## 4.4 ESTIMATION OF PROBABILITY DENSITY FUNCTIONS

The most important task in implementing a statistical approach for solving pattern classification problems is estimating the density functions $p(\boldsymbol{x}|C_i)$, $1 \le i \le m$. We will first show how to use the *maximum entropy principle* to obtain the *form* of probability density functions.

### 4.4.1 Form of the Density Function

The principle of maximum entropy states that in the case where a probability density function of a random variable is not known, the function which maximizes the entropy of this variable subject to known specified constraints is an appropriate choice. Any other choice would show a bias to some information obtained from the given data. The maximum entropy solution is easily derived when the constraints are given in the form of averages associated with the probability density function. Given a probability density function $p(\boldsymbol{x})$, the associated entropy is

$$E = -\int_{\boldsymbol{x}} p(\boldsymbol{x}) \ln[p(\boldsymbol{x})] d\boldsymbol{x} \tag{4.4.1}$$

and we assume the constraints

$$\int_{\boldsymbol{x}} f_i(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} = \alpha_i \ , \quad 0 \le i \le M \tag{4.4.2}$$

where $f_0(\boldsymbol{x}) = 1$ and $\alpha_0 = 1$. We wish to obtain $p(\boldsymbol{x})$ which satisfies Eq. (4.4.2) while minimizing the entropy $E$ of Eq. (4.4.1). This is done using Lagrange multipliers. Define

$$E_1 = E + \sum_{i=0}^{M} \lambda_i [\int_{\boldsymbol{x}} f_i(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} - \alpha_i] \tag{4.4.3}$$

where the constants $\lambda_i$, $0 \le i \le M$ are yet to be determined. By virtue of Eq. (4.4.1) we get

$$E_1 = -\int_x p(x)[\ln[p(x)] - \sum_{i=0}^{M} \lambda_i f_i(x)]dx - \sum_{i=0}^{M} \lambda_i \alpha_i \tag{4.4.4}$$

The partial derivative of $E_1$ with respect to $p(x)$ is

$$\frac{\partial E_1}{\partial [p(x)]} = -\int_x [\ln[p(x)] - \sum_{i=0}^{M} \lambda_i f_i(x) + 1]dx \tag{4.4.5}$$

and to obtain the maximum entropy solution the integrand must vanish, i.e.

$$p(x) = \exp[\sum_{i=0}^{M} \lambda_i f_i(x) - 1] \tag{4.4.6}$$

We still have freedom of choosing $\lambda_i$, $a \le i \le m$ and these coefficients are chosen so that Eq. (4.4.2) holds. Once the form of the probability density function is known, we may turn and perform the next step: estimating the parameters of this density.

■ **Example 4.4.1** Consider a random variable $x$ which is characterized by

$$a < x < b \ , \quad \int_0^{\infty} p(x)dx = 1$$

By virtue of Eq. (4.4.6) we obtain

$$p(x) = \exp(\lambda_0 - 1) \ , \quad \int_a^b \exp(\lambda_0 - 1)dx = 1$$

and therefore

$$p(x)=\begin{cases}\dfrac{1}{b-a}, & a<x<b\\ 0, & \text{otherwise}\end{cases}$$

♠

■ **Example 4.4.2** Assume that the *a priori* information about $x$ is

$$x\geq 0\ ,\ \int_0^\infty p(x)dx=1\ ,\ \int_0^\infty xp(x)dx=\mu$$

Using Eq. (4.4.6) we obtain

$$p(x)=\exp(\lambda_0-1+\lambda_1 x)$$

and in order to satisfy the two constraints the density function must be

$$p(x)=\begin{cases}(1/\mu)\exp(-x/\mu), & x\geq 0\\ 0, & \text{otherwise}\end{cases}$$

♠

### 4.4.2 Estimating the Mean Vector and Covariance Matrix

Consider a pattern population with probability density function $p(\boldsymbol{x})$. The mean vector of this population is given by

$$\boldsymbol{\mu}=E(\boldsymbol{x})=\int_{\boldsymbol{x}}\boldsymbol{x}p(\boldsymbol{x})d\boldsymbol{x} \tag{4.4.7}$$

If the patterns are in $R^n$, then $\boldsymbol{\mu}$ is a vector with $n$ components $(\mu_1,\ldots,\mu_n)$. Let $\{\boldsymbol{x}_i\}_{i=1}^N$ denote the given patterns. An approximate to $\boldsymbol{\mu}$ is simply

$$\boldsymbol{\mu}\approx\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i \tag{4.4.8}$$

The covariance matrix $\boldsymbol{C} = (c_{jk})$ , $1 \leq j,k \leq n$ satisfies

$$c_{jk} = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_j - \mu_j)(x_k - \mu_k) p(x_j, x_k) dx_j dx_k \tag{4.4.9}$$

We can also rewrite $\boldsymbol{C}$ as

$$\begin{aligned} \boldsymbol{C} &= E[(\boldsymbol{x} - \mu)(\boldsymbol{x} - \mu)^T] \\ &= E[\boldsymbol{x}\boldsymbol{x}^T - 2x\mu^T + \mu\mu^T] \\ &= E[\boldsymbol{x}\boldsymbol{x}^T] - \mu\mu^T \end{aligned} \tag{4.4.10}$$

and use the new expression to approximate $\boldsymbol{C}$ as

$$\boldsymbol{C} \approx \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^T - \mu\mu^T \tag{4.4.11}$$

Both estimates for $\mu$ and for $\boldsymbol{C}$ can be conveniently used in a recursive manner. Let $N$ be the current number of sample patterns and assume on additional incoming pattern. Denote by $\mu(N), \boldsymbol{C}(N)$ the current mean vector and covariance matrix. Then

$$\begin{aligned} \mu(N+1) &= \frac{1}{N+1} \sum_{i=1}^{N+1} \boldsymbol{x}_i \\ &= \frac{1}{N+1} \left( \sum_{i=1}^{N+1} \boldsymbol{x}_i + \boldsymbol{x}_{N+1} \right) \\ &= \frac{1}{N+1} (N\mu(N) + \boldsymbol{x}_{N+1}) \end{aligned} \tag{4.4.12}$$

where $\mu(1) = \boldsymbol{x}_1$. This recursive expression updates the mean vector.

In the case of the covariance matrix we obtain

$$\begin{aligned} \boldsymbol{C}(N+1) &= \frac{1}{N+1}\sum_{i=1}^{N+1}\boldsymbol{x}_i\boldsymbol{x}_i^T - \mu(N+1)\mu^T(N+1) \\ &= \frac{1}{N+1}\left(\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^T + \boldsymbol{x}_{N+1}\boldsymbol{x}_{N+1}^T\right) - \mu(N+1)\mu^T(N+1) \\ &= \frac{1}{N+1}(N\boldsymbol{C}(N) + N\mu(N)\mu^T(N) + \boldsymbol{x}_{N+1}\boldsymbol{x}_{N+1}^T) \\ &\quad - \frac{1}{(N+1)^2}(N\mu(N) + \boldsymbol{x}_{N+1})(N\mu^T(N) + \boldsymbol{x}_{N+1}^T) \end{aligned} \tag{4.4.13}$$

To start the calculation of $\boldsymbol{C}(N)$ we use the relation

$$\boldsymbol{C}(1) = \boldsymbol{x}_1\boldsymbol{x}_1^T - \mu(1)\mu^T(1)$$

to obtain $\boldsymbol{C}(1)$=0.

■ **Example 4.4.3** Consider the sample patterns

$$\boldsymbol{x}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \boldsymbol{x}_2 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \boldsymbol{x}_3 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \boldsymbol{x}_4 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$$

To start the recursive procedure we set

$$\mu(1) = \boldsymbol{x}_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \boldsymbol{C}_1 = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$$

and then, using Eqs. (4.4.12-13), obtain

$$\mu(2) = \begin{pmatrix} 1/2 \\ 0 \end{pmatrix}, \quad \mu(3) = \begin{pmatrix} 2/3 \\ 1/3 \end{pmatrix}, \quad \mu(4) = \begin{pmatrix} 3/4 \\ 3/4 \end{pmatrix},$$

$$\boldsymbol{C}(2) = \begin{pmatrix} 1/4 & 0 \\ 0 & 0 \end{pmatrix}, \quad \boldsymbol{C}(3) = \begin{pmatrix} 2/9 & 1/9 \\ 1/9 & 2/9 \end{pmatrix}, \quad \boldsymbol{C}(4) = \begin{pmatrix} 3/16 & 3/16 \\ 3/16 & 11/16 \end{pmatrix}$$

♠

### 4.4.3 Estimation by Functional Approximation

If the form of the density function is not known we may estimate it directly using functional approximation.

Let $p(\boldsymbol{x})$ denote the probability density function $p(\boldsymbol{x} \mid C)$ and consider an approximate $\tilde{p}(\boldsymbol{x})$ given by

$$\tilde{p}(\boldsymbol{x}) = \sum_{i=1}^{m} a_i \phi_i(\boldsymbol{x}) \tag{4.4.14}$$

where $\{\phi_i(\boldsymbol{x})\}_{i=1}^{m}$ are specified basis functions. We wish to minimize

$$E = \int_{\boldsymbol{x}} w(\boldsymbol{x})[p(\boldsymbol{x}) - \tilde{p}(\boldsymbol{x})]^2 d\boldsymbol{x} \tag{4.4.15}$$

or

$$E = \int_{\boldsymbol{x}} w(\boldsymbol{x}) \left[ p(\boldsymbol{x}) - \sum_{i=1}^{m} a_i \phi_i(\boldsymbol{x}) \right]^2 d\boldsymbol{x} \tag{4.4.16}$$

where $w(x)$ is a specified weight function. Solving the system

$$\frac{\partial E}{\partial a_i} = 0 \ , \quad i = 1, \ldots, m \tag{4.4.17}$$

provides a set of linear equations

$$\sum_{j=1}^{m} a_i \int_{\boldsymbol{x}} w(\boldsymbol{x}) \phi_j(\boldsymbol{x}) \phi_i(\boldsymbol{x}) d\boldsymbol{x} = \int_{\boldsymbol{x}} w(\boldsymbol{x}) \phi_i(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x}, \quad 1 \le i \le m \tag{4.4.18}$$

The right-hand sides of these equations are simply the expected values of $w(\boldsymbol{x})\phi_i(\boldsymbol{x})$, $1 \le i \le m$. If $\{\boldsymbol{x}_k\}_{k=1}^{N}$ are given samples which belong to $C$, these expected values can be estimated as

$$\int_{\boldsymbol{x}} w(\boldsymbol{x}) \phi_i(\boldsymbol{x}) p(\boldsymbol{x}) d\boldsymbol{x} \approx \frac{1}{N} \sum_{k=1}^{N} w(\boldsymbol{x}_k) \phi_i(\boldsymbol{x}_k) \tag{4.4.19}$$

We thus replace the system given by Eq. (4.4.18) by

$$\sum_{j=1}^{m} a_i \int_x w(\boldsymbol{x})\phi_j(\boldsymbol{x})\phi_i(\boldsymbol{x})d\boldsymbol{x} = \frac{1}{N}\sum_{k=1}^{N} w(\boldsymbol{x}_k)\phi_i(\boldsymbol{x}_k)\ , \quad 1 \le i \le m \tag{4.4.20}$$

In the particular case where $\phi_i(\boldsymbol{x})$, $1 \le i \le k$ are orthogonal with respect to $w(\boldsymbol{x})$, we have

$$\int_x w(\boldsymbol{x})\phi_j(\boldsymbol{x})\phi_i(\boldsymbol{x})d\boldsymbol{x} = \begin{cases} A_i\ , & j = i \\ 0\ , & j \ne i \end{cases} \tag{4.4.21}$$

and consequently

$$a_i = \frac{1}{NA_i}\sum_{k=1}^{N} w(\boldsymbol{x}_k)\phi_i(\boldsymbol{x}_k) \tag{4.4.22}$$

The expression given by Eq. (4.4.22) provides an easy way to obtain $a_i(N+1)$ from $a_i(N)$. Indeed

$$a_i(N+1) = \frac{1}{(N+1)A_i}\sum_{k=1}^{N+1} w(\boldsymbol{x}_k)\phi_i(\boldsymbol{x}_k)$$

$$= \frac{1}{(N+1)A_i}\left[NA_i a_i(N) + w(\boldsymbol{x}_{N+1})\phi_i(\boldsymbol{x}_{N+1})\right] \tag{4.4.23}$$

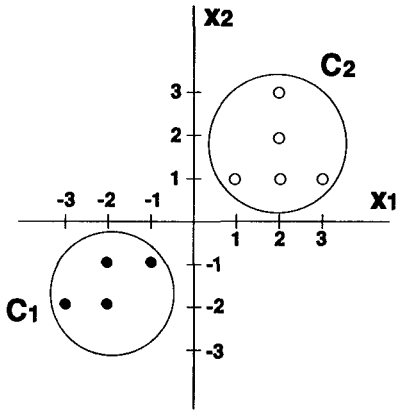
In decision making problems, since the terms $w(\boldsymbol{x}_k)$ in Eq. (4.4.22) are independent of $i$ and are therefore common to all the coefficients, they can usually be eliminated from the process, without violating the discriminatory characteristics of the coefficients. We may usually therefore, for such problems, apply a simpler relation

$$a_i = \frac{1}{N}\sum_{k=1}^{N} \phi_i(\boldsymbol{x}_k) \tag{4.4.24}$$

i.e. relate to the orthogonal basis functions, as if they were orthonormal and simplifying the computations.

Since $p(\boldsymbol{x})$ is not known, one may not be able to decide how large $m$ should be. Usually we start with a prefixed $m$ and experiment with the training set to determine whether $\tilde{p}(\boldsymbol{x})$ is an acceptable approximate to $p(\boldsymbol{x})$. If the classification performance of $\tilde{p}(\boldsymbol{x})$ is poor, we increase $m$ until we reach a 'saturation' state, i.e. until adding new terms has no effect on the classification quality of $\tilde{p}(\boldsymbol{x})$. It can be shown that in general $\tilde{p}(\boldsymbol{x})$ approaches $p(\boldsymbol{x})$ as $m \to \infty$ and $N \to \infty$.

■ **Example 4.4.4** Consider the two-class classification problem given in Fig. 4.4.1




■ **Figure 4.4.1** Bayes classification using functional approximation.

Assuming that the entire domain of the patterns is the whole plane, one is tempted to use the Hermite polynomials which are orthogonal over the interval $(-\infty, \infty)$. The first two polynomials are $H_0(x) = 1$, $H_1(x) = 2x$. If four basis functions are considered we may choose

$$\phi_1(\boldsymbol{x}) = H_0(x_1)H_0(x_2) = 1$$

$$\phi_2(\boldsymbol{x}) = H_1(x_1)H_0(x_2) = 2x_1$$

$$\phi_3(\boldsymbol{x}) = H_0(x_1)H_1(x_2) = 2x_2$$

$$\phi_4(\boldsymbol{x}) = H_1(x_1)H_1(x_2) = 4x_1x_2$$

We may treat the functions as if they were orthonormal and obtain

$$a_i^{(l)} = \frac{1}{N_l}\sum_{k=1}^{N_l}\phi_i\left(\boldsymbol{x}_k^{(l)}\right), \quad 1 \le l \le 2$$

where $a_i^{(l)}$ are the coefficients associated with class $l$, $N_l$ – the number of patterns in class $l$, and $\boldsymbol{x}_k^{(l)}$ are the patterns in class $l$. Thus

$$a_1^{(1)} = \frac{1}{4}(1+1+1+1) = 1\ , \qquad a_2^{(1)} = \frac{1}{4}(-2-4-4-6) = -4$$

$$a_3^{(1)} = \frac{1}{4}(-2-2-4-4) = -3\ , \qquad a_4^{(1)} = \frac{1}{4}(4+8+16+24) = 13$$

$$a_1^{(2)} = \frac{1}{5}(1+1+1+1+1) = 1\ , \qquad a_2^{(2)} = \frac{1}{5}(2+4+4+4+6) = 4$$

$$a_3^{(2)} = \frac{1}{5}(2+2+2+4+6) = 3.2\ , \qquad a_4^{(2)} = \frac{1}{5}(4+8+12+16+24) = 12.8$$

and consequently

$$\tilde{p}(\boldsymbol{x} \mid C_1) = 1 - 8x_1 - 6x_2 + 52x_1x_2$$

$$\tilde{p}(\boldsymbol{x} \mid C_1) = 1 + 8x_1 + 6.4x_2 + 51.2x_1x_2$$

The decision functions are

$$d_1(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x} \mid C_1)p(C_1)$$

$$d_2(\boldsymbol{x}) = \tilde{p}(\boldsymbol{x} \mid C_2)p(C_2)$$

and by assuming $p(C_1) = p(C_2) = 1/2$ we get

$$d_1(\boldsymbol{x}) = \frac{1}{2} - 4x_1 - 3x_2 + 26x_1x_2$$

$$d_2(\boldsymbol{x}) = \frac{1}{2} + 4x_1 + 3.2x_2 + 25.6x_1x_2$$

The decision boundary is therefore

$$d_{12}(\boldsymbol{x}) = d_1(\boldsymbol{x}) - d_2(\boldsymbol{x}) = -8x_1 - 6.2x_2 + 0.4x_1x_2 = 0$$

♠

## PROBLEMS

1. Use the maximum entropy principle to obtain the probability density function if the information

$$-\infty < x < \infty\,,\quad \int_{-\infty}^{\infty} p(x)dx = 1\,,\quad \int_{-\infty}^{\infty} xp(x)dx = \mu\,,\quad \int_{-\infty}^{\infty} x^2 p(x)dx = \sigma^2$$

   is *a priori* known.

2. Given the sample patterns

$$\boldsymbol{x}_1 = \begin{pmatrix} 1 \\ 0 \\ 1 \end{pmatrix},\quad \boldsymbol{x}_2 = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix},\quad \boldsymbol{x}_3 = \begin{pmatrix} 1 \\ -1 \\ 1 \end{pmatrix},\quad \boldsymbol{x}_4 = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}$$

Use Eqs. (4.4.12-13) to estimate $\mu(i)$, $\boldsymbol{C}(i)$, $1 \le i \le 4$.

3. Apply the method of functional approximation to get estimates of $p(\boldsymbol{x}|C_1)$ and $p(\boldsymbol{x}|C_2)$ where

$$C_1 = \{(1,0)^T,\ (1,1)^T, (2,1)^T, (3,0)^T, (4,1)^T\}$$

$$C_2 = \{(-1,0)^T,\ (-2,0)^T, (-2,-1)^T, (-3,1)^T, (-3,2)^T\}$$

Use the first three 2-D Hermite polynomials and Eq. (4.4.24) to obtain the coefficients.

4. Repeat problem 3 but use the first four Hermite polynomials.

5. Repeat problems 3 and 4 using Hermite orthonormal functions. Replace $H_k(x)$ by

$$H_k^*(x) = \frac{\exp(-x^2/2)}{\sqrt{2^k k! \sqrt{\pi}}} H_k(x)$$

but still use Eq. (4.4.24).