, -3.5, -0.2, -12.7]

14. Discussion

Dataset in "Penguins" is classified based on their sex (male, female) using Logistic regression. Three different hyperparameters are used to determine the most accurate hyperparameters. From the results it is shown that case III (learning rate = 0.01 and number of iterations = 1,000) with least learning rate and number of iterations gives the highest accuracy. The main advantage of logistic regression is that is easy to train and interpret, however only used to classify data that are linearly separable so can't be used to solve nonlinear problems.

The link for python coded script is in the following link:

https://github.com/MohamedHassanLasheen/coding-assigment-1-CES574/upload/main