Logistic Regression

Project3 Report

By: Shivam Pandit

Problem Description

Given a data set representing given the body length and dorsal fin length of a fish, we have to create a Logistic Regression program that will predict if it is TigerFish1 or TigerFish0.

Data Description

The initial training data consisted of 300 records representing features of either TigerFish0 or TigerFish1. Each record had three tab-separated entries. The first line 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).

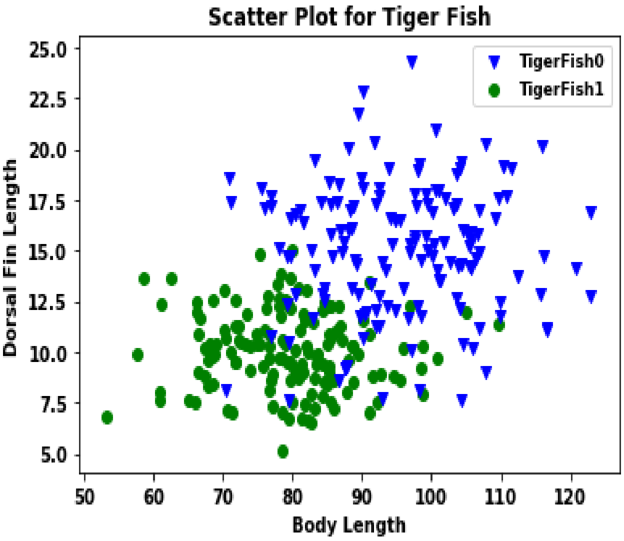


Figure 1: Initial Data Set

A plot of the data is shown in Figure 1.

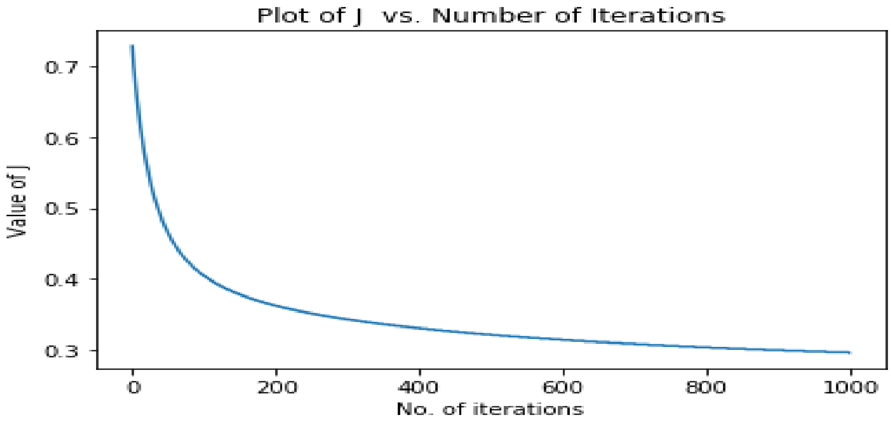
Initial Values for weights, alpha and J

Initially, I took random values for weights using random.seed(4) which gave weights as [0.44, 0.03, 0.55] and I chose alpha as 0.01. Initial cost(J) came out to be 0.758

Final values for alpha, weights, number of iterations and final value of J on training set

After splitting dataset into 70% training and 30% test we had 210 examples in training set and 90 examples in test set. After training model with different values of alpha, I chose alpha as 0.01, number of iterations as 1000 as that gave me reasonable plots. My final values for weights after 1000 iterations were [2.06, -5.67, -17.36]. Final value of cost(J) on training data set came as 0.296

Plot of *J* (vertical axis) vs. number of iterations (horizontal axis)



Feature Scaling

Since the features were not normally distributed, so I used standardization to scale down features using formula shown below. For all the fields in the dataset, mean was subtracted from them and then divided by standard deviation of the dataset. Standardizing the features so that they are centered around 0 with a standard deviation of 1 is important so that smaller features don't get neglected.

$$z = \frac{x - \mu}{\sigma}$$

Value of J in Test Dataset

Test Dataset had 90 examples or rows along with 90 true labels. Value of J on test set came out to be 0.256 using final weights.

Confusion Matrix for Logistic Regression on Test Set

		Predicted Tiger Fish	
		N	Y
Actual TigerFish	N	TN=38	FP=6
	Y	FN=2	TP=44

Confusion Matrix for Logistic Regression

Final Results

We got 82 out of 90 predictions correct which gave accuracy as 0.91. We got precision as 0.88 which means 88% of time, we predicted a fish as tigerfish1 to be correct. We got recall as 0.96 which means 2 records were misclassified. We got F1_Score as 0.92

kNN & Logistic Regression Results Comparison

Predicted Tiger Fish

		N	Y
Actual TigerFish	N	TN=38	FP=6
	Y	FN=2	TP=44

Confusion Matrix for Logistic Regression

Predicted TigerFish

		N	Y
Actual TigerFish	N	TN=29	FP=3
	Y	FN=3	TP=25

Confusion Matrix for KNN

Results	Logistic Regression	kNN
Accuracy	0.91	0.90
Precision	0.88	0.89
Recall	0.96	0.89
F1 Score	0.92	0.89

From the table above, we can see that there is small difference in performance parameters of Logistic Regression and kNN. Logistic Regression has better accuracy, recall and F1Score which shows that Logistic Regression performed slightly better as compared to kNN.