K-NEAREST NEIGHBOR PROJECT 2 REPORT

ERIC CAI

Problem Description

Develop a k-NN program to predict whether capacitors from a fabrication plant pass quality control based (QC) on two different tests. Train your system and determine its reliability with a set of 118 examples.

Data Description

The initial data consisted of 118 examples of capacitors (Figure 1). Each record had three values. The first value was a float that represented the capacitor's score on the first test. The second value was a float that represented the capacitor's score on the second test. The third value was an int of either 0 or 1 that represents whether the capacitor failed or passed quality control, respectively. These 118 examples were split into a training set of 85 examples and a test set of 33 examples.

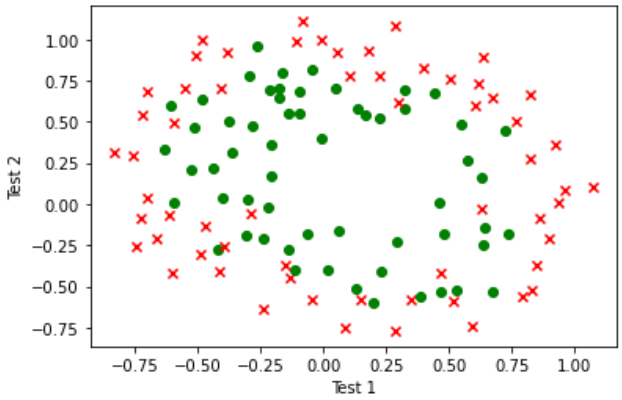

Figure 1: The initial data set.

Procedure

A k-Nearest Neighbor algorithm was developed using 5-Fold Validation. First, the data was randomized into 5 folds of 17 examples each. These folds were used to create four different training sets of 68 examples with the leftover 17 examples used as the validation set. Each example in the validation sets was categorized based on k-NN with values of 1 through 21. The number of misclassifications were recorded for each fold and each k value (Figure 2). From this data, the cross-validated accuracy was plotted for each value of k (Figure 3). The k value with the highest value was 3, so k = 3 was chosen for the test set.

k	1	3	5	7	9	11	13	15	17	19	21
Test 1 Errors	8	4	6	5	11	10	12	12	13	12	12
Test 2 Errors	8	3	5	6	7	7	6	7	9	9	8
Test 3 Errors	7	6	5	8	8	9	9	9	9	8	7
Test 4 Errors	7	7	8	6	6	7	6	5	5	5	6
Test 5 Errors	7	4	3	2	1	3	4	4	8	7	7
Total	37	24	27	27	33	36	37	37	44	41	40

Figure 2: The number of misclassifications for each k and fold.

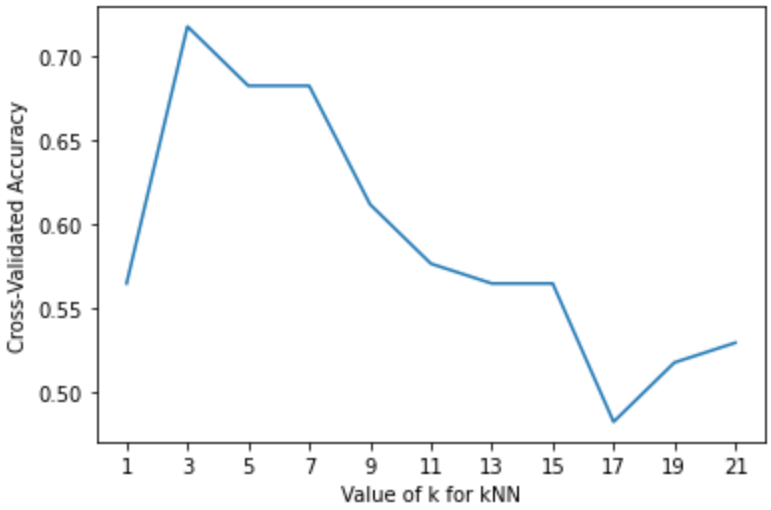

Figure 3: The plot of the cross-validated accuracy for each k.

Results

A confusion matrix was created for the results on the test set of a k-Nearest Neighbor algorithm with k = 3 (Figure 4).

The test set consisted of 16 capacitors that did not pass quality control and 17 capacitors that did pass quality control for a total of 33 capacitors. 21/33 capacitors were correctly classified for an accuracy of about 0.6364. 11/17 capacitors predicted to pass actually passed for a precision of about 0.6471. 11/17 capacitors that actually passed were predicted to pass for a recall of about 0.6471. The overall F1 score was about 0.6471.

		Predicted Pass	
		N	Y
Actual Pass	N	TN: 10	FP: 6
	Y	FN: 6	TP: 11

Figure 4: A confusion matrix for the results on the test set of a k-Nearest Neighbor algorithm with k = 3.