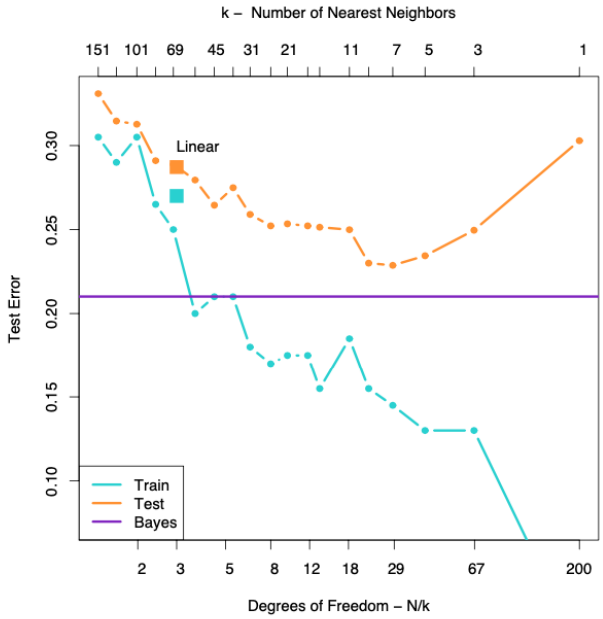
DATA 221
Homework 4 (rev 0)

Trimble/Nussbaum

Due: Friday 2023-02-03

In the last homework, you generated two lumpy distributionis that were mixtures of multivariate normal samples in 2d. 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. (This large set of points for evaluation causes the $\frac{1}{\sqrt{N}}$ sampling-based uncertainty in the correctly-classified proportion to be small, $\sim 1\%$.) 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 and plot the accuracy vs. k for the training and the testing data on the same graph.