SD 372 Pattern Recognition

Winter 2009

Lab 1: Clusters and Classification Boundaries
Due Friday, February 9, 2008 (by 5:00pm)

(Lab is to be done in groups of up to 3 students. One report per group).

1 Purpose

This lab investigates three related areas: calculating orthonormal transformations, creating decision boundaries, and assessing classification error.

You may use whatever software you like in order to complete the lab, but Matlab is strongly encouraged.

Online Resources:

- A brief overview of Matlab, as well as a routine for plotting ellipses (plot_ellipse.m), are available on the course web page:
  http://ocho.uwaterloo.ca/∼pfieguth/Teaching/372/sd372.html
- Grammar, report writing, and figure advice:
  http://ocho.uwaterloo.ca/∼pfieguth/Teaching/grammar.html
- "Getting Started with Matlab " tutorial in PDF format:
  http://www.mathworks.com/access/helpdesk/help/pdf_doc/matlab/getstart.pdf
- "Matlab Summary and Tutorial":
  http://www.math.ufl.edu/help/matlab-tutorial/

Class Data

In this lab, consider five classes with the following bivariate (i.e., $n = 2$) Gaussian distribution parameters:

CASE 1:

$$\text{Class A:} \quad N_A = 200 \quad \mu_A = [5 \;\; 10]^T \quad \Sigma_A = \begin{bmatrix} 8 & 0 \\ 0 & 4 \end{bmatrix}$$

$$\text{Class B:} \quad N_B = 200 \quad \mu_B = [10 \;\; 15]^T \quad \Sigma_B = \begin{bmatrix} 8 & 0 \\ 0 & 4 \end{bmatrix}$$

CASE 2:

$$\text{Class C:} \quad N_C = 100 \quad \mu_C = [5 \;\; 10]^T \quad \Sigma_C = \begin{bmatrix} 8 & 4 \\ 4 & 40 \end{bmatrix}$$

$$\text{Class D:} \quad N_D = 200 \quad \mu_D = [15 \;\; 10]^T \quad \Sigma_D = \begin{bmatrix} 8 & 0 \\ 0 & 8 \end{bmatrix}$$

$$\text{Class E:} \quad N_E = 150 \quad \mu_E = [10 \;\; 5]^T \quad \Sigma_E = \begin{bmatrix} 10 & -5 \\ -5 & 20 \end{bmatrix}$$

2 Generating Clusters

1. Use the MATLAB function **randn** to assist in the generation of the 2D clusters above.

   The **randn** function will produce normally distributed data with mean 0 and variance 1.0. To create the correlated data as required, you will need to apply a transformation to the uncorrelated, equal-variance data.

2. Plot the samples and the unit standard deviation contour for each of the four classes. Put Classes A and B together on one plot; C, D, and E together on another. Visually, how does the unit contour relate to the cluster data?

3 Classifiers

For the two cases, plot the classification boundaries between the classes using

1. Minimum Euclidean Distance (MED), using the true means as the prototypes.
2. Generalized Euclidean Distance (GED), using the true means and covariances.
3. Maximum A Posterioi (MAP), using the true statistics. Set the *a priori* class probabilities proportional to the number of samples in each class.
4. Nearest neighbor (NN), using Euclidean distance.
5. k-Nearest neighbor (kNN) for $k = 5$, using Euclidean distance.

For each case, plot the class samples, unit standard deviation contours, and the MED, MICD and MAP boundaries on the same plot, and the class samples with the NN, 5NN boundaries on a separate plots. An analytical expression for the classification boundaries is not required nor ever desired; approach the problem numerically (e.g. create a 2D grid, classify each point, then do a contour plot). Using different plot line styles (try `help plot` in Matlab) will make the figure clearer.

Comment on the classification boundaries. How do the different boundaries compare?

4 Error Analysis

For each of the two cases, determine

1. The experimental error rate $P(\epsilon)$, and
2. The confusion matrix for each classifier (MED, MICD, MAP, NN, 5NN)

As NN, 5NN are a function of the individual data points, you will need to generate separate training and testing sets.

Compare the results. Which error is smallest? What do you observe in the confusion matrices for CASE2?

5 Report

Include in your report:

- A brief introduction.
- Discussion of your implementations and results (Include *brief* derivations, as appropriate, for equations implemented in M-files. Don't bother generating equations using a word processor. Handwritten equations are ok as long as they are readily legible.)
- Printouts of pertinent graphs.
- M-files for each section.
- Include answers to all questions.
- A brief summary of your results with conclusions.

Keep your report short! We are *not* looking for length.