CHAPTER

2

Classifiers Based on Bayes Decision Theory

2.1 INTRODUCTION

This is the first chapter, out of three, dealing with the design of the classifier in a pattern recognition system. The approach to be followed builds upon probabilistic arguments stemming from the statistical nature of the generated features. As has already been pointed out in the introductory chapter, this is due to the statistical variation of the patterns as well as to the noise in the measuring sensors. Adopting this reasoning as our kickoff point, we will design classifiers that classify an unknown pattern in the most probable of the classes. Thus, our task now becomes that of defining what "most probable" means.

Given a classification task of M classes, $\omega_1, \omega_2, \ldots, \omega_M$, and an unknown pattern, which is represented by a feature vector $\boldsymbol{x}$, we form the M conditional probabilities $P(\omega_i|\boldsymbol{x})$, $i = 1, 2, \ldots, M$. Sometimes, these are also referred to as *a posteriori probabilities*. In words, each of them represents the probability that the unknown pattern belongs to the respective class ω_i, given that the corresponding feature vector takes the value $\boldsymbol{x}$. Who could then argue that these conditional probabilities are not sensible choices to quantify the term *most probable*? Indeed, the classifiers to be considered in this chapter compute either the maximum of these M values or, equivalently, the maximum of an appropriately defined function of them. The unknown pattern is then assigned to the class corresponding to this maximum.

The first task we are faced with is the computation of the conditional probabilities. The Bayes rule will once more prove its usefulness! A major effort in this chapter will be devoted to techniques for estimating probability density functions (pdf), based on the available experimental evidence, that is, the feature vectors corresponding to the patterns of the training set.

2.2 BAYES DECISION THEORY

We will initially focus on the two-class case. Let ω_1, ω_2 be the two classes in which our patterns belong. In the sequel, we assume that the *a priori probabilities*

$P(\omega_1), P(\omega_2)$ are known. This is a very reasonable assumption, because even if they are not known, they can easily be estimated from the available training feature vectors. Indeed, if N is the total number of available training patterns, and N_1, N_2 of them belong to ω_1 and ω_2, respectively, then $P(\omega_1) \approx N_1/N$ and $P(\omega_2) \approx N_2/N$.

The other statistical quantities assumed to be known are the class-conditional probability density functions $p(\boldsymbol{x}|\omega_i), i = 1, 2$, describing the distribution of the feature vectors in each of the classes. If these are not known, they can also be estimated from the available training data, as we will discuss later on in this chapter. The pdf $p(\boldsymbol{x}|\omega_i)$ is sometimes referred to as the *likelihood function of ω_i with respect to $\boldsymbol{x}$*. Here we should stress the fact that an implicit assumption has been made. That is, the feature vectors can take any value in the l-dimensional feature space. In the case that feature vectors can take only discrete values, density functions $p(\boldsymbol{x}|\omega_i)$ become probabilities and will be denoted by $P(\boldsymbol{x}|\omega_i)$.

We now have all the ingredients to compute our conditional probabilities, as stated in the introduction. To this end, let us recall from our probability course basics the *Bayes rule* (Appendix A)

$$P(\omega_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\omega_i)P(\omega_i)}{p(\boldsymbol{x})} \tag{2.1}$$

where $p(\boldsymbol{x})$ is the pdf of $\boldsymbol{x}$ and for which we have (Appendix A)

$$p(\boldsymbol{x}) = \sum_{i=1}^{2} p(\boldsymbol{x}|\omega_i)P(\omega_i) \tag{2.2}$$

The *Bayes classification rule* can now be stated as

$$\begin{aligned}
&\text{If} \quad P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x}), \quad \boldsymbol{x} \text{ is classified to } \omega_1 \\
&\text{If} \quad P(\omega_1|\boldsymbol{x}) < P(\omega_2|\boldsymbol{x}), \quad \boldsymbol{x} \text{ is classified to } \omega_2
\end{aligned} \tag{2.3}$$

The case of equality is detrimental and the pattern can be assigned to either of the two classes. Using (2.1), the decision can equivalently be based on the inequalities

$$p(\boldsymbol{x}|\omega_1)P(\omega_1) \gtrless p(\boldsymbol{x}|\omega_2)P(\omega_2) \tag{2.4}$$

$p(\boldsymbol{x})$ is not taken into account, because it is the same for all classes and it does not affect the decision. Furthermore, if the *a priori* probabilities are equal, that is, $P(\omega_1) = P(\omega_2) = 1/2$, Eq. (2.4) becomes

$$p(\boldsymbol{x}|\omega_1) \gtrless p(\boldsymbol{x}|\omega_2) \tag{2.5}$$

Thus, the search for the maximum now rests on the values of the conditional pdfs evaluated at $\boldsymbol{x}$. Figure 2.1 presents an example of two equiprobable classes and shows the variations of $p(x|\omega_i), i = 1, 2$, as functions of x for the simple case of a single feature ($l = 1$). The dotted line at x_0 is a threshold partitioning the feature space into two regions, R_1 and R_2. According to the Bayes decision rule, for all values of x in R_1 the classifier decides ω_1 and for all values in R_2 it decides ω_2. However, it is obvious from the figure that decision errors are unavoidable. Indeed, there is

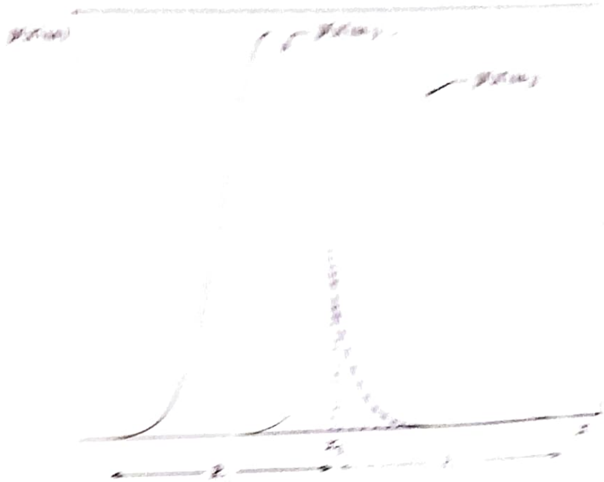


FIGURE 2.1

Example of the two regions R_1 and R_2 formed by the Bayesian classifier for the case of two equiprobable classes.

a finite probability for an x to lie in the R_2 region and at the same time to belong in class ω_1. Then our decision is in error. The same is true for points originating from class ω_2. It does not take much thought to see that the total probability, P_e, of committing a decision error for the case of two equiprobable classes, is given by

$$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x|\omega_2)\,dx + \frac{1}{2}\int_{x_0}^{+\infty} p(x|\omega_1)\,dx \tag{2.6}$$

which is equal to the total shaded area under the curves in Figure 2.1. We have now touched on a very important issue. Our starting point to arrive at the Bayes classification rule was rather empirical, via our interpretation of the term *most probable*. We will now see that this classification test, though simple in its formulation, has a sounder mathematical interpretation.

Minimizing the Classification Error Probability

We will show that the Bayesian classifier is optimal with respect to minimizing the classification error probability. Indeed, the reader can easily verify, as an exercise, that moving the threshold away from x_0, in Figure 2.1, always increases the corresponding shaded area under the curves. Let us now proceed with a more formal proof.

Proof. Let R_1 be the region of the feature space in which we decide in favor of ω_1 and R_2 be the corresponding region for ω_2. Then an error is made if $\mathbf{x} \in R_1$, although it belongs to ω_2 or if $\mathbf{x} \in R_2$, although it belongs to ω_1. That is

$$P_e = P(\mathbf{x} \in R_2, \omega_1) + P(\mathbf{x} \in R_1, \omega_2) \tag{2.7}$$

P(A ∩ B) = P(A, B)

P(A | B) P(B) =

P(B | A) P(A)

where $P(\cdot,\cdot)$ is the joint probability of two events. Recalling, once more, our probability basics (Appendix A), this becomes

$$P_e = P(\boldsymbol{x} \in R_2|\omega_1)P(\omega_1) + P(\boldsymbol{x} \in R_1|\omega_2)P(\omega_2)$$
$$= P(\omega_1)\int_{R_2} p(\boldsymbol{x}|\omega_1)\,d\boldsymbol{x} + P(\omega_2)\int_{R_1} p(\boldsymbol{x}|\omega_2)\,d\boldsymbol{x} \tag{2.8}$$

or using the Bayes rule

$$P_e = \int_{R_2} P(\omega_1|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} + \int_{R_1} P(\omega_2|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} \tag{2.9}$$

It is now easy to see that the error is minimized if *the partitioning regions R_1 and R_2 of the feature space are chosen so that*

$$R_1: P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$$
$$R_2: P(\omega_2|\boldsymbol{x}) > P(\omega_1|\boldsymbol{x}) \tag{2.10}$$

Indeed, since the union of the regions R_1, R_2 covers all the space, from the definition of a probability density function we have that

$$\int_{R_1} P(\omega_1|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} + \int_{R_2} P(\omega_1|\boldsymbol{x})p(\boldsymbol{x})\,d\boldsymbol{x} = P(\omega_1) \tag{2.11}$$

Combining Eqs. (2.9) and (2.11), we get

$P(\omega_1) - \int_{R_1} P(\omega_1|x)p(x)dx$

$$P_e = P(\omega_1) - \int_{R_1} (P(\omega_1|\boldsymbol{x}) - P(\omega_2|\boldsymbol{x}))\,p(\boldsymbol{x})\,d\boldsymbol{x} \tag{2.12}$$

This suggests that the probability of error is minimized if R_1 is the region of space in which $P(\omega_1|\boldsymbol{x}) > P(\omega_2|\boldsymbol{x})$. Then, R_2 becomes the region where the reverse is true. □

So far, we have dealt with the simple case of two classes. Generalizations to the multiclass case are straightforward. In a classification task with M classes, $\omega_1, \omega_2, \ldots, \omega_M$, an unknown pattern, represented by the feature vector $\boldsymbol{x}$, is assigned to class ω_i if

$$P(\omega_i|\boldsymbol{x}) > P(\omega_j|\boldsymbol{x}) \quad \forall j \neq i \tag{2.13}$$

It turns out that such a choice also minimizes the classification error probability (Problem 2.1).

Minimizing the Average Risk

The classification error probability is not always the best criterion to be adopted for minimization. This is because it assigns the same importance to all errors. However, there are cases in which some wrong decisions may have more serious implications than others. For example, it is much more serious for a doctor to make a wrong decision and a malignant tumor to be diagnosed as a benign one, than the other way round. If a benign tumor is diagnosed as a malignant one, the wrong decision will be cleared out during subsequent clinical examinations. However, the results

malignant tumor more harmful than benign from the wrong decision concerning a malignant tumor may be fatal. Thus, in such cases it is more appropriate to assign a penalty term to weigh each error. For our example, let us denote by ω_1 the class of malignant tumors and as ω_2 the class of the benign ones. Let, also, R_1, R_2 be the regions in the feature space where we decide in favor of ω_1 and ω_2, respectively. The error probability P_e is given by Eq. (2.8). Instead of selecting R_1 and R_2 so that P_e is minimized, we will now try to minimize a modified version of it, that is,

$$r = \lambda_{12}P(\omega_1)\int_{R_2} p(\boldsymbol{x}|\omega_1)d\boldsymbol{x} + \lambda_{21}P(\omega_2)\int_{R_1} p(\boldsymbol{x}|\omega_2)d\boldsymbol{x} \tag{2.14}$$

where each of the two terms that contributes to the overall error probability is weighted according to its significance. For our case, the reasonable choice would be to have $\lambda_{12} > \lambda_{21}$. Thus errors due to the assignment of patterns originating from class ω_1 to class ω_2 will have a larger effect on the cost function than the errors associated with the second term in the summation.

Let us now consider an M-class problem and let $R_j, j = 1, 2, \ldots, M$, be the regions of the feature space assigned to classes ω_j, respectively. Assume now that a feature vector $\boldsymbol{x}$ that belongs to class ω_k lies in $R_i, i \neq k$. Then this vector is misclassified in ω_i and an error is committed. A penalty term λ_{ki}, known as *loss*, is associated with this wrong decision. The matrix L, which has at its (k, i) location the corresponding penalty term, is known as the *loss matrix*.[1] Observe that in contrast to the philosophy behind Eq. (2.14), we have now allowed weights across the diagonal of the loss matrix (λ_{kk}), which correspond to correct decisions. In practice, these are usually set equal to zero, although we have considered them here for the sake of generality. The *risk or loss* associated with ω_k is defined as

$$r_k = \sum_{i=1}^{M} \lambda_{ki} \int_{R_i} p(\boldsymbol{x}|\omega_k)\, d\boldsymbol{x} \tag{2.15}$$

Observe that the integral is the overall probability of a feature vector from class ω_k being classified in ω_i. This probability is weighted by λ_{ki}. Our goal now is to choose the partitioning regions R_j so that the *average risk*

$$r = \sum_{k=1}^{M} r_k P(\omega_k)$$

$$= \sum_{i=1}^{M} \int_{R_i} \left(\sum_{k=1}^{M} \lambda_{ki}\, p(\boldsymbol{x}|\omega_k)P(\omega_k) \right) d\boldsymbol{x} \tag{2.16}$$

is minimized. This is achieved if each of the integrals is minimized, which is equivalent to selecting partitioning regions so that

$$\boldsymbol{x} \in R_i \quad \text{if} \quad l_i \equiv \sum_{k=1}^{M} \lambda_{ki}\, p(\boldsymbol{x}|\omega_k)P(\omega_k) < l_j \equiv \sum_{k=1}^{M} \lambda_{kj}\, p(\boldsymbol{x}|\omega_k)P(\omega_k) \quad \forall j \neq i \tag{2.17}$$

[1] The terminology comes from the general decision theory.

$W_1 \to$ malignant
$W_2 \to$ benign

$\lambda_{12} \to$ w1 of class 1 but assigning it to class 2

$x \to \omega_k$ but assign to R_i
$i \neq K$

λ_{ki}

It is obvious that if $\lambda_{ki} = 1 - \delta_{ki}$, where δ_{ki} is *Kronecker's delta* (0 if $k \neq i$ and 1 if $k = i$), then minimizing the average risk becomes equivalent to minimizing the classification error probability.

The two-class case. For this specific case we obtain

$$l_1 = \lambda_{11} p(\boldsymbol{x}|\omega_1)P(\omega_1) + \lambda_{21} p(\boldsymbol{x}|\omega_2)P(\omega_2)$$
$$l_2 = \lambda_{12} p(\boldsymbol{x}|\omega_1)P(\omega_1) + \lambda_{22} p(\boldsymbol{x}|\omega_2)P(\omega_2) \tag{2.18}$$

We assign $\boldsymbol{x}$ to ω_1 if $l_1 < l_2$, that is,

$$(\lambda_{21} - \lambda_{22})p(\boldsymbol{x}|\omega_2)P(\omega_2) < (\lambda_{12} - \lambda_{11})p(\boldsymbol{x}|\omega_1)P(\omega_1) \tag{2.19}$$

It is natural to assume that $\lambda_{ij} > \lambda_{ii}$ (correct decisions are penalized much less than wrong ones). Adopting this assumption, the decision rule (2.17) for the two-class case now becomes

$$\boldsymbol{x} \in \omega_1(\omega_2) \quad \text{if} \quad l_{12} \equiv \frac{p(\boldsymbol{x}|\omega_1)}{p(\boldsymbol{x}|\omega_2)} > (<) \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{21} - \lambda_{22}}{\lambda_{12} - \lambda_{11}} \tag{2.20}$$

The ratio l_{12} is known as the ***likelihood ratio*** and the preceding test as the ***likelihood ratio test***. Let us now investigate Eq. (2.20) a little further and consider the case of Figure 2.1. Assume that the loss matrix is of the form

$$L = \begin{bmatrix} 0 & \lambda_{12} \\ \lambda_{21} & 0 \end{bmatrix}$$

If misclassification of patterns that come from ω_2 is considered to have serious consequences, then we must choose $\lambda_{21} > \lambda_{12}$. Thus, patterns are assigned to class ω_2 if

$$p(\boldsymbol{x}|\omega_2) > p(\boldsymbol{x}|\omega_1)\frac{\lambda_{12}}{\lambda_{21}}$$

where $P(\omega_1) = P(\omega_2) = 1/2$ has been assumed. That is, $p(\boldsymbol{x}|\omega_1)$ is multiplied by a factor less than 1 and the effect of this is to move the threshold in Figure 2.1 to the left of x_0. In other words, region R_2 is increased while R_1 is decreased. The opposite would be true if $\lambda_{21} < \lambda_{12}$.

An alternative cost that sometimes is used for two class problems is the Neyman-Pearson criterion. The error for one of the classes is now constrained to be fixed and equal to a chosen value (Problem 2.6). Such a decision rule has been used, for example, in radar detection problems. The task there is to detect a target in the presence of noise. One type of error is the so-called *false alarm*—that is, to mistake the noise for a signal (target) present. Of course, the other type of error is to miss the signal and to decide in favor of the noise (***missed detection***). In many cases the error probability of false alarm is set equal to a predetermined threshold.

Example 2.1

In a two-class problem with a single feature x the pdfs are Gaussians with variance $\sigma^2 = 1/2$ for both classes and mean values 0 and 1, respectively, that is,

$\sigma^2 = -$

$$p(x|\omega_1) = \frac{1}{\sqrt{\pi}}\exp(-x^2)$$

$$p(x|\omega_2) = \frac{1}{\sqrt{\pi}}\exp(-(x-1)^2)$$

If $P(\omega_1) = P(\omega_2) = 1/2$, compute the threshold value x_0 (a) for minimum error probability and (b) for minimum risk if the loss matrix is

$$L = \begin{bmatrix} 0 & 0.5 \\ 1.0 & 0 \end{bmatrix}$$

Taking into account the shape of the Gaussian function graph (Appendix A), the threshold for the minimum probability case will be

$$x_0 : \exp(-x^2) = \exp(-(x-1)^2)$$

Taking the logarithm of both sides, we end up with $x_0 = 1/2$. In the minimum risk case we get

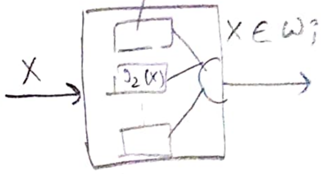
$$x_0 : \exp(-x^2) = 2\exp(-(x-1)^2)$$

or $x_0 = (1 - \ln 2)/2 < 1/2$; that is, the threshold moves to the left of $1/2$. If the two classes are not equiprobable, then it is easily verified that if $P(\omega_1) > (<) P(\omega_2)$ the threshold moves to the right (left). That is, we expand the region in which we decide in favor of the most probable class, since it is better to make fewer errors for the most probable class.

$\underbrace{P(\omega_i|X)}_{\lambda_{ii}}$

(functional units)

$g_1(x)$



Classifier (black box)

$g_1(x), g_2(x) \ldots g_n(x)$ Discriminant function.

2.3 DISCRIMINANT FUNCTIONS AND DECISION SURFACES

It is by now clear that minimizing either the risk or the error probability or the Neyman-Pearson criterion is equivalent to partitioning the feature space into M regions, for a task with M classes. If regions R_i, R_j happen to be contiguous, then they are separated by a *decision surface* in the multidimensional feature space. For the minimum error probability case, this is described by the equation

$$P(\omega_i|x) - P(\omega_j|x) = 0 \tag{2.21}$$

From the one side of the surface this difference is positive, and from the other it is negative. Sometimes, instead of working directly with probabilities (or risk functions), it may be more convenient, from a mathematical point of view, to work with an equivalent function of them, for example, $g_i(x) \equiv f(P(\omega_i|x))$, where $f(\cdot)$ is a monotonically increasing function. $g_i(x)$ is known as a *discriminant function*. The decision test (2.13) is now stated as

$g_i(x) \equiv f(P(\omega_i|x))$

$f(\cdot) \to$ monotonically inc. func

$$\text{classify } x \text{ in } \omega_i \quad \text{if} \quad g_i(x) > g_j(x) \quad \forall j \neq i \tag{2.22}$$

$g_i(x) > g_j(x)$ $\forall j \neq i$ $X \in \omega_i$

The decision surfaces, separating contiguous regions, are described by

$$g_{ij}(x) \equiv g_i(x) - g_j(x) = 0, \quad i, j = 1, 2, \ldots, M, \quad i \neq j \tag{2.23}$$

For minimum risk classifier :- $R(\alpha_i/X)$

$g_i(X) = -R(\alpha_i/X)$

Min error rate classification

$g_i(X) = P(\omega_i/X)$