SYDE 675 Pattern Recognition

Assignment 3
Due Mar 15th 2017

You may want to read Duda, Hart, & Stork, pages 476-478.
You'll need the data file assign3.mat from the course home page on LEARN.
The assign3.mat file contains samples from two classes, A and B.
We will call this initial data set $\mathcal{D}$.

MED Classifier

Using $\mathcal{D}$ as a training set, find the two class MED classifier, using the class sample means as prototypes. How many of the training samples are erroneously classified using this MED classifier?

Boosting

We'll take a slightly simplified approach:

1. Select, at random, one quarter of the training samples from each of class A and B.
   Call this set $\mathcal{D}_1$. Train base classifier $\mathcal{C}_1$ using $\mathcal{D}_1$.
2. Find the samples in $\mathcal{D}$ which are erroneously classified by $\mathcal{C}_1$.
   Select , at random, half of these erroneous samples and an equal number of samples correctly classified by $\mathcal{C}_1$.
   Call this set $\mathcal{D}_2$. Train base classifier $\mathcal{C}_2$ using $\mathcal{D}_2$.
3. Finally, find all samples in $\mathcal{D}$ which are classified differently by $\mathcal{C}_1$ and $\mathcal{C}_2$.
   Call this set $\mathcal{D}_3$ Train base classifier $\mathcal{C}_3$ using $\mathcal{D}_3$.
4. The final boosted classifier is found by keeping the majority vote of $\mathcal{C}_1, \mathcal{C}_2, \mathcal{C}_3$.

Given training set $\mathcal{D}$, evaluate the above described boosting algorithm using each of the following two types of base classifier:

1. MED classifier, using the sample means of the provided class training samples as prototypes.
2. A selected-linear classifier as described below:
   First do the following q times:
   - At random, select one sample from each set of provided class training samples.
   - Define a corresponding MED classifier based on the distance to these selected samples.
   - Approximate $P(\epsilon)$ based on the rest of the training data.

   Then, of the q classifiers, select the one with the lowest $P(\epsilon)$.

The boosted MED classifier method will probably not work very well.
Explain why the defined boosting algorithm fails to significantly improve performance when a sample-means based MED classifier is used as a base classifier.

Next, let $q = 10$ and test boosting using a selected-linear classifier as a base classifier.
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 the probability of error for the boosted selected-linear classifier method to the choice of q?
How large/small does q need to be for reasonable results?

How does $P(\epsilon)$ compare across the unboosted MED, boosted MED, unboosted selected-linear, and boosted selected-linear (each using $q = 10$) classifier methods?