TTT4275 Estimation, Detection and Classification
Problem Set Class1 Solution

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)?

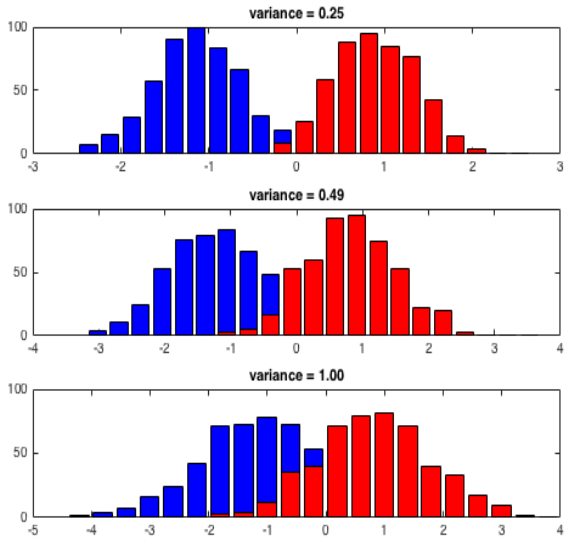

Figure 1: *Class histograms for the three different variances*

A: As expected the case with largest variance has most overlap and thus largest 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.

A: As the variance is equal for the two classes the treshold between the classes is given by $(m_1 + m_2)/2 = 0$. *Thus we have an error if a)* $x \in \omega_1$ *and* $x > 0$ *or b)* $x \in \omega_2$ *and* $x < 0$. *The two error cases contribute the same thus the probability of error is given by (using* ω_1*)*

$$P_e = P_1 * \int_0^\infty N(x, m_1, \sigma^2)dx + P_2 * \int_{-\infty}^0 N(x, m_2, \sigma^2)dx = \int_0^\infty N(x, m_1, \sigma^2)dx \quad (1)$$

Normalizing, i.e. $y = (x - m_1)/\sigma$ *means y has a N(0,1) distribution and the integral can be found by*

$$P_e = \int_\alpha^\infty N(y, 0, 1)dx = erfc(\alpha) \quad (2)$$

where $\alpha = -m_1/\sigma = +1/\sigma$ *and* $erfc$ *is the complementary error funcion*

Thus we have for the three cases : a) $P_e = erfc(1/0.5) = 0.0047$ *b)* $P_e = erfc(1/0.7) = 0.0434$ *c)* $P_e = erfc(1/1.00) = 0.1573$

(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 rate over the five experiments for each given N_T.

A: The matrixes and the error rates will be dependent on the values of the randomly generated data; i.e. will vary from trial to trial. This, combined with a final size of the test data, will lead to results both better and worse than the true error rates. However, the typically trend is that the variation decreases with increasing N_T. *Further we will see the same trend as for the true error rates; the error rates decrease with decreasing class variance* σ^2.

(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?

A: Since we use the same classifier (true class distributions) for all five test sets, we will get excactly the same error rates if we merge all five test sets into one.

(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 error rates and compare them.
- Also find the averaged error rate over the five tests for each given N_D.

A: When using a small amount N_D of training data (and no "reuse"), we can expect large variation in class model parameters (i.e. mean and variance). This again will lead to large variation in test error rates and confusion matrixes for the five different experiments. When N_D increases these differences will be smaller and smaller.

(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?

A: Now we have different models (estimated parameters) for each of the five experiments. If we merge all five training sets we will get a sixt model version.Thus the averaged error rate over the five models will differ from the error rate of these new models.