Gina Yang

QBIO 499

10 April 2019

Homework 10

I used the iris data set ( data(“iris”) ) for classification. This data set contains 50 instances each of 3 classes with each class representing a species of iris. Each sample has four features: sepal length, sepal width, petal length, and petal width.

Optimal neighborhood size k for k nearest neighbors: 7 (found using cross-validation, knn3() with neighbor values ranging from 1 to 149. 7 showed the lowest mean error of 0.0333)

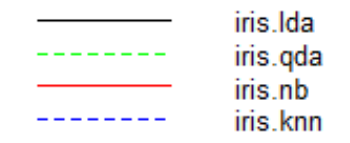
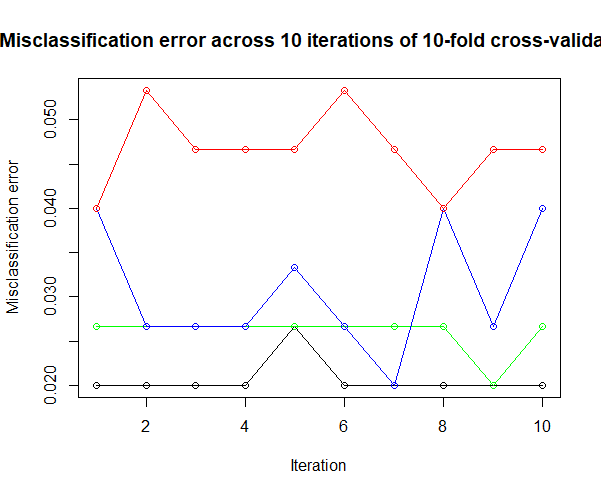


Table 1: Mean misclassification error of 10 iterations of 10-fold cross-validation for each classification method

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | lda() | qda() | knn3() | naive\_bayes() |
| Mean | 0.02 | 0.024667 | 0.030667 | 0.047333 |

Linear discriminant analysis showed the lowest mean misclassification error. Quadratic discriminant analysis performed slightly worse than lda, followed closely by k-nearest neighbors (carried out with 7 neighbors). Naïve Bayes had the highest mean error by far. This is likely because it assumes all features are independent of each other—an assumption appears not to work as well in practice for the iris flower.