Assignment 3 SYDE 675

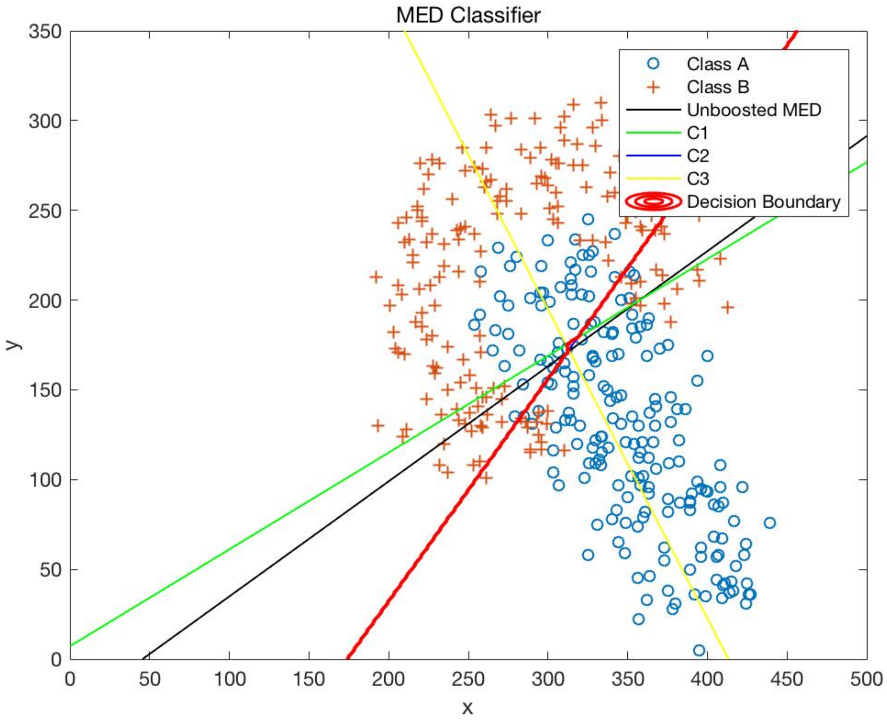
XIN CHEN 20705271

March 13

1 MED Classifier

Using 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?
Answer: using this MED classifier, there are 82 training samples are erroneously classified.

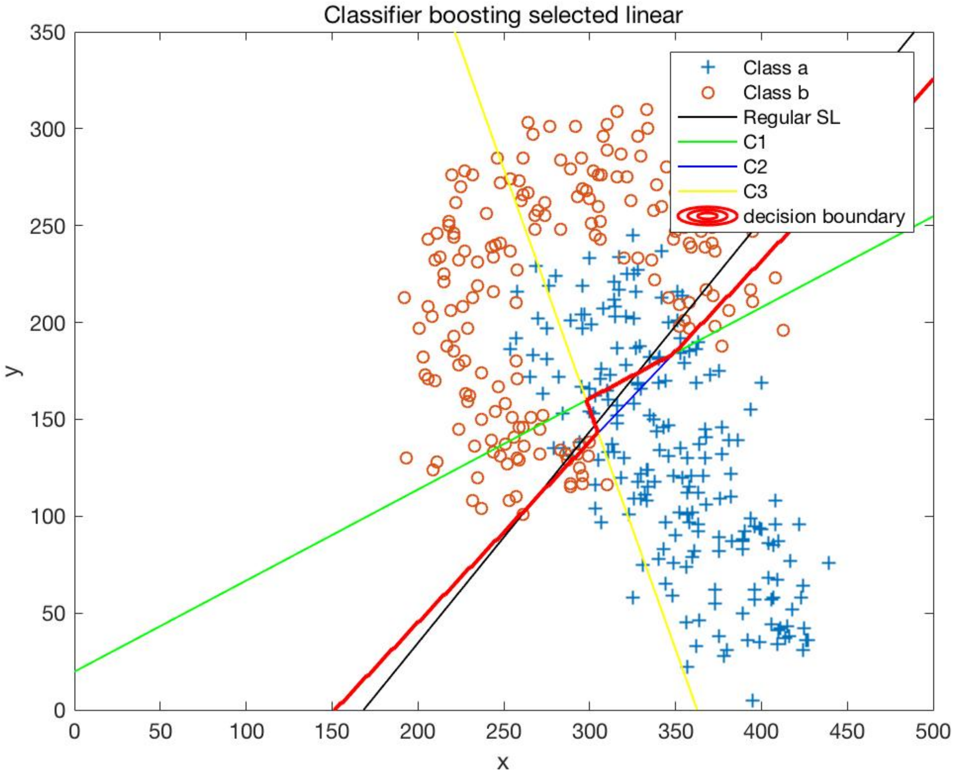
2 Boosting



Explain why the defined boosting algorithm fails to significantly improve performance when a sample-means based MED classifier is used as a base classifier.

Answer: The boosting algorithm didn't significantly improve performance of the MED classifier (from 0.2050 to 0.1950). The main reason is that the voting of c1 and c2 are almost similar, and this result in the final majority voting result was almost the same as c1 and c2. Moreover, this also make c3 sometimes become meaningless.

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?

Answer: Yes. From the figure, it is clear that the c1, c2, c3 boundaries varied a lot. Moreover, the probability of classification error varied much between runs as well. Since we only selected one sample to train the classifier and there must be chance, the boundary of c1, c2, c3 classifiers varied a lot.

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?

Answer: As for the sensitivity of the probability of error for the boosted selected-linear classifier method to q, training size is the main influence factor. If we choose a large q for a large training set, there will be a chance for us to look for a lowest error rate sample. While if we generate a small training set, it is waste of time to choose the large q, which is because that we will fine the same data sample again and again. Therefore, if q is larger than half of size of the dataset and smaller than the whole dataset, the result will be reasonable.

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

Answer:

classifier	unboosted MED	boosted MED	unboosted selected-linear	boosted selected-linear
P(ε)	0.2050	0.1950	0.2100	0.1875

As the above table showed, the mean-based classification does better than the selected- linear one. It is because there are input randomness, noise and outlier in selected-linear classifier, which will influence the result a lot.