

Problem Set 4

Handed Out: 3/3/2009

Due: 3/17/2004

This problem set is intended to be a machine problem set. You are prohibited from using Excel or other spreadsheet software for purposes other than plotting the data sets. You may find it convenient to code this machine problem set in MATLAB. The purpose of this exercise is to gain some familiarity with Pattern Recognition Algorithms. Instead of using data from the real world, we will do a Monte Carlo simulation based on pseudo-random numbers generated by a computer. Also, for the sake of simplicity, we will work on a four-class, two-dimensional problem.

The four classes are $\Omega = \{\omega_1, \omega_2, \omega_3, \omega_4\}$ and the measurements are of the form $\vec{x} \in \mathbf{R}^2$, that is $\vec{x} = (x_1, x_2)$. Let $p(\vec{x}|\omega_i) = N(\vec{x}, \vec{\mu}_i, \Sigma_i)$ for $i = 1 \dots 4$. In other words, $p(\vec{x}|\omega_i)$ is the normal distribution on observation \vec{x} with mean vector $\vec{\mu}_i$ and covariance matrix Σ_i that are as follows:

$$\left\{ \begin{array}{ll} \vec{\mu}_1 = (1, 1) & \Sigma_1 = \begin{bmatrix} 0.8 & 0.0 \\ 0.0 & 0.6 \end{bmatrix} \\ \vec{\mu}_2 = (1, -1) & \Sigma_2 = \begin{bmatrix} 0.5 & 0.0 \\ 0.0 & 0.8 \end{bmatrix} \\ \vec{\mu}_3 = (-1, -1) & \Sigma_3 = \begin{bmatrix} 0.7 & 0.0 \\ 0.0 & 0.7 \end{bmatrix} \\ \vec{\mu}_4 = (-1, 1) & \Sigma_4 = \begin{bmatrix} 0.3 & 0.0 \\ 0.0 & 0.5 \end{bmatrix} \end{array} \right. \quad (1)$$

Recall from lecture that we can generate samples from these distributions by using the univariate Gaussian random number generator:

$$y = \left[\left(\sum_{i=1}^{12} x_i \right) - 6 \right] \sigma + \mu \quad (2)$$

Each x_i is a random number uniformly distributed on $[0, 1]$, and μ and σ , respectively, are the mean and standard deviation of the distribution we would like to simulate. Most programming environments have an intrinsic function that will return one sample for each function call. Make sure the range of the samples is correct. You may have to scale them and adjust the mean value if the range is different from $[0, 1]$.

1. [Bayesian Classification]

- (a) Generate the data sets Y_i consisting of twenty samples each from the classes ω_i . Plot these 80 points on a common \mathbf{R}^2 plane. What does this plot represent? How do you know the plot is correct simply by looking at the way the points lie in the plane?
- (b) Compute the sample means and covariances of the sets Y_i and use these values to write formulas for $p(\vec{x}|\omega_i)$. Are these formulas identical to eq. (1) with values of (2) substituted for the appropriate parameters? Why or why not?

- (c) Now generate new data sets Z_i exactly as before except draw 100 samples per class. Again compute estimates of the parameter values from the sample means and covariances. How do these values compare with the previous estimates and the values in eq. (2)? Explain your results.
- (d) Let $Z = Z_1 \cup Z_2 \cup Z_3 \cup Z_4$. Classify the test data samples in Z according to the rule:

$$\vec{x} \in \omega^* \text{ if and only if } \omega^* = \operatorname{argmax}_{\omega_i} p(\omega_i | \vec{x}) \quad (3)$$

The parameter values in (4) are the estimates computed from Y_i . Plot the results. (Be sure to use different symbols for each class).

- (e) Compute the probability of classification error, P_e by counting the number of incorrect classifications and dividing by 400. Why is this an appropriate estimate for P_e ?

2. [Nearest Neighbor Classification]

- (a) Now classify the same set of test data Z using the nearest neighbor decision rule. Use the Euclidean metric

$$d(\vec{x}, \vec{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2} \quad (4)$$

to measure the necessary distances to data in the training set Y_i . Plot the results. (Be sure to use different symbols for each class)

- (b) Compute the probability of classification error P_e .
- (c) How doe P_e compare with the previous estimate? Explain your results.

3. [Unsupervised learning]

- (a) Write a program that implements the K-means algorithm and applies it to the set Z . Plot the results for the 2 and 4 cluster cases. (Be sure to use different symbols for each class)
- (b) Compute the ratio of average between-cluster to average between-cluster distance for both configurations. Explain your results.
- (c) Compute the probability of classification error P_e .
- (d) How doe P_e compare with the previous estimates? Explain your results.