

**Fall 2008, CPSC 689-607
Special Topics in Pattern Classification
Homework #1**

Due date: 9/18/2008

In recognition of the Texas A&M University policies of academic integrity, I certify that I have neither given nor received dishonest aid in this homework assignment.

Name: _____ Signature: _____

PLEASE FOLLOW THESE GUIDELINES:

1. Download the compressed file 'hw1.zip' from the course web page
2. Start each problem on a separate page
3. Provide hardcopies of all plots, MATLAB code and derivations
4. Sign and return this page with your finished assignment.
5. It is very important that you show your work and discuss your findings. This ensures full credit if your results are correct, and allows you to earn partial credit otherwise!

Problem 1 (25%)

(1) Given the set of patterns in Figure 1, which correspond to the vowels in the Korean alphabet, design a set of features (not more than 8) that provides good separability between all classes. Discuss your rationale.

(2) Generate MATLAB code to extract these features for each pattern. Generate two-dimensional scatter plots for each pair of features (e.g., f1 vs. f2, f3 vs. f4, etc). Discuss your results.

(3) Compute the Euclidean distance between each pair of examples in the original space (i.e., along 16,384 dimensions), and display it as a 50x50 matrix (HINT: use "imagesc"). Repeat the above, but in the low-dimensional space defined by your features. By comparing these two distance matrices, what can you tell about class separability in the original image space and in your final feature space? Discuss your results.



Figure 1. Dataset of handwritten Korean vowels

NOTES:

- An electronic copy of these patterns can be loaded from the file “hw1p1_data.mat” (“load hw1p1_data”). Each example is a 1x16,384 row vector. You may use the command “reshape” to convert each example into a 128x128 image, and the command “imagesc” to display the image.
- To generate scatter plots, use the command “text” and labels ‘0’ through ‘9’ for each of the ten classes. Use the command “axes” to adjust the range of the two axes.

Problem 2 (20%)

Assume that you are to build a probability density function of the produce from a California farm that grows two types of strawberries: maincrops and alpiners. The maincrops are larger, with an average size of 12 grams and a standard deviation of 2 grams. The alpiners are smaller, with an average size of 5 grams and a standard deviation of 3 gram. As an approximation, assume that the distributions are Gaussian. On an average season, the production of maincrops is twice as large as that of alpiners (in number of fruits, not weight)

- (a) Generate a single pdf to model the distribution of weights regardless of strawberry size. Generate a plot of this theoretical density.
- (b) Generate N=6,000 random samples according to this distribution. Generate a histogram of this sampled distribution using an appropriate number of bins.
- (c) Do the theoretical and experimental distributions match? If so, why? If not, why not?

HINT: Remember that the mass (area) of the histogram and the theoretical pdf have to be EACH equal to ONE.

Problem 3 (25%)

Consider a three-class recognition problem with equal priors, where the likelihood density functions are given by $N(2,3)$, $N(5,6)$ and $N(10,1)$, respectively, and the following cost function:

$$C_{ij} = \begin{bmatrix} 0 & 1 & \lambda \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

- (a) Derive an analytical expression for the conditional risk of each class $\mathfrak{R}(\alpha_i | x)$
- (b) Assuming a value of $\lambda=1$, generate a graph that illustrates the conditional risk $\mathfrak{R}(\alpha_i | x)$, the likelihood density $p(x | \omega_i)$, and the decision region for each class within $-10 \leq x \leq +10$.
- (c) Illustrate the asymptotic behavior of the decision regions as λ approaches infinity. Discuss your results.

Problem 4 (15%)

Let $p(x|\omega_i) = N(\mu_i, \Sigma)$ for a two-category d -dimensional problem with $p(\omega_1) = p(\omega_2) = 1/2$. It can be shown that the minimum probability of error is given by:

$$P_e = \frac{1}{\sqrt{2\pi}} \int_{r/2}^{\infty} e^{-u^2/2} du$$

where $r = (\mu_1 - \mu_2)^T \Sigma^{-1} (\mu_1 - \mu_2)$. Assuming that the features are independent (i.e., Σ is diagonal), discuss the asymptotic behavior of P_e as the dimensionality d approaches infinity. What are the theoretical implications for pattern-recognition purposes? What are the issues that may prevent these predictions to be met in practice?

Problem 5 (15%)

Given a multivariate density defined by its mean and covariance matrix:

$$\mu = \begin{bmatrix} 2 \\ 4 \end{bmatrix}; \quad \Sigma = \begin{bmatrix} 10 & 6 \\ 6 & 6 \end{bmatrix}$$

- (a) Generate $N=1,000$ random vectors, and estimate their mean and covariance matrix. Do they match the theoretical ones? If so, why? If not, why not?
- (b) Generate a 2D scatterplot of this data, and comment on the relationship between the structure of the scatter plot and that of the covariance matrix
- (c) Draw the locus of equidistant points from the mean at Mahalanobis distances of 1, 2, and 3 units. HINT: find those points in the data which are approximately $(\pm\varepsilon)$ at those distances.
- (d) Repeat parts (a), (b) and (c) for $\mu = \begin{bmatrix} 2 \\ 4 \end{bmatrix}; \Sigma = \begin{bmatrix} 10 & 0 \\ 0 & 6 \end{bmatrix}$
- (e) Discuss your results.