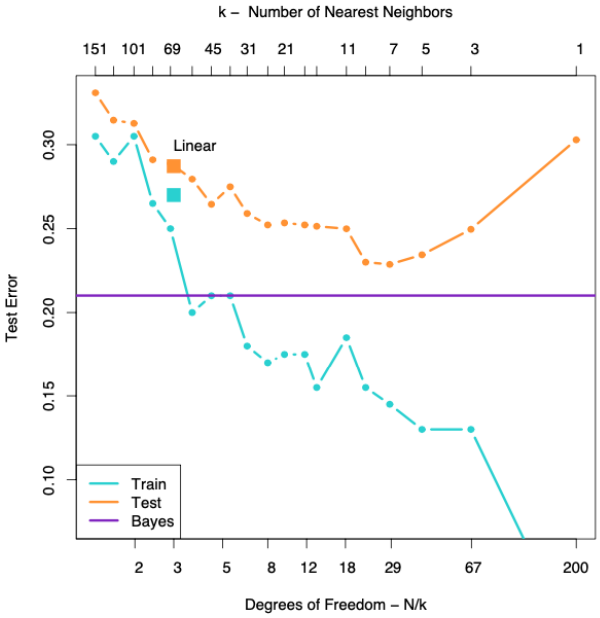
DATA 221
Homework 4 (rev 0)

Trimble/Nussbaum

Due: Friday 2023-02-03

In the last homework, you generated two lumpy distributions that were mixtures of multivariate normal samples in two dimensions. Using this distribution, we can reproduce the graph of accuracy vs. model complexity of kNN classification :



1. Perform k-nearest-neighbor classification for at least six values of k ranging from 1 to 100; use the neighbors of each point to predict the class identity. Evaluate accuracy as a function of k for the training set (with 200 points).

2. Generate a large ”testing” sample of 10,000 points from each class. (We will use the ”true” generated class values of this dataset to judge the accuracy of the KNN classifier.) Evaluate the accuracy of KNN classification trained on the 200-point training set as a function of k for the 20,000 points in the test set. Plot the accuracy vs. N/k for the training and the testing data on the same graph.

3. Become familiar with the MNIST digits dataset; 60,000 28 pixel by 28 pixel 8-bit greyscale images for training and 10,000 images for testing: Display 40 or so of the digits, 4 examples of each arabic digit 0 through 9.

4. Put the class labels (”true” digit labels) into one-hot encoding and run linear regression on the 60000 x 10 digit label matrix against the 784 x 60000 matrix of pixel values. Report the crude accuracy and the confusion matrix of unregularized linear regression.