TTT4275 Estimation, Detection and Classification
Problem Set Class1

The main topics for this problem set are investigation of the importance of the data set sizes. The task is to be implemented in Matlab.
We shall work with computer generated data, i.e we know the true performance. In order to visualize we use two classes of Gaussian distributed scalar data x. The mean of the two classes are repectively $m_1 = -1$ and $m_2 = 1$. The two class variances are identical σ^2. Further assume equal probability (priors) $P_1 = P_2 = 0.5$ for the two classes.
Generate three datasets each of 1000 samples from each class. The three datasets of size 2000 shall have the following values for the variance for the classes :

1) $\sigma_1^2 = 0.25$

2) $\sigma_2^2 = 0.49$

3) $\sigma_3^2 = 1.00$

Save the three datasets. You shall use these datasets several times during this exercise!. The first 500 samples for each class will be used for training and the last 500 for testing.

Problem 1

(a) Plot histograms for the three **test data** sets using different colour for the two classes. What can you say about the difficulty of classifying them (in terms of the error rate)?

(b) Assume the true class densities are known. From the theory of statistics do calculate the true error rate for the three variance cases.

(c) Use the **true** densities as class models. Pick N_T samples for each class from the test data set and find the corresponding estimated error rate. Do this for :

- Three sizes $N_T = 5, 20, 100$ from each class
- Repeat the experiment 4 more times for each N_T. Be sure not to "reuse" any samples during these totally 5 tests
- Find the confusion matrixes and compare them.
- Also find the averaged confusion matrixes and the corresponding error rates over the five experiments for each given N_T.

One will typically find that

(d) Will the average error rate above differ from the error rate you would get if you used all the five test sets in one single test?

(e) Repeat all the above for the two other datasets (different variances).

Problem 2

Instead of using the true class models we will now train the class models. For each of the three datasets do the following :

(a) Use the first N_D training samples of each class to train/estimate the parameters of a Gaussian classifier. Use all 500 test samples for each class during a subsequent test. Do this for :

- Three training set sizes $N_D = 5, 20, 100$ from each class
- Repeat the experiment 4 times for each N_D by using a different training set samples (no "reuse"!).
- Find the confusion matrixes and compare them.
- Also find the averaged error rate over the five tests for each given N_D.

(b) Will this error rate differ from the error rate you would get if you used all the five training sets in one single training?

Problem 3

We will now inspect the leave-one-out testing procedure for small datasets.

(a) Assume that the three databases have only a total of 6 samples from each class. Use the first 6 samples from the training set.

(b) For each of the three datasets do the following :

- Take away sample number $i = 1, \ldots, 6$ for both classes for testing and train with the remaining 5 samples per class. Classify the two test samples.
- Find the average error rate over the 6 experiments. Compare with the performance of the comparable classifier (same number of training samples, $N_D = 5$) in problem 2.