SYDE 675 Pattern Recognition

Assignment 2
Due Feb 23th 2016
(Assignments are to be done individually. Do not write a formal report.)

This assignment looks at experimental and analytical error rates.

We have the same three cases from the previous assignment, all of which are Gaussian. You should be able to re-use your code with minimal modifications.

To keep answers consistent, do *not* generate your own data points. Please use file assign2.mat from the LEARN.

1. $\mu_A = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\Sigma_A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ $\mu_B = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$ $\Sigma_B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
2. $\mu_A = \begin{bmatrix} -1 \\ 0 \end{bmatrix}$ $\Sigma_A = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$ $\mu_B = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ $\Sigma_B = \begin{bmatrix} 4 & 3 \\ 3 & 4 \end{bmatrix}$
3. $\mu_A = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\Sigma_A = \begin{bmatrix} 3 & 1 \\ 1 & 2 \end{bmatrix}$ $\mu_B = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$ $\Sigma_B = \begin{bmatrix} 7 & -3 \\ -3 & 4 \end{bmatrix}$
4. Two non-Gaussian classes, the data points found in assign2.mat.

In all cases the two classes are equally likely.

Analytical Error Rates:

First we consider deriving analytical error rates for the first three cases.

Derive the exact probability of error $P(\epsilon)$ for GED and MED for each of cases 1 and 2. Keep your derivations brief.

Case 3 is harder, because the covariances are not the same. Derive the exact probability of error $P(\epsilon)$ for MED. However don't try to work out $P(\epsilon)$ for GED, you will find it quite hard. Instead, discuss how you might attack this problem. That is, if I didn't let you use a computer, how would you go about setting up the problem to figure out $P_{GED}(\epsilon)$. Would you try the exact value or a bound? Why?

Experimental Error Rates:

Now we'll continue with three studies of experimental classification error:

1. **Classifier variability given very limited data:**

   We will use only five samples from each class (i.e., total of ten) to learn the classifier. So, for each of the four cases, and for each of the five distance-based classifiers (MED, GED, 1NN, 3NN, 5NN, but not MAP), we want to estimate the probability of error $P(\epsilon)$ based on the remaining $200 - 5 = 195$ data points per class.

   Since we have $N = 200$ data points per class, we can do this $200/5 = 40$ times, giving us 40 separate estimates of $P(\epsilon)$, allowing us to compute an overall average and standard deviation.

   Prepare a table showing $\mu_{P(\epsilon)}$ and $\sigma_{P(\epsilon)}$ for each of the four cases and five classifiers.

2. **Classifier accuracy and assessment using all of the data as training:**

   We want to use as many training samples as possible. We will use $N - 1 = 199$ data points per class, leaving two points (one from each class) for testing. Since two points are too little to meaningfully estimate $P(\epsilon)$, we will jackknife:

   ```
   for i=1:200
      Learn Classifier C_i, skipping data point i in each class
      Use the learned classifier C_i to classify the skipped points i
      Let m_i be the number of these which are misclassified, 0 ≤ m_i ≤ 2
   Estimate P(ε) = (1/200)(1/2) Σ_{i=1}^{200} m_i
   ```

   Prepare a table showing $P(\epsilon)$ for each of the four cases and five classifiers.

3. **Experimental vs. Analytical Errors:**

   Prepare a table comparing the experimental and analytical values of $P(\epsilon)$ for each of the first three cases for the MED and GED classifiers.

Discuss the results and comment on your observations for all three studies.