DATA 221
Homework 3 (rev 1)

Trimble/Nussbaum

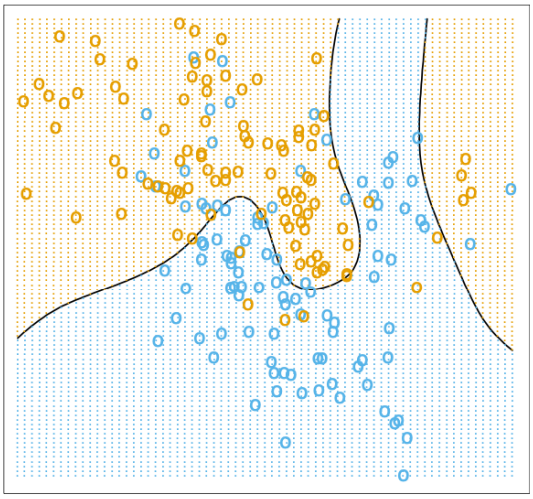
Due: Friday 2023-01-26 11:59pm

This question asks to you produce a graph like of overfitting given in Hastie Elements of Statistical Learning Figure 2.4. The underlying source of the points in the graph was a lumpy mixture of normal distributions. We will have to generate random parameters for this distribution, generate samples from the distribution for training, and generate another (large) set of samples for testing. The random dataset you generate should look like a shotgun target.

To start, you need to make a distribution of 100 orange and 100 blue points. Start by generating 10 means for class 1 (orange) and 10 means for class 2 (blue) from a normal distribution with variance (1, 1) and centered at (x_1 ,x_2) = (0,1) for blue and (1,0) for orange and no correlation between x_1 and x_2. For the trainign data, generate 10 data points from a 2d normal with standard deviation 1/3 for each (of the 20) clusters. This is now a lumpy distribution in two dimensions with 10 clusters for class 1 and 10 clusters for class 2.

1. Generate 200 points from the lumpy-Gaussian-mixture dataset as described above. Plot a scatterplot.

2. Visualize the Bayes decision boundary between the two classes, the surfaces where the (true) density in class 1 equals the density in class 2. You can use contour maps to approximate the boundary or you can solve for the boundaries numerically.



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)

3. Split the dataset 50/50 into training and test, and try several logistic regression models to predict the `default.payment.next.month` field.

4. Present a summary of a handful of the models that gave the best accuracy on the test set. Which variables did you include, and which ones mattered the most?

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