H3: Density Estimation and Classifiers

(1) Given the training patterns $x_1, x_2, \ldots, x_n$ from $K$ categories, where $n_1 + n_2 + \ldots + n_K = n$. Let the between-class scatter matrix $B$, the within-class scatter matrix $W$, and the total scatter matrix $T$ be defined by

\[ B = \sum_{i=1}^{K} n_i (u_i - u)(u_i - u)^t, \text{ where } u_i \text{ is the mean of ith category, } u = \frac{1}{n} \sum_{i=1}^{n} x_i, \]

\[ W = \sum_{i=1}^{K} \sum_{x \in \omega_i} (x - u_i)(x - u_i)^t. \]

\[ T = \sum_{i=1}^{n} (x_i - u)(x_i - u)^t. \]

Show that $B + W = T$.

(2) Problem 13 on page 144: Let $X \sim N(u, C)$, where $u$ is known and $C$ is unknown. Show that the MLE for $C$ is given by

\[ \hat{C} = \frac{1}{n} \sum_{k=1}^{n} (x_k - u)(x_k - u)^t \]

by carrying the following argument:

(a) Prove that $y^t A y = \text{tr}[A y y^t]$.

(b) Prove that the likelihood function can be written in the form

\[ p(x_1, x_2, \ldots, x_n | C) = (\sqrt{2\pi})^{-nd} |C|^{-n/2} \exp \left\{ -\frac{1}{2} \text{tr} \left[ C^{-1} \sum_{k=1}^{n} (x_k - u)(x_k - u)^t \right] \right\} \]

(c) Let $A = C^{-1} \hat{C}$ and $\lambda_1, \lambda_2, \ldots, \lambda_d$ be the eigenvalues of $A$; show that (b) leads to

\[ p(x_1, x_2, \ldots, x_n | \hat{C}) = (\sqrt{2\pi})^{-nd} |\hat{C}|^{-n/2} (\lambda_1 \lambda_2 \ldots \lambda_d)^{n/2} \exp \left\{ -\frac{n}{2} (\lambda_1 + \lambda_2 + \ldots + \lambda_d) \right\} \]

(d) Complete the proof by showing that the likelihood is maximized by the choice of $\lambda_1 = \lambda_2 = \ldots = \lambda_d = 1$. 
(3) Report the leave-one-out errors and the computation time of the Fisher’s linear classifier, quadratic classifier, and 1-nn classifier, respectively, on the data sets IMOX and iris as introduced in class.

http://www.cs.nthu.edu.tw/~cchen

(4) This problem is to learn how far the Bayes decision theory is from a practical application by simulations. Let \( X|\omega_1 \sim N(\mathbf{u}_1, C) \) and \( X|\omega_2 \sim N(\mathbf{u}_2, C) \), where

\[
\mathbf{u}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \mathbf{u}_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 \\ 0 & 0.25 \end{bmatrix}
\]

Assume the equal priors. An optimal decision rule and its theoretical error rate was reported in the last assignment. Now do the following tasks.

(a) Generate 100 patterns \( \{x_i\} \) independently from \( N(\mathbf{u}_1, C) \).
(b) Generate 100 patterns \( \{y_i\} \) independently from \( N(\mathbf{u}_2, C) \).
(c) Apply MLE to find \( \hat{\mathbf{u}}_1 \) and \( \hat{C}_1 \).
(d) Apply MLE to find \( \hat{\mathbf{u}}_2 \) and \( \hat{C}_2 \).
(e) Compute the error rate based on MAP (a quadratic classifier).
(f) How close is the result in (e) to that in the last assignment?

(5) This problem is to verify the density estimation methods by the Parzen window and the K-nn distance approaches.

(a) Generate 400 patterns independently from \( N(0, 1) \).
(b) Compute the estimated probability density \( \hat{p}(x) \) using the density function of the standard normal distribution as a kernel with some appropriate values of \( h \) and \( \alpha \) introduced in class (report your \( h \) and \( \alpha \) values).
(c) Compute the estimated probability density \( \hat{f}(x) \) using the K-nn distance approach with an appropriate value \( K \).
(d) Compare the results of (b) and (c) with the true density by plotting \( \hat{p}(x) \) and \( \hat{f}(x) \) vs. \( x \), respectively.
(e) Repeat steps (a) \( \sim \) (d) for the exponential distribution \( f(x) = \frac{1}{2}e^{-x/2}, x \geq 0 \).