Brandon Langley
CPSC 4820
k-Nearest Neighbor
Project 1: Report
September 16,2019

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 k-Nearest Neighbor(kNN) 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.

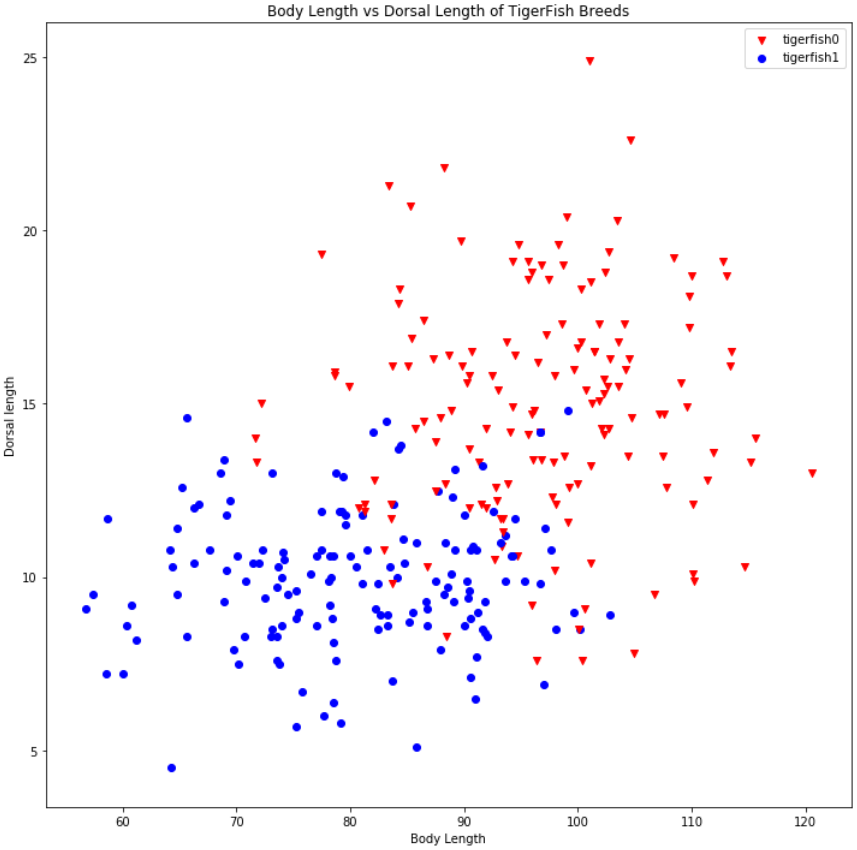

Figure 1. Initial Data Set

kNN 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 240 records and a test set containing the remaining 60 records. Next to further improve the algorithm, 5-fold Cross validation was used, meaning the 240 records in the training set were further broken into 5 folds containing 48 records each. These 5 folds were used to create 5 smaller training/validation pairs. Each pair containing a different combination of 4 of these folds to be the training set for a total of 192 record with the final 48 records of the 5th fold used as a validation set. After these Training/Validation combinations were created each pair was processed with kNN with each k value from 1 to 31. The errors for each test were recorded. After running the algorithm 40 times I found that consistently the best value for k was 3, it provided the lowest error for 13 out of the 40 test runs which was far more than any other single k value. Using the data gathered during the test runs the accuracy was plotted for each value of k which can be seen in Figure 2. k=3 was chosen for kNN for the test set.

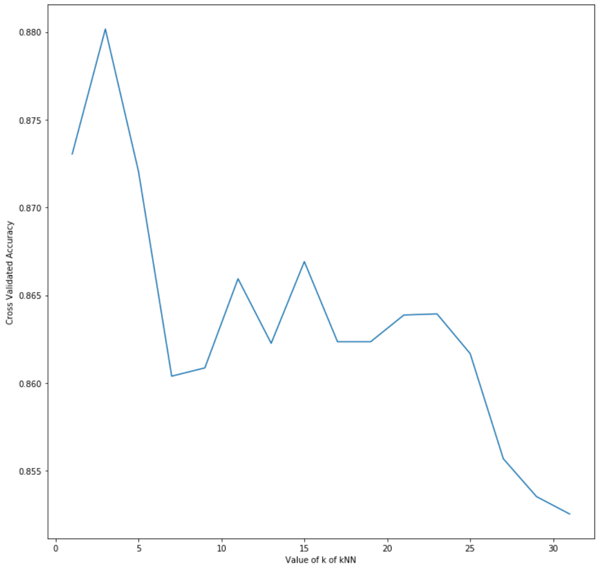

Figure 2. Average Accuracy for different Values of k

Results

Finally, after completing the cross fold validation kNN was run on the test set with k=3. The results of this test can be seen in the confusion matrix in Figure 3. The test set consisted of 60 values 34 of which were TigerFish0 and 26 which were TigerFish1. As can be seen in the confusion matrix 53 of 60 records were correctly identified for an accuracy of .883. Precision was .909, 3 fish were identified as TigerFish0 which were actually TigerFish1. Recall was .882, four TigerFish0 were incorrectly identified as TigerFish1. Taking all these values into consideration the overall F1 score was .895

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

Figure 3. Confusion Matrix