Project3 Report: Binary Classification using Logistic Regression

Simeon Babatunde

Problem Description

Given a TigerFish dataset containing the body length, dorsal fin length and species of fish, we want to create a logistic regression classification algorithm that, given the body length and dorsal fin length of a fish, will predict if the fish is **TigerFish1** or **TigerFish0**.

Data Description

The first line of the dataset contains a single integer indicating how many sets of labelled data we have to work with. Each line after that contains three tab-separated entries. The first is a float representing the body length in centimeters, followed by a float representing the dorsal fin length in centimeters, then an integer identifying the fish as either TigerFish0 (with a 0) or TigerFish1 (with a 1).

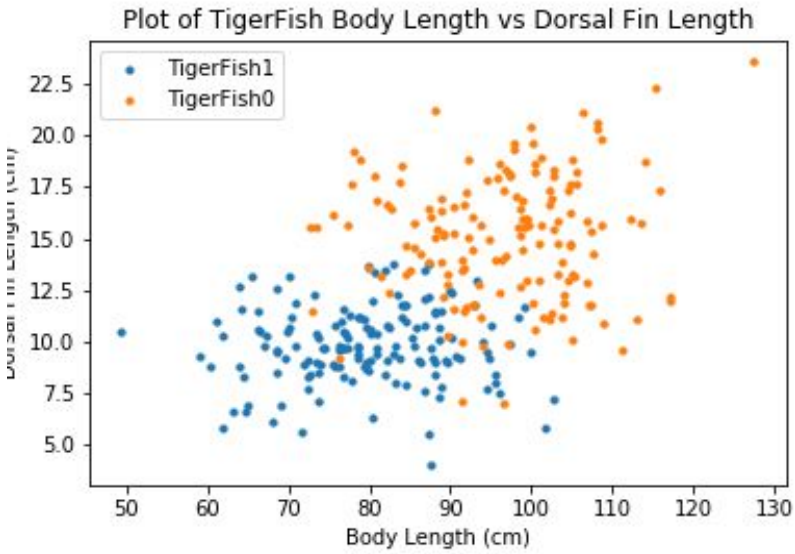

Figure 1. The Initial Dataset

Feature Scaling and Decision Boundary

In order to speed up gradient descent convergence, we scaled the dataset using the standardization technique. $(x_i - \mu)/\sigma$. The feature scaling was carried out after randomizing the dataset. We defined the decision boundary for this algorithms as $\mathbf{w_0 + w_1x_1 + w_2x_2 + w_3x_1x_2}$. The data was split into train (70%) and test(30%).

Model Training

While training the regression model, we computed the error J(w) to define the weights that were being used in the Zigmoid function. Figure 2 shows a plot of error J and number of iterations it takes to minimize the error while. The initial values chosen were; $\boldsymbol{w_0 = 0.5\ w_1 = 1.2\ w_2 = 2.8\ w_3 = 0.8}$, ***alpha = 0.01 J=3.07655227.*** Final values includes; $\boldsymbol{w_0 = -0.32039835972176695\ w_1 =}$

-2.411093368019604 w_2 = -3.3780505239011958 w_3 = 1.084723542839459, alpha = 0.05 J=0.187076148 and a total of **5000** epochs was carried out to ensure minimal error.

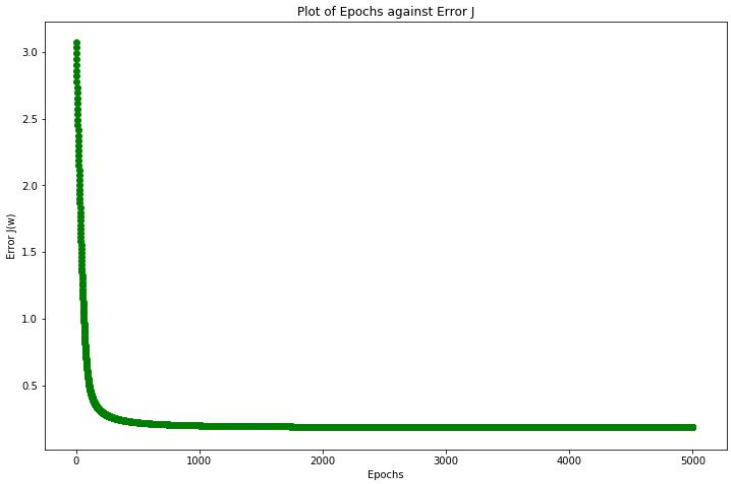

Figure 2. Error J against Number of Iterations

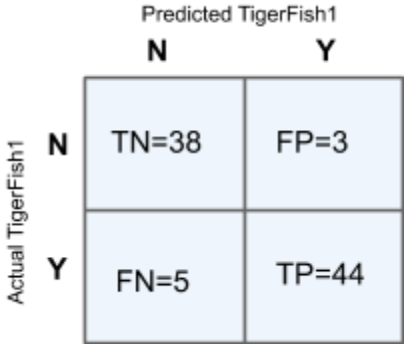

Figure 3. Confusion Matrix

Results

The value of ***J*** on test data is ***0.23679725110557281.*** Figure 3 shows the confusion matrix of the prediction results. The test set consisted of a total of 90 records with 49 being TigerFish1 and 41 TigerFish0. 82 of 90 test records were correctly classified for an **accuracy** of 0.91. **Precision** (tp/tp+fp) was 0.94 i.e., every time the algorithm predicted that a fish was TigerFish1 it was correct. **Recall** (tp/tp+fn) was 0.90 i.e., percentage of TigerFish1 that are correctly classified. The overall **F1** score was 0.92. Table 1 below shows a comparison of the logistic regression classifier results against that of KNN classifier.

	KNN Classifier	Logistic Regression Classifier
Accuracy	0.93	0.91
Precision	0.94	0.94
Recall	0.94	0.90
F1 Score	0.94	0.92

Table 1. Comparison of major performance metrics

The KNN classifier seems to outperform the Logistic Regression classifier based on all the major performance metrics (accuracy, precision, recall and F1 score) as shown in the table above.