Assignment29

John Bute

2024-11-08

R Markdown

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

Our main goal is to reduce the test error rate, as it reflects how a model generalizes to unseen data. Logistic regression has a test error rate of 30%, while the 1NN has an average rate across both training and test sets of 18%.

Thus we cannot simply compare both model's test error rates as we do not know it for 1NN.

However, we do know that 1NN is highly flexible, and prone to extreme overfitting, thus the training rate is typically 0%. Thus if the average between the training and test set error is 18%, then our test error would be approximately:

18 * 2 - 0 = 36% test error, which is higher than our logistic regression model (as an approximation, we cannot truly confirm the test error of 1NN)

Thus, we would prefer logistic regression, because despite it having maybe a higher test error, its nature allows it to be more robust and less likely to overfit, thus better at generalizing to new observations, while 1NN may definitely perform well on the training set, but it is a high variance model, that is prone to overfitting, and thus cannot generalize as well.