kNN Project1 Report

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 k-Nearest Neighbor 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).

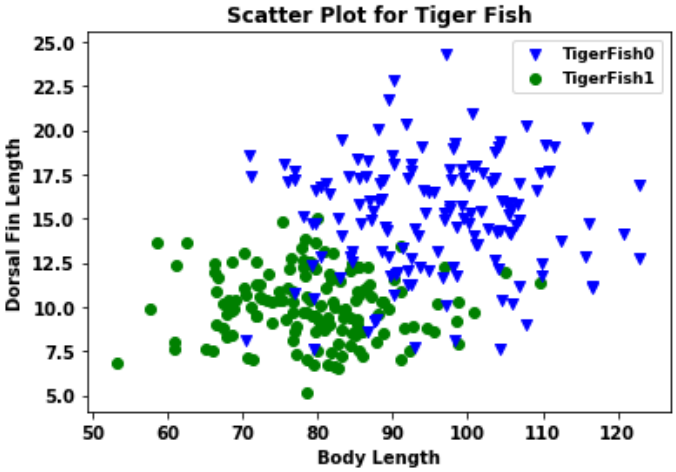

Figure 1: Initial Data Set

A plot of the data is shown in Figure 1.

Training a kNN Algorithm

A k Nearest Neighbor algorithm was developed using 5-fold Cross Validation. First, the data was read from .txt file and stored in nd array ignoring first row that had total rows mentioned. Then, that nd array was randomly shuffled and split into train and test with 80:20 ratio. That resulted into training set with 240 records (80%) and test set with 60 records (20%). The training set of 240 records was further divided into five folds of 48 records each. The five folds were used to create five smaller training sets of four folds (192 records) each, with the

K	1	3	5	7	9	11	13	15	17	19	21
Test 1 Errors	8	7	5	7	6	5	5	4	4	4	4
Test 2 Errors	8	5	4	5	4	5	5	5	5	5	6
Test 3 Errors	9	8	8	9	8	8	8	8	8	8	8
Test 4 Errors	5	7	7	7	7	6	7	8	7	7	7
Test 5 Errors	5	5	4	3	3	3	5	5	5	5	5
Totals	35	32	28	31	28	27	30	30	29	29	30

Figure 2. Misclassifications for different values of k on the five training sets

leftover fold (48 records) in each case used as the validation set. Each training set was then executed via k-NN with odd values of k of 1 through 21. For each value of k, the number of misclassifications were recorded for all five training/validation set combinations (Figure 2). From this data the cross-validated accuracy was plotted for each value of k. k = 9 and k = 11 provided the best accuracy (Figure 3). k=11 was chosen for kNN for the test set.

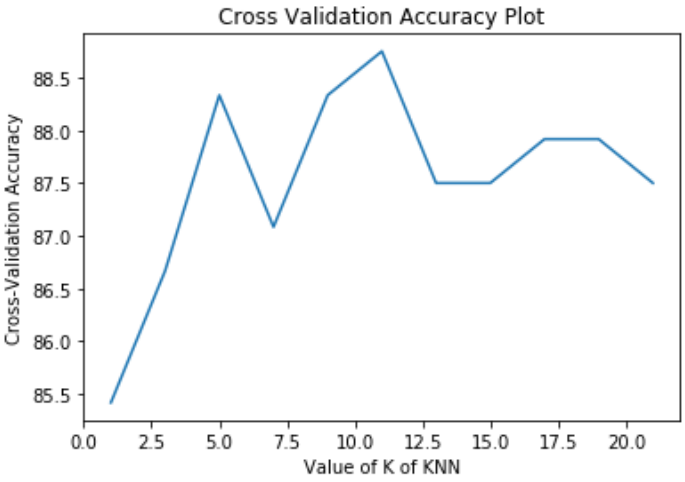

Figure 3. Average accuracy for different values of k

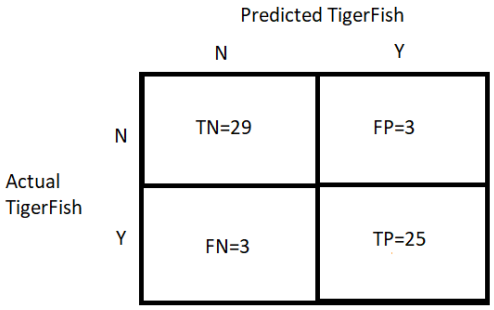

Figure4: Confusion Matrix

Results

A Confusion matrix for the results of the Nearest Neighbor algorithm with k = 11 is shown in Figure 4. The test set consisted of 60 records, 34 records represented TigerFish 0, and 26 records represented TigerFish 1. 54 out of 60 records were correctly identified for an accuracy of 0.90. Precision was equal to 0.89, meaning each time we predicted a fish as TigerFish 1, 89% of the time it was correct. Recall was 0.89; 3 TigerFish 0 records were misidentified as TigerFish 1. The overall F1 score was 0.89.