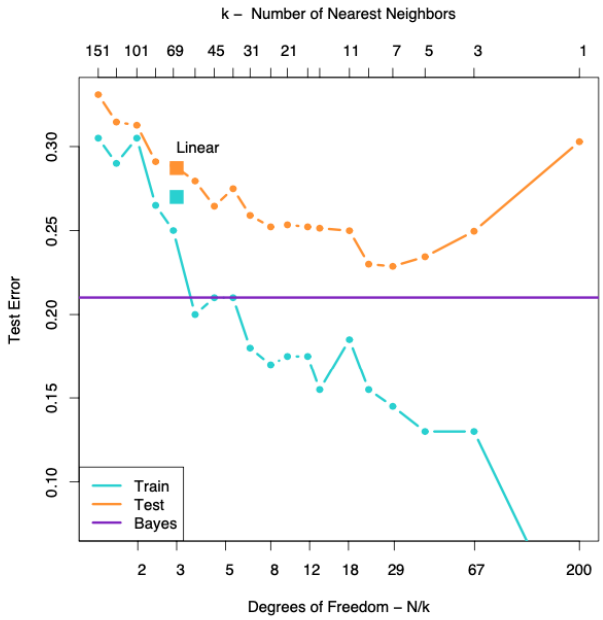
DATA 221
Homework 4

Trimble

Due: Friday 2023-04-21 11:59pm

In the last homework, you generated two lumpy distributions 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. (a) 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 knn with the 200-point the training set. You can generate a larger test dataset to make your plot more beautiful.

   (b) Estimate the Bayes error rate by drawing samples (as in HW3) and add a line showing this error rate to the plot of accuracy (or error) vs. log k.

2. Become familiar with the fashion-MNIST dataset; 60,000 28 pixel by 28 pixel 8-bit greyscale images for training and 10,000 images for testing in 10 classes.

   (a) Put the class labels into one-hot encoding and run logistic regression on the 60000 x 10 label matrix against the 784 x 60000 matrix of pixel values. Report the crude accuracy (on the holdout test set) and the confusion matrix of unregularized logistic regression. (This is essentially a zero-layer neural network.)

   (b) Compare the accuracy of unregularized logistic regression on the test data to that of regularized logistic regression, using cross-validation to choose the best regularization constant.

   (c) Train a 2-layer neural network on the fashion-MNIST data. Report the confusion matrix. Do you have any thoughts on the hard-to-discern pairs of labels?