CPSC 6820 Project 1 Report : k Nearest Neighbor

Problem Description: Given a data set representing body and dorsal fin length of TigerFish0 and TigerFish1 fish species, develop a kNN classification 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.

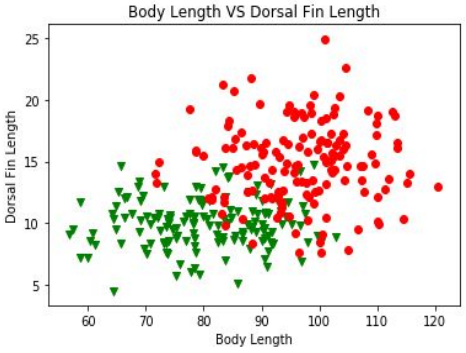

Fig. 1 The initial data set (red: TigerFish0)

Training a kNN Algorithm:

A k Nearest Neighbor algorithm was developed using 5-fold Cross Validation. First, the data was randomized and 60 randomly selected records were put into a test set (20 percent). The remaining 240 records were 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 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=7 and k=9 provided the best accuracy (Figure 3). **k = 7** was chosen for kNN for the test set.

k	**1**	**3**	**5**	**7**	**9**	**11**	**13**	**15**	**17**	**19**	**21**
Test1 Errors	4	3	3	2	1	5	4	3	4	5	3
Test1 Errors	3	5	4	3	3	4	3	6	2	5	3
Test1 Errors	3	5	4	3	4	5	4	5	5	5	6
Test1 Errors	5	4	4	4	3	4	5	6	6	2	5
Test1 Errors	4	4	4	3	4	4	5	3	4	4	6
Totals	**19**	**21**	**19**	**15**	**15**	**22**	**21**	**23**	**21**	**21**	**23**

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

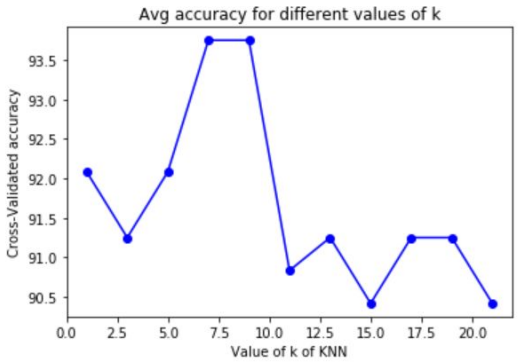

Fig. 3 Avg accuracy for different values of k

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

Fig. 4 Confusion Matrix

Results:

A Confusion matrix for the results of the Nearest Neighbor algorithm with **k = 7** is shown in Figure 4.
The test set consisted of 8 records representing TigerFish0, and 52 records representing TigerFish1 species. **57 of 60 records were correctly identified for an accuracy of 0.95**.

Of the 60 records:

- 5 TigerFish0 species were correctly identified
- 52 TigerFish1 species were correctly identified
- 3 TigerFish0 species were misclassified as TigerFish1

Precision was equal to 0.94.
Recall was 1; all positive cases (TigerFish1) were correctly predicted.
The over F1 score was 0.97.