Brandon Langley
CPSC 4820
Logistic Regression
Project 3: Report
October 17, 2019

Problem Description

Tigerfish0 and TigerFish1 are two species of fish who appear very similar but have slight differences in body and dorsal fin length. Given a data set representing the body and dorsal fin measurements of both TigerFish0 and TigerFish1 develop and validate a Logistic Regression algorithm to classify a given pair of unknown measurements as either TigerFish0 or TigerFish1.

Data Description

Measurements were taken from a population of both TigerFish0 and TigerFish1, the body length and Dorsal Fin Length were recorded for 150 individuals from each species respectively. This data was given in a file where each line consisted of the record from a single fish with three-tab separated columns corresponding to the Body Length, the Dorsal fin Length and then an integer representing the species, a 0 for TigerFish0 or a 1 for TigerFish1. This Initial data set can be seen represented graphically in Figure 1.

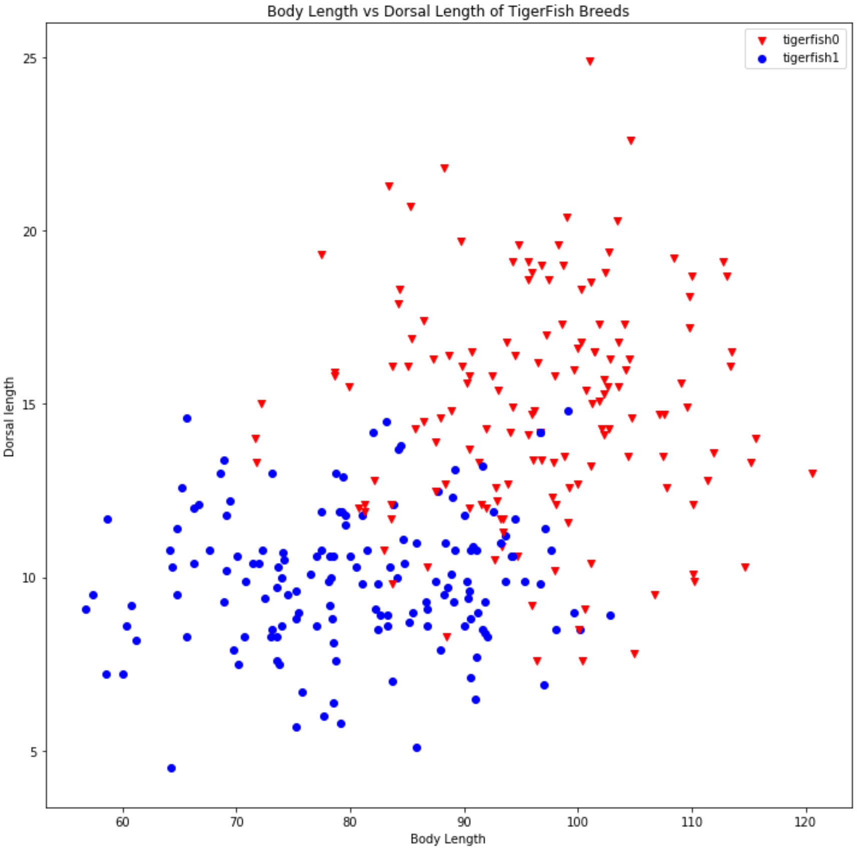

Figure 1. Initial Data Set

Logistic Algorithm Training Development

When the initial data set was read into the program it was grouped by species such that all records for each respective species were adjacent, to improve how the algorithm generalizes from the initial set the data was immediately randomized. Next the data set was split into a training set containing 70% of the data (210 records) and a test set containing the remaining 30% (90 records). Then the logistic regression model was trained using the training set. The initial values passed to the model were as follows; weights: W_0=0, W_1=0, W_2=0 and learning rate alpha=0.0001. However, when attempting to train the model the cost seemed to behave unpredictably and was not converging so the learning rate was adjusted to find a more appropriate value, eventual alpha=0.000001 was chosen. With this value for alpha the initial value for J was 0.69314718 from this point the algorithm was further trained. In this instance the algorithm was iterated 70,000 times. After training, the cost for the weights after each iteration was plotted and can been seen below in Figure 2. As can be seen in the graph the cost converges around .60.

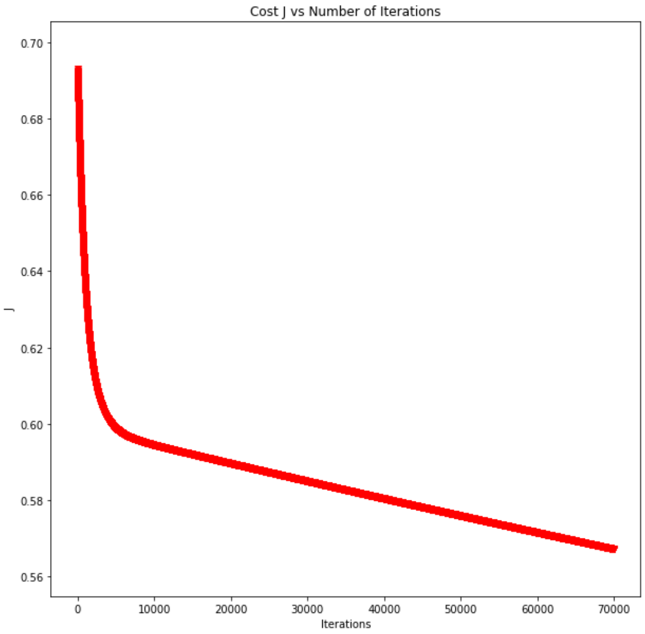

Figure 2. Cost J vs number of Iteration

Results

After completing training, the algorithm had the following weights W_0=1.27705108, W_1=0.02545808, W_2=-0.2986095. The Final value for J on the training set was 0.58951353 the value on the test set was J=0.5203803. The Test Set consisted of 90 records 46 of which represented Tigerfish1 and 44 which were Tigerfish0. As can be seen in the confusion matrix 67 of the 90 records were correctly identified for an accuracy of 0.7444. Precision was .8286, 6 fish were identified as Tigerfish1 which were actually Tigerfish0. Recall was .6304, 17 TigerFish1 were incorrectly identified as TigerFish0. Taking all these values into consideration the overall F1 score was .7160. Compared to my previous results with kNN, which were accuracy of .883, Precision of .909 and an F1 score of .895, the Logistic regression seems to be slightly worse at predicting the type of tigerfish based on its body measurements.

		Predicted TigerFish1	
		Y	N
Actual TigerFish1	Y	TP=29	FN=17
	N	FP=6	TN=38

Figure 3. Confusion Matrix

		Predicted TigerFish0	
		Y	N
Actual TigerFish0	Y	TP=30	FN=4
	N	FP=3	TN=23

Figure 4. Confusion Matrix for kNN