kNN Class Example Report

Larry Hodges

Problem Description

Given a data set representing petal and sepal length of Versicolor and Virginica Iris flowers, develop a kNN classification algorithm that will determine if a flower is a Versicolor (the positive case) or a Virginica.

Data Description

The initial training data consisted of 100 records representing features of either Versicolor Iris flowers or Virginia Iris flowers. Each record had three tab-separated entries. The first is a float representing the sepal length in centimeters, followed by a float represent the petal length in centimeters, then a string identifying the flower as either “versicolor” or “virginica”. A plot of the data is shown in Figure 1.

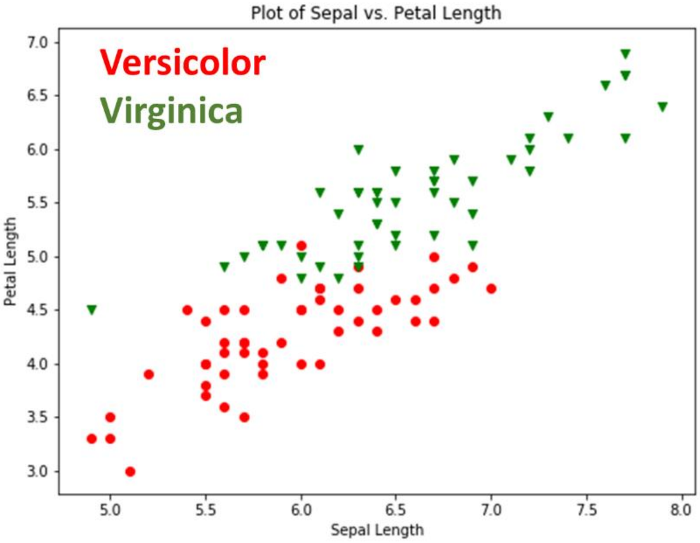
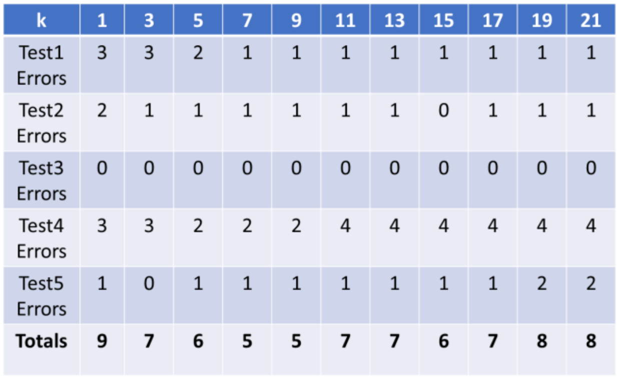


Figure 1. The Initial Data Set

Training a kNN Algorithm

A k Nearest Neighbor algorithm was developed using 5-fold Cross Validation. First, the data was randomized and the string, ”versicolor,” was replaced by a 1, and the string, “virginica,” was replaced by a 0. Twenty randomly selected records were put into a test set. The remaining 80 records were divided into five folds of 16 records each. The five folds were used to create five smaller training sets of four folds (64 records) each, with the leftover fold (16 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



k	1	3	5	7	9	11	13	15	17	19	21
Test1 Errors	3	3	2	1	1	1	1	1	1	1	1
Test2 Errors	2	1	1	1	1	1	1	0	1	1	1
Test3 Errors	0	0	0	0	0	0	0	0	0	0	0
Test4 Errors	3	3	2	2	2	4	4	4	4	4	4
Test5 Errors	1	0	1	1	1	1	1	1	1	2	2
Totals	**9**	**7**	**6**	**5**	**5**	**7**	**7**	**6**	**7**	**8**	**8**

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

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.

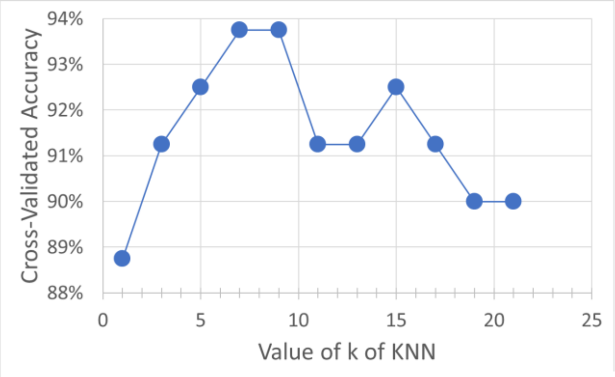


Figure 3. Average accuracy for different values of k

		Predicted Versicolor	
		N	Y
Actual Versicolor	N	TN=6	FP=0
	Y	FN=2	TP=12

Figure 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 6 records representing Virginica flowers, and 14 records representing Versicolor flowers. 18 of 20 records were correctly identified for an accuracy of 0.9. Precision was equal to 1.0. I.e., every time the algorithm said a flower was Versicolor it was correct. Recall was 0.86; two Versicolor records were misidentified as Virginica. The over ***F1*** score was 0.92.