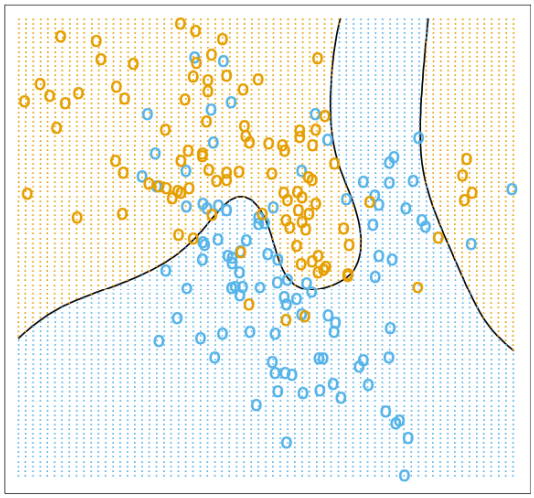
DATA 221
Homework 4 (rev 0)

Trimble/Nussbaum

Due: Friday 2023-02-02

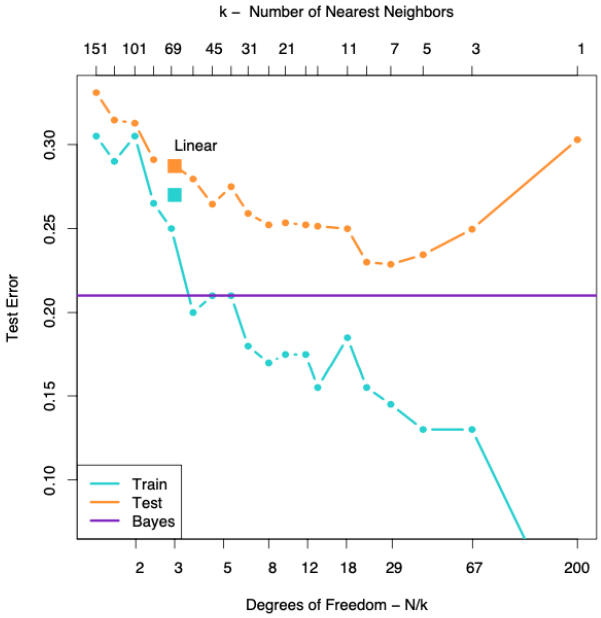
Using the distribution of orange and blue data points from a mixture-of-10-normals model from HW3:



1.

2. 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).

3. Generate a large "testing" sample of 10,000 points from each class. 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.



Looking at the UCI "default of credit card clients Data Set" contains various fields describing 30,000 credit card customers in Taiwan in 2005. (Yeh & Lien, doi://10.1016/j.eswa.2007.12.020)

4. Try to predict the `default.payment.next.month` using KNN classifiers for four different values of k.

   You have to choose how to measure distance in a vector space that includes indicator variables and payment amounts ($NT 100k).

   Report accuracy on the testing and training datasets.

   `https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients`

5. Compare the KNN accuracy to the accuracy of logistic x