Q29

Rahul Atre

2023-11-09

Q29 [Classification]

Suppose that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use logistic regression and get an error rate of 20% on the training data and 30% on the test data. Next we use 1NN (i.e. K = 1) and get an average error rate (averaged over both test and training data sets) of 18%. Based on these results, which method should we prefer to use for classification of new observations? Why?

Ans: In the case of logistic regression, we are given both the training error and test error rate. For the KNN classification method, we are only provided with the average error rate, which is 18%. However, it is mentioned that the specific classification used is 1NN, where K = 1. In 1NN, we know that the single nearest neighbor from the dataset is selected. So, for the training set, K=1 will always select itself as the closest neighbour, thereby making the training error rate 0%.

Since the training error rate for K=1 of 1NN is 0%, the test error rate can be calculated as:

$\frac{P(training)+P(test)}{2} = 18\%$

$\Rightarrow 0\% + P(test) = 36\%$

$\Rightarrow P(test) = 36\%$

Thus, the test error rate of 1NN is 36%. In contrast, the test error rate for logistic regression is 30%, which is 6% less than 1NN.

Therefore, the preferred method for classifying new observations is **logistic regression**, since it has a lower test error rate.