DATA 221
Homework 3 (rev 1)

Trimble

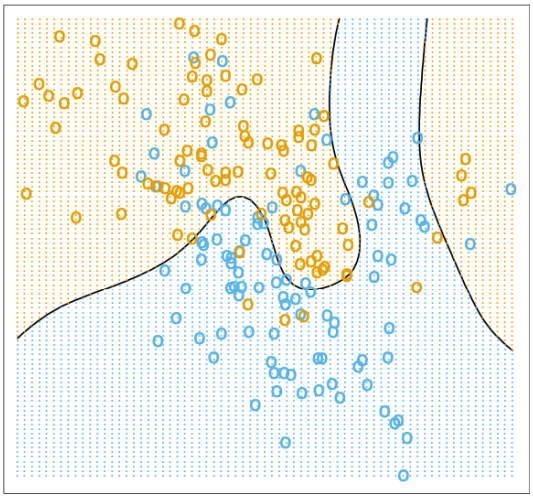
Due: Monday 2023-04-17 11:59pm

1. This question asks to you produce an artificial dataset like that used in Hastie Elements of Statistical Learning. The underlying source of the points in the graph was a mixture of normal distributions. We will have to generate random parameters for the underlying distributions, and generate samples from the mixture distribution for training and 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 training 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.

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

   (b) 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.



2. 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)

   (a) Split the dataset 50/50 into training and test. Use forward/backward model selection (you will need to write code for this) to create two sequences of models of varying complexity. Present a table summarizing at least 12 of the models, the parameters each includes, and the goodness-of-fit for each.

(b) Create ROC curves (plots of True Positive Rate vs. False Positive Rate) for a handful of the models from your backward/forward selection process. 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? Note: to generate these at all you need to vary the false positive/false negative balance. This is best done by extracting probabilities from your lostistic regression and applying a varying threshold to the probability-of-default estimate.

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