CPSC 6820 Project 3 Report : Logistic Regression

Problem Description: Given a data set representing body and dorsal fin length of TigerFish0 and TigerFish1 fish species, develop a logistic regression algorithm that will determine if a fish is TigerFish1 (the positive case) or a TigerFish0.

Data Description:

The initial training data consisted of 300 records representing features of either TigerFish1 species or TigerFish0 species.
Each record had 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). A plot of the data is shown in Figure 1.

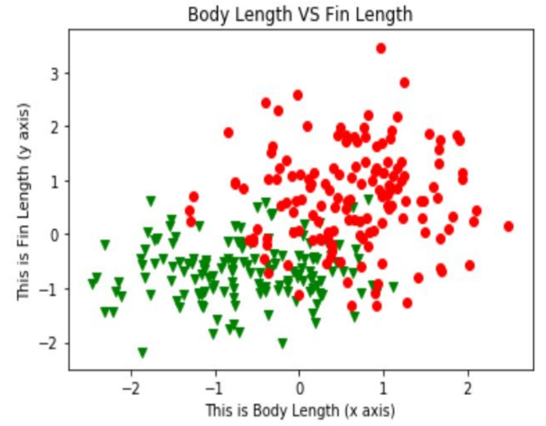


Fig. 1 The initial data set (red: TigerFish0)

Model Description and procedure:

Data of 300 records was divided into a training set (80%) and test set (20%). Logistic regression hypothesis function was used with initial values of weights and alpha on randomized training dataset and continued until convergence. Used the final values on test set and calculated the TP,TN,FP,FN values. The accuracy obtained with several runs was more than 90% in all cases.

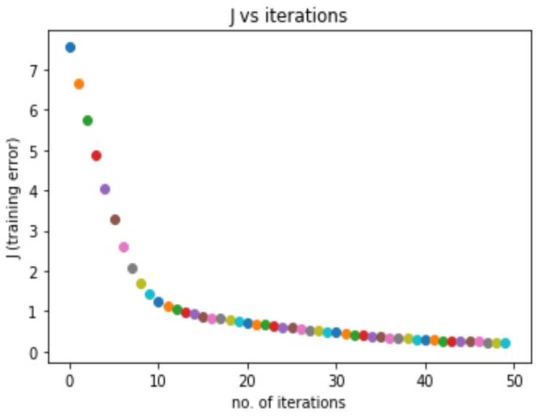
Initial Values:

Initial Values for weights, alpha were chosen as follows:
w0=1, w1=5, w2=7, alpha=0.003
Value of J with these initial values: **7.574**

Final Values:

After 50 iterations,when gradient descent started to converge, final values for weights, alpha were obtained as follows:
w0: 1.988, w1: -1.530, w2: -0.925, alpha=0.003
Final value of J on training set: **0.226**

Plot of J vs number of iterations for gradient descent:



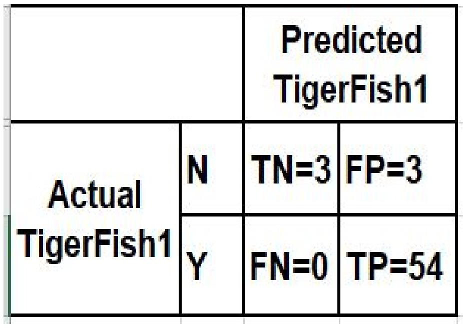
After 50 iterations, the gradient descent converges and J values start to decrease very slowly, almost constant, hence 50 was chosen to get final values of weights and alpha.

Fig 2. J vs iterations

Feature scaling:
Since the two features had a different range of values, I used standardization method of feature scaling and scaled all values to a single range. (x1 and x2)
This way, the values were in the same range for training and higher value did not dominate.

Value of J on test set:
After using final weight values on the test set, the J on test set was observed : **0.157**

Results:
A Confusion matrix for the results of the logistic regression algorithm is shown in Fig 3.
The test set consisted of 3 records representing TigerFish0, and 57 records representing TigerFish1 species. 57 of 60 records were correctly identified for an **accuracy of 0.95** .
Of the 60 records:



		Predicted TigerFish1	
Actual TigerFish1	N	TN=3	FP=3
	Y	FN=0	TP=54

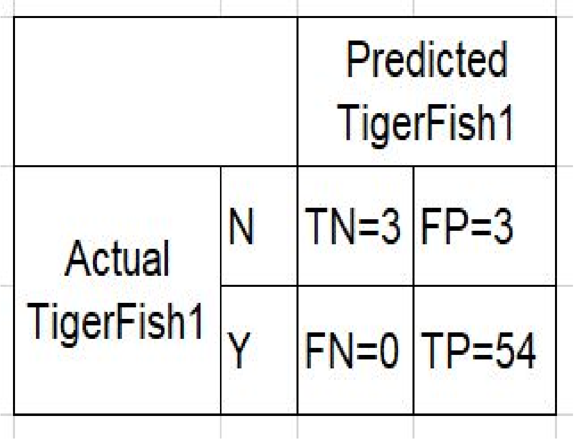
- 3 TigerFish0 species were correctly identified
- 54 TigerFish1 species were correctly identified
- 3 TigerFish0 species were misclassified as TigerFish1

Precision : 0.95.
Recall : 1
all positive cases (TigerFish1) were correctly predicted.
The over F1 score was 0.98.

Fig 3. Confusion matrix

A comparison of results with a logistic regression approach as compared to previous k-nearest neighbor approach (including kNN confusion matrix, accuracy, precision, recall and F1 values) :

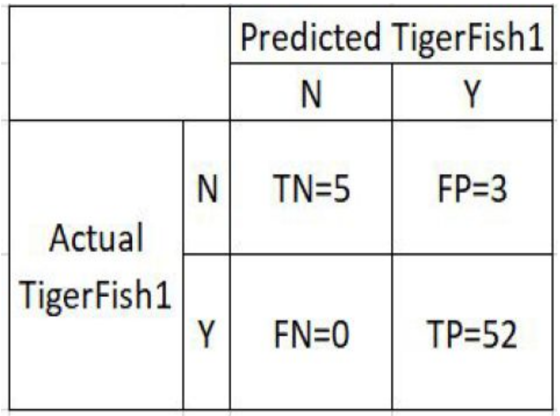
Logistic Regression



		Predicted TigerFish1	
Actual TigerFish1	N	TN=3	FP=3
	Y	FN=0	TP=54

Accuracy: 0.95
Precision: 0.95
Recall: 1
F1: 0.98

K Nearest Neighbors



		Predicted TigerFish1	
		N	Y
Actual TigerFish1	N	TN=5	FP=3
	Y	FN=0	TP=52

Accuracy: 0.95
Precision: 0.94
Recall: 1
F1: 0.97