SD 675 Pattern Recognition

You may want to read Duda, Hart, & Stork, pages 476-478.

You'll need the data assign4.mat from the course home page. There are two classes, A and B. We will call this initial data set $\mathcal{D}$.

MED

Find the MED classifier for the two clusters, using the sample mean as the prototype. How many of the sample points are misclassified?

Boosting

We'll take a slightly simplified approach:

1. Select, at random, one quarter of the points in both A and B. Call this set $\mathcal{D}_1$. Learn classifier $\mathcal{C}_1$ based on $\mathcal{D}_1$.
2. Find the points in $\mathcal{D}$ which are misclassified by $\mathcal{C}_1$. At random, keep half of these bad points and an equal number of points where $\mathcal{C}_1$ classifies correctly. Call this new set $\mathcal{D}_2$. Learn classifier $\mathcal{C}_2$ based on $\mathcal{D}_2$.
3. Finally, find all points in $\mathcal{D}$ which are classified differently by $\mathcal{C}_1$ and $\mathcal{C}_2$. Based on all of these points, $\mathcal{D}_3$, learn classifier $\mathcal{C}_3$.
4. The final classifier is found by keeping the majority vote of $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$.

How will we learn the classifiers? First try MED; it will probably not do very well — explain why not. Instead of MED, I propose the following:

- Given training sets $\mathcal{D}_A$ and $\mathcal{D}_B$, do the following $i = 1 \ldots q = 10$ times:
  - Pick one random point in each of $\mathcal{D}_A$ and $\mathcal{D}_B$.
  - Let trial classifier i be the minimum Euclidean distance to these points.
  - Approximate $P(\epsilon)$ based on the rest of the training data.
- Of the q trial classifiers, keep the one with the lowest $P(\epsilon)$.

Plot the resulting classification boundary. Does the boundary vary very much from one run to the next? Does the probability of classification error vary much between runs? Comment.

How sensitive is this approach to the choice of q? How large / small does q need to be for reasonable results?

How does $P(\epsilon)$ compare between unboosted MED, boosted MED, unboosted proposed classifier, and the boosted proposed classifier?