Overfitting

This question asks to you produce a graph demonstrating overfitting like that given in Hastie Elements of Statistical Learning Figure 2.4 (pictured below). The underlying source of the points in the graph is a mixture of normal distributions. We will have to 1. Generate random parameters for this distribution, 2. Generate samples from the distribution for training, and 3. Generate another (large) set of samples for testing. The random data sets you generate should look like a shotgun target.

The method for generating the data sets we need is as follows:

- Start by generating 10 means for Class A (orange) and 10 means for Class B (blue) from a bivariate normal distribution. The parameter values for each class are given below:
  - Class 1: $\mu_A = (0, 1), \mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
  - Class 2: $\mu_B = (1, 0), \mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- For the training data, generate 10 data points from a 2-dimensional normal with standard deviation 1/3 for each (of the 20) clusters, a.k.a., $\mu_{Ak} = (x_{1k}, x_{2k}), \mathbf{\Sigma} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$ for data generated in Class A, cluster k and $\mu_{2k} = (\mu_{1k}, \mu_{2k}), \mathbf{\Sigma} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{bmatrix}$ for data generated in Class B.

This is now a distribution in two dimensions with 10 clusters for Class A and 10 clusters for Class B! Be ready to generate 10,000 more samples from these distributions; keep track of the cluster means since you will need them later. You may also want to set a seed to maintain reproducibility.

1. Generate 200 points (100 from Class A, 100 from Class B) from the normal-mixture data set as described above. Plot a scatterplot.
2. Visualize the Bayes decision boundary between the two classes, the surfaces where the (true) density in Class A equals the density in Class B. You can use contour maps to approximate the boundary or you can solve for the boundaries numerically.