CHAPTER 2

BAYESIAN DECISION THEORY

2.1 INTRODUCTION

Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. This approach is based on quantifying the tradeoffs between various classification decisions using probability and the costs that accompany such decisions. It makes the assumption that the decision problem is posed in probabilistic terms, and that all of the relevant probability values are known. In this chapter we develop the fundamentals of this theory and we show how it can be viewed as being simply a formalization of common-sense procedures; in subsequent chapters we will consider the problems that arise when the probabilistic structure is not completely known.

While we will give a quite general, abstract development of Bayesian decision theory in Section 2.2, we begin our discussion with a specific example. Let us reconsider the hypothetical problem posed in Chapter 1 of designing a classifier to separate two kinds of fish: sea bass and salmon. Suppose that an observer watching fish arrive along the conveyor belt finds it hard to predict what type will emerge next and that the sequence of types of fish appears to be random. In decision-theoretic terminology we would say that as each fish emerges nature is in one or the other of the two possible states: Either the fish is a sea bass or the fish is a salmon. We let ω denote the *state of nature*, with $\omega = \omega_1$ for sea bass and $\omega = \omega_2$ for salmon. Because the state of nature is so unpredictable, we consider ω to be a variable that must be described probabilistically.

STATE OF NATURE

If the catch produced as much sea bass as salmon, we would say that the next fish is equally likely to be sea bass or salmon. More generally, we assume that there is some *a priori probability* (or simply *prior*) $P(\omega_1)$ that the next fish is sea bass, and some prior probability $P(\omega_2)$ that it is salmon. If we assume there are no other types of fish relevant here, then $P(\omega_1)$ and $P(\omega_2)$ sum to one. These prior probabilities reflect our prior knowledge of how likely we are to get a sea bass or salmon before the fish actually appears. It might, for instance, depend upon the time of year or the choice of fishing area.

PRIOR

Suppose for a moment that we were forced to make a decision about the type of fish that will appear next without being allowed to see it. For the moment, we shall assume that any incorrect classification entails the same cost or consequence, and that the only information we are allowed to use is the value of the prior probabilities. If a decision must be made with so little information, it seems logical to use the following *decision rule*: Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2.

DECISION RULE

This rule makes sense if we are to judge just one fish, but if we are to judge many fish, using this rule repeatedly may seem a bit strange. After all, we would always make the same decision even though we know that *both* types of fish will appear. How well it works depends upon the values of the prior probabilities. If $P(\omega_1)$ is very much greater than $P(\omega_2)$, our decision in favor of ω_1 will be right most of the time. If $P(\omega_1) = P(\omega_2)$, we have only a fifty-fifty chance of being right. In general, the probability of error is the smaller of $P(\omega_1)$ and $P(\omega_2)$, and we shall see later that under these conditions no other decision rule can yield a larger probability of being right.

In most circumstances we are not asked to make decisions with so little information. In our example, we might for instance use a lightness measurement x to improve our classifier. Different fish will yield different lightness readings, and we express this variability in probabilistic terms; we consider x to be a continuous random variable whose distribution depends on the state of nature and is expressed as $p(x|\omega)$.* This is the *class-conditional probability density* function, the probability density function for x given that the state of nature is ω. (It is also sometimes called state-conditional probability density.) Then the difference between $p(x|\omega_1)$ and $p(x|\omega_2)$ describes the difference in lightness between populations of sea bass and salmon (Fig. 2.1).†

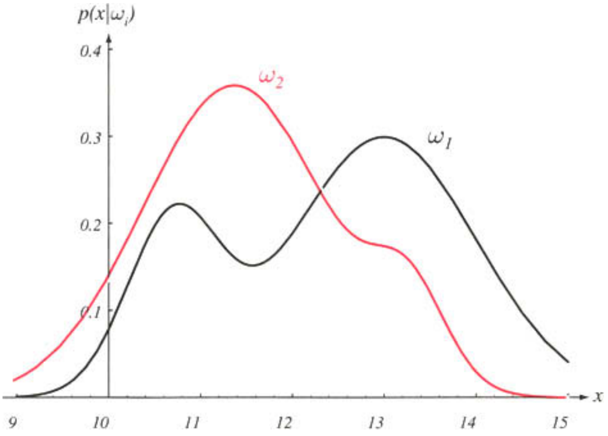


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0.

*We generally use an uppercase $P(\cdot)$ to denote a probability mass function and use a lowercase $p(\cdot)$ to denote a probability density function.

†Strictly speaking, the probability density function $p(x|\omega)$ should be written as $p_x(x|\omega)$ to indicate that we are speaking about a particular density function for the random variable X. This more elaborate subscripted notation makes it clear that $p_x(\cdot)$ and $p_y(\cdot)$ denote two different functions, a fact that is obscured when writing $p(x)$ and $p(y)$. Because this potential confusion rarely arises in practice, we have elected to adopt the simpler notation. Readers who are unsure of our notation or who would like to review probability theory should see Section A.4 of the Appendix.

Suppose that we know both the prior probabilities $P(\omega_j)$ and the conditional densities $p(x|\omega_j)$ for $j = 1, 2$. Suppose further that we measure the lightness of a fish and discover that its value is x. How does this measurement influence our attitude concerning the true state of nature—that is, the category of the fish? We note first that the (joint) probability density of finding a pattern that is in category ω_j *and* has feature value x can be written in two ways: $p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j)$. Rearranging these leads us to the answer to our question, which is called *Bayes formula*:

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}, \tag{1}$$

where in this case of two categories

$$p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j). \tag{2}$$

Bayes formula can be expressed informally in English by saying that

$$\mathit{posterior} = \frac{\mathit{likelihood} \times \mathit{prior}}{\mathit{evidence}}. \tag{3}$$

POSTERIOR

LIKELIHOOD

EVIDENCE

Bayes formula shows that by observing the value of x we can convert the prior probability $P(\omega_j)$ to the *a posteriori* probability (or *posterior*) $P(\omega_j|x)$—the probability of the state of nature being ω_j given that feature value x has been measured. We call $p(x|\omega_j)$ the *likelihood* of ω_j with respect to x, a term chosen to indicate that, other things being equal, the category ω_j for which $p(x|\omega_j)$ is large is more "likely" to be the true category. Notice that it is the product of the likelihood and the prior probability that is most important in determining the posterior probability; the *evidence* factor, $p(x)$, can be viewed as merely a scale factor that guarantees that the posterior probabilities sum to one, as all good probabilities must. The variation of $P(\omega_j|x)$ with x is illustrated in Fig. 2.2 for the case $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$.

If we have an observation x for which $P(\omega_1|x)$ is greater than $P(\omega_2|x)$, we would naturally be inclined to decide that the true state of nature is ω_1. Conversely, if $P(\omega_2|x)$ is greater than $P(\omega_1|x)$, we would be inclined to choose ω_2. To justify this decision procedure, let us calculate the probability of error whenever we make a decision. Whenever we observe a particular x, the probability of error is

$$P(\mathit{error}|x) = \begin{cases} P(\omega_1|x) & \text{if we decide } \omega_2 \\ P(\omega_2|x) & \text{if we decide } \omega_1. \end{cases} \tag{4}$$

Clearly, for a given x we can minimize the probability of error by deciding ω_1 if $P(\omega_1|x) > P(\omega_2|x)$ and ω_2 otherwise. Of course, we may never observe exactly the same value of x twice. Will this rule minimize the average probability of error? Yes, because the average probability of error is given by

$$P(\mathit{error}) = \int_{-\infty}^{\infty} P(\mathit{error}, x)\, dx = \int_{-\infty}^{\infty} P(\mathit{error}|x)p(x)\, dx \tag{5}$$

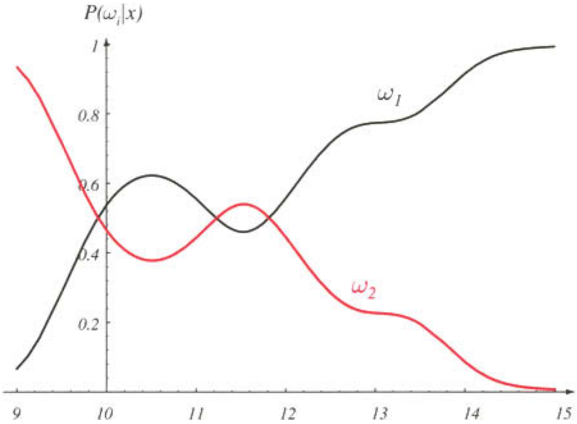


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0.

BAYES DECISION RULE

and if for every x we ensure that $P(error|x)$ is as small as possible, then the integral must be as small as possible. Thus we have justified the following *Bayes decision rule* for minimizing the probability of error:

$$\text{Decide } \omega_1 \text{ if } P(\omega_1|x) > P(\omega_2|x); \quad \text{otherwise decide } \omega_2. \tag{6}$$

Under this rule Eq. 4 becomes

$$P(error|x) = \min[P(\omega_1|x), P(\omega_2|x)]. \tag{7}$$

EVIDENCE

This form of the decision rule emphasizes the role of the posterior probabilities. By using Eq. 1 we can instead express the rule in terms of the conditional and prior probabilities. First note that the *evidence*, $p(x)$, in Eq. 1 is unimportant as far as making a decision is concerned. It is basically just a scale factor that states how frequently we will actually measure a pattern with feature value x; as mentioned above, its presence in Eq. 1 assures us that $P(\omega_1|x) + P(\omega_2|x) = 1$. By eliminating this scale factor, we obtain the following completely equivalent decision rule:

$$\text{Decide } \omega_1 \text{ if } p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2); \quad \text{otherwise decide } \omega_2. \tag{8}$$

Some additional insight can be obtained by considering a few special cases. If for some x we have $p(x|\omega_1) = p(x|\omega_2)$, then that particular observation gives us no information about the state of nature; in this case, the decision hinges entirely on the prior probabilities. On the other hand, if $P(\omega_1) = P(\omega_2)$, then the states of nature are equally probable; in this case the decision is based entirely on the likelihoods $p(x|\omega_j)$. In general, both of these factors are important in making a decision, and the Bayes decision rule combines them to achieve the minimum probability of error.

2.2 BAYESIAN DECISION THEORY—CONTINUOUS FEATURES

We shall now formalize the ideas just considered, and generalize them in four ways:

- By allowing the use of more than one feature
- By allowing more than two states of nature
- By allowing actions other than merely deciding the state of nature
- By introducing a loss function more general than the probability of error

FEATURE SPACE

LOSS FUNCTION

These generalizations and their attendant notational complexities should not obscure the central points illustrated in our simple example. Allowing the use of more than one feature merely requires replacing the scalar x by the *feature vector* $\mathbf{x}$, where $\mathbf{x}$ is in a d-dimensional Euclidean space $\mathbf{R}^d$, called the *feature space*. Allowing more than two states of nature provides us with a useful generalization for a small notational expense. Allowing actions other than classification primarily allows the possibility of rejection—that is, of refusing to make a decision in close cases; this is a useful option if being indecisive is not too costly. Formally, the *loss function* states exactly how costly each action is, and is used to convert a probability determination into a decision. Cost functions let us treat situations in which some kinds of classification mistakes are more costly than others, although we often discuss the simplest case, where all errors are equally costly. With this as a preamble, let us begin the more formal treatment.

Let $\{\omega_1, \ldots, \omega_c\}$ be the finite set of c states of nature ("categories") and let $\{\alpha_1, \ldots, \alpha_a\}$ be the finite set of a possible actions. The loss function $\lambda(\alpha_i|\omega_j)$ describes the loss incurred for taking action α_i when the state of nature is ω_j. Let the feature vector $\mathbf{x}$ be a d-component vector-valued random variable and let $p(\mathbf{x}|\omega_j)$ be the state-conditional probability density function for $\mathbf{x}$, with the probability density function for $\mathbf{x}$ conditioned on ω_j being the true state of nature. As before, $P(\omega_j)$ describes the prior probability that nature is in state ω_j. Then the posterior probability $P(\omega_j|\mathbf{x})$ can be computed from $p(\mathbf{x}|\omega_j)$ by Bayes formula:

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})}, \tag{9}$$

where the evidence is now

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j). \tag{10}$$

Suppose that we observe a particular $\mathbf{x}$ and that we contemplate taking action α_i. If the true state of nature is ω_j, by definition we will incur the loss $\lambda(\alpha_i|\omega_j)$. Because $P(\omega_j|\mathbf{x})$ is the probability that the true state of nature is ω_j, the expected loss associated with taking action α_i is merely

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}). \tag{11}$$

RISK

In decision-theoretic terminology, an expected loss is called a *risk*, and $R(\alpha_i|\mathbf{x})$ is called the *conditional risk*. Whenever we encounter a particular observation $\mathbf{x}$, we can minimize our expected loss by selecting the action that minimizes the conditional risk. We shall now show that this *Bayes decision procedure* actually provides the optimal performance.

DECISION RULE

Stated formally, our problem is to find a decision rule against $P(\omega_j)$ that minimizes the overall risk. A general *decision rule* is a function $\alpha(\mathbf{x})$ that tells us which action to take for every possible observation. To be more specific, for every $\mathbf{x}$ the *decision function* $\alpha(\mathbf{x})$ assumes one of the a values $\alpha_1, \ldots, \alpha_a$. The overall risk R is the expected loss associated with a given decision rule. Because $R(\alpha_i|\mathbf{x})$ is the conditional risk associated with action α_i and because the decision rule specifies the action, the overall risk is given by

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})\, d\mathbf{x}, \qquad (12)$$

where $d\mathbf{x}$ is our notation for a d-space volume element and where the integral extends over the entire feature space. Clearly, if $\alpha(\mathbf{x})$ is chosen so that $R(\alpha_i(\mathbf{x}))$ is as small as possible for every $\mathbf{x}$, then the overall risk will be minimized. This justifies the following statement of the *Bayes decision rule*: To minimize the overall risk, compute the conditional risk

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \qquad (13)$$

BAYES RISK

for $i = 1, \ldots, a$ and then select the action α_i for which $R(\alpha_i|\mathbf{x})$ is minimum.* The resulting minimum overall risk is called the *Bayes risk*, denoted R^*, and is the best performance that can be achieved.

2.2.1 Two-Category Classification

Let us consider these results when applied to the special case of two-category classification problems. Here action α_1 corresponds to deciding that the true state of nature is ω_1, and action α_2 corresponds to deciding that it is ω_2. For notational simplicity, let $\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ be the loss incurred for deciding ω_i when the true state of nature is ω_j. If we write out the conditional risk given by Eq. 13, we obtain

$$R(\alpha_1|\mathbf{x}) = \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x}) \qquad (14)$$

$$R(\alpha_2|\mathbf{x}) = \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x}). \qquad (15)$$

There are a variety of ways of expressing the minimum-risk decision rule, each having its own minor advantages. The fundamental rule is to decide ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$. In terms of the posterior probabilities, we decide ω_1 if

$$(\lambda_{21} - \lambda_{11})P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2|\mathbf{x}). \qquad (16)$$

Ordinarily, the loss incurred for making an error is greater than the loss incurred for being correct, and both of the factors $\lambda_{21} - \lambda_{11}$ and $\lambda_{12} - \lambda_{22}$ are positive. Thus in practice, our decision is generally determined by the more likely state of nature, although

*Note that if more than one action minimizes $R(\alpha|\mathbf{x})$, it does not matter which of these actions is taken, and any convenient tie-breaking rule can be used.

we must scale the posterior probabilities by the loss differences. By employing Bayes formula, we can replace the posterior probabilities by the prior probabilities and the conditional densities. This results in the equivalent rule, to decide ω_1 if

$$(\lambda_{21} - \lambda_{11})p(\mathbf{x}|\omega_1)P(\omega_1) > (\lambda_{12} - \lambda_{22})p(\mathbf{x}|\omega_2)P(\omega_2), \tag{17}$$

and otherwise decide ω_2.

Another alternative, which follows at once under the reasonable assumption that $\lambda_{21} > \lambda_{11}$, is to decide ω_1 if

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}. \tag{18}$$

LIKELIHOOD RATIO

This form of the decision rule focuses on the $\mathbf{x}$-dependence of the probability densities. We can consider $p(\mathbf{x}|\omega_j)$ a function of ω_j (i.e., the likelihood function) and then form the *likelihood ratio* $p(\mathbf{x}|\omega_1)/p(\mathbf{x}|\omega_2)$. Thus the Bayes decision rule can be interpreted as calling for deciding ω_1 if the likelihood ratio exceeds a threshold value that is independent of the observation $\mathbf{x}$.

2.3 MINIMUM-ERROR-RATE CLASSIFICATION

In classification problems, each state of nature is usually associated with a different one of the c classes, and the action α_i is usually interpreted as the decision that the true state of nature is ω_i. If action α_i is taken and the true state of nature is ω_j, then the decision is correct if $i = j$ and in error if $i \neq j$. If errors are to be avoided, it is natural to seek a decision rule that minimizes the probability of error, that is, the *error rate*.

ZERO-ONE LOSS

The loss function of interest for this case is hence the so-called *symmetrical* or *zero-one* loss function,

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c. \tag{19}$$

This loss function assigns no loss to a correct decision, and assigns a unit loss to any error; thus, all errors are equally costly.* The risk corresponding to this loss function is precisely the average probability of error because the conditional risk is

$$\begin{aligned} R(\alpha_i|\mathbf{x}) &= \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \\ &= \sum_{j\neq i} P(\omega_j|\mathbf{x}) \\ &= 1 - P(\omega_i|\mathbf{x}) \end{aligned} \tag{20}$$

*We note that other loss functions, such as quadratic and linear difference, find greater use in regression tasks where there is a natural ordering on the predictions and we can meaningfully penalize predictions that are "more wrong" than others.

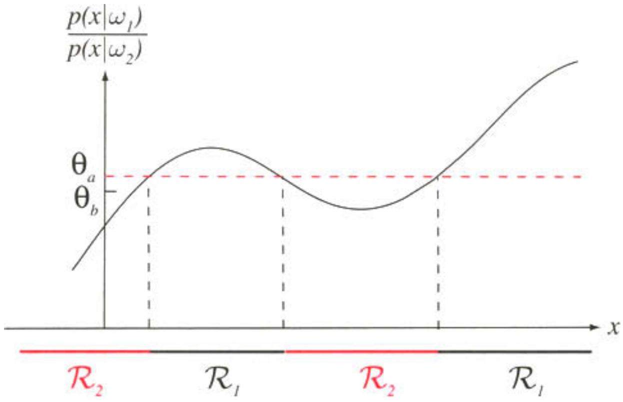


FIGURE 2.3. The likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the distributions shown in Fig. 2.1. If we employ a zero-one or classification loss, our decision boundaries are determined by the threshold θ_a. If our loss function penalizes miscategorizing ω_2 as ω_1 patterns more than the converse, we get the smaller threshold θ_b, and hence $\mathcal{R}_1$ becomes smaller.

and $P(\omega_i|\mathbf{x})$ is the conditional probability that action α_i is correct. The Bayes decision rule to minimize risk calls for selecting the action that minimizes the conditional risk. Thus, to minimize the average probability of error, we should select the i that *maximizes* the posterior probability $P(\omega_i|\mathbf{x})$. In other words, for *minimum error rate*:

$$\text{Decide } \omega_i \text{ if } P(\omega_i|\mathbf{x}) > P(\omega_j|\mathbf{x}) \qquad \text{for all } j \neq i. \tag{21}$$

This is the same rule as in Eq. 6. The region in the input space where we decide ω_i is denoted $\mathcal{R}_i$; such a region need not be simply connected.

We saw in Fig. 2.2 some class-conditional probability densities and the posterior probabilities; Fig. 2.3 shows the likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the same case. In general, this ratio can range between zero and infinity. The threshold value θ_a marked is from the same prior probabilities but with a zero-one loss function. Notice that this leads to the same decision boundaries as in Fig. 2.2, as it must. If we penalize mistakes in classifying ω_1 patterns as ω_2 more than the converse (i.e., $\lambda_{21} > \lambda_{12}$), then Eq. 18 leads to the threshold θ_b marked. Note that the range of x values for which we classify a pattern as ω_1 gets larger, as it should.

*2.3.1 Minimax Criterion

Sometimes we must design our classifier to perform well over a *range* of prior probabilities. For instance, in our fish categorization problem we can imagine that whereas the physical properties of lightness and width of each type of fish remain constant, the prior probabilities might vary widely and in an unpredictable way, or alternatively we want to use the classifier in a different plant where we do not know the prior probabilities. A reasonable approach is then to design our classifier so that the *worst* overall risk for any value of the priors is as small as possible—that is, minimize the maximum possible overall risk.

In order to understand this, we let $\mathcal{R}_1$ denote that (as yet unknown) region in feature space where the classifier decides ω_1 and likewise for $\mathcal{R}_2$ and ω_2, and then we write our overall risk Eq. 12 in terms of conditional risks:

$$
\begin{aligned}
R = &\int_{\mathcal{R}_1} [\lambda_{11}P(\omega_1)\,p(\mathbf{x}|\omega_1) + \lambda_{12}P(\omega_2)\,p(\mathbf{x}|\omega_2)]\,d\mathbf{x} \\
&+ \int_{\mathcal{R}_2} [\lambda_{21}P(\omega_1)\,p(\mathbf{x}|\omega_1) + \lambda_{22}P(\omega_2)\,p(\mathbf{x}|\omega_2)]\,d\mathbf{x}.
\end{aligned} \tag{22}
$$

We use the fact that $P(\omega_2) = 1 - P(\omega_1)$ and that $\int_{\mathcal{R}_1} p(\mathbf{x}|\omega_1)\,d\mathbf{x} = 1 - \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)\,d\mathbf{x}$ to rewrite the risk as:

$$
\begin{aligned}
R(P(\omega_1)) = &\overbrace{\lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)\,d\mathbf{x}}^{=R_{mm},\ \text{minimax risk}} \\
&+ P(\omega_1) \underbrace{\left[(\lambda_{11} - \lambda_{22}) + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)\,d\mathbf{x} - (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)\,d\mathbf{x} \right]}_{=0 \text{ for minimax solution}}.
\end{aligned} \tag{23}
$$

This equation shows that once the decision boundary is set (i.e., $\mathcal{R}_1$ and $\mathcal{R}_2$ determined), the overall risk is linear in $P(\omega_1)$. If we can find a boundary such that the constant of proportionality is 0, then the risk is independent of priors. This is the *minimax solution*, and the *minimax risk*, R_{mm}, can be read from Eq. 23:

MINIMAX RISK

$$
\begin{aligned}
R_{mm} &= \lambda_{22} + (\lambda_{12} - \lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)\,d\mathbf{x} \\
&= \lambda_{11} + (\lambda_{21} - \lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)\,d\mathbf{x}.
\end{aligned} \tag{24}
$$

Figure 2.4 illustrates the approach. Briefly stated, we search for the prior for which the Bayes risk is *maximum*, and the corresponding decision boundary then gives the minimax solution. The value of the minimax risk, R_{mm}, is hence equal to the worst Bayes risk. In practice, finding the decision boundary for minimax risk may be difficult, particularly when distributions are complicated. Nevertheless, in some cases the boundary can be determined analytically (Problem 4).

The minimax criterion finds greater use in game theory than it does in traditional pattern recognition. In game theory you have a hostile opponent who can be expected to take an action maximally detrimental to you. Thus it makes great sense for you to take an action (e.g., make a classification) where your costs—due to your opponent's subsequent actions—are minimized.

*2.3.2 Neyman-Pearson Criterion

In some problems, we may wish to minimize the overall risk subject to a constraint; for instance, we might wish to minimize the total risk subject to the constraint $\int R(\alpha_i|\mathbf{x})\,d\mathbf{x} < constant$ for some particular i. Such a constraint might arise when there is a fixed resource that accompanies one particular action α_i, or when we must not misclassify a pattern from a particular state of nature ω_i at more than some limited frequency. For instance, in our fish example, there might be some government

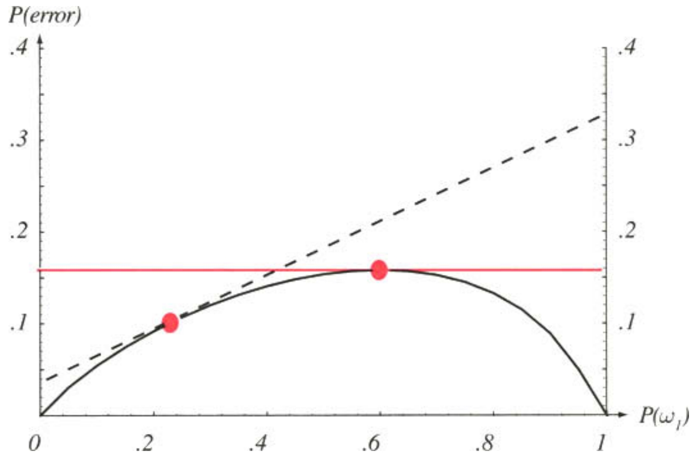


FIGURE 2.4. The curve at the bottom shows the minimum (Bayes) error as a function of prior probability $P(\omega_1)$ in a two-category classification problem of fixed distributions. For each value of the priors (e.g., $P(\omega_1) = 0.25$) there is a corresponding optimal decision boundary and associated Bayes error rate. For any (fixed) such boundary, if the priors are then changed, the probability of error will change as a linear function of $P(\omega_1)$ (shown by the dashed line). The maximum such error will occur at an extreme value of the prior, here at $P(\omega_1) = 1$. To minimize the maximum of such error, we should design our decision boundary for the maximum Bayes error (here $P(\omega_1) = 0.6$), and thus the error will not change as a function of prior, as shown by the solid red horizontal line.

regulation that we must not misclassify more than 1% of salmon as sea bass. We might then seek a decision that minimizes the chance of classifying a sea bass as a salmon subject to this condition.

We generally satisfy such a *Neyman-Pearson criterion* by adjusting decision boundaries numerically. However, for Gaussian and some other distributions, Neyman-Pearson solutions can be found analytically (Problems 6 and 7). We shall have cause to mention Neyman-Pearson criteria again in Section 2.8.3 on operating characteristics.

2.4 CLASSIFIERS, DISCRIMINANT FUNCTIONS, AND DECISION SURFACES

2.4.1 The Multicategory Case

There are many different ways to represent pattern classifiers. One of the most useful is in terms of a set of *discriminant functions* $g_i(\mathbf{x}), i = 1, \ldots, c$. The classifier is said to assign a feature vector $\mathbf{x}$ to class ω_i if

$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \qquad \text{for all } j \neq i. \tag{25}$$

Thus, the classifier is viewed as a network or machine that computes c discriminant functions and selects the category corresponding to the largest discriminant. A network representation of a classifier is illustrated in Fig. 2.5.

A Bayes classifier is easily and naturally represented in this way. For the general case with risks, we can let $g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x})$, because the maximum discriminant

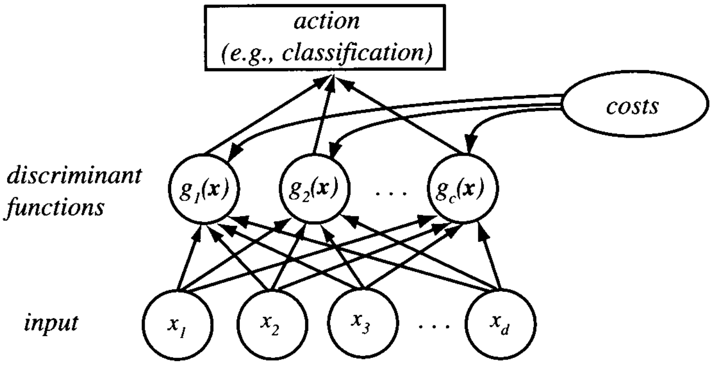


FIGURE 2.5. The functional structure of a general statistical pattern classifier which includes d inputs and c discriminant functions $g_i(\mathbf{x})$. A subsequent step determines which of the discriminant values is the maximum, and categorizes the input pattern accordingly. The arrows show the direction of the flow of information, though frequently the arrows are omitted when the direction of flow is self-evident.

function will then correspond to the minimum conditional risk. For the minimum-error-rate case, we can simplify things further by taking $g_i(\mathbf{x}) = P(\omega_i|\mathbf{x})$, so that the maximum discriminant function corresponds to the maximum posterior probability.

Clearly, the choice of discriminant functions is not unique. We can always multiply all the discriminant functions by the same positive constant or shift them by the same additive constant without influencing the decision. More generally, if we replace every $g_i(\mathbf{x})$ by $f(g_i(\mathbf{x}))$, where $f(\cdot)$ is a monotonically increasing function, the resulting classification is unchanged. This observation can lead to significant analytical and computational simplifications. In particular, for minimum-error-rate classification, any of the following choices gives identical classification results, but some can be much simpler to understand or to compute than others:

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j)} \tag{26}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i) \tag{27}$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i), \tag{28}$$

where ln denotes natural logarithm.

DECISION REGION

Even though the discriminant functions can be written in a variety of forms, the decision rules are equivalent. The effect of any decision rule is to divide the feature space into c *decision regions*, $\mathcal{R}_1, \ldots, \mathcal{R}_c$. If $g_i(\mathbf{x}) > g_j(\mathbf{x})$ for all $j \neq i$, then $\mathbf{x}$ is in $\mathcal{R}_i$, and the decision rule calls for us to assign $\mathbf{x}$ to ω_i. The regions are separated by *decision boundaries*, surfaces in feature space where ties occur among the largest discriminant functions (Fig. 2.6).

2.4.2 The Two-Category Case

DICHOTOMIZER

While the two-category case is just a special instance of the multicategory case, it has traditionally received separate treatment. Indeed, a classifier that places a pattern in one of only two categories has a special name—a *dichotomizer*.* Instead of using

*A classifier for more than two categories is called a polychotomizer.

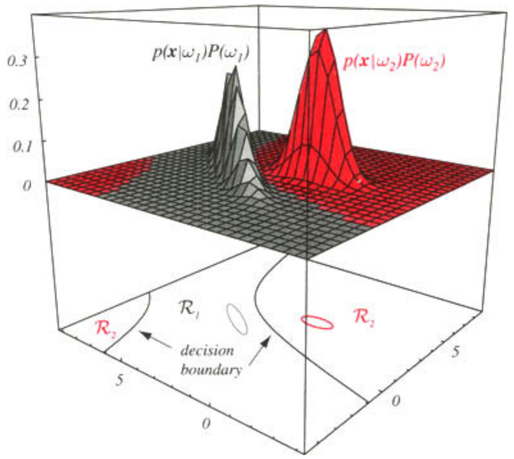


FIGURE 2.6. In this two-dimensional two-category classifier, the probability densities are Gaussian, the decision boundary consists of two hyperbolas, and thus the decision region $\mathcal{R}_2$ is not simply connected. The ellipses mark where the density is $1/e$ times that at the peak of the distribution.

two discriminant functions g_1 and g_2 and assigning $\mathbf{x}$ to ω_1 if $g_1 > g_2$, it is more common to define a single discriminant function

$$g(\mathbf{x}) \equiv g_1(\mathbf{x}) - g_2(\mathbf{x}), \tag{29}$$

and to use the following decision rule: Decide ω_1 if $g(\mathbf{x}) > 0$; otherwise decide ω_2. Thus, a dichotomizer can be viewed as a machine that computes a single discriminant function $g(\mathbf{x})$, and classifies $\mathbf{x}$ according to the algebraic sign of the result. Of the various forms in which the minimum-error-rate discriminant function can be written, the following two (derived from Eqs. 26 and 28) are particularly convenient:

$$g(\mathbf{x}) = P(\omega_1|\mathbf{x}) - P(\omega_2|\mathbf{x}) \tag{30}$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)}. \tag{31}$$

2.5 THE NORMAL DENSITY

The structure of a Bayes classifier is determined by the conditional densities $p(\mathbf{x}|\omega_i)$ as well as by the prior probabilities $P(\omega_i)$. Of the various density functions that have been investigated, none has received more attention than the multivariate normal or Gaussian density. To a large extent this attention is due to its analytical tractability. However, the multivariate normal density is also an appropriate model for an important situation, namely, the case where the feature vectors $\mathbf{x}$ for a given class ω_i are continuous-valued, randomly corrupted versions of a single typical or prototype vector $\boldsymbol{\mu}_i$. In this section we provide a brief exposition of the multivariate normal density, focusing on the properties of greatest interest for classification problems.

EXPECTATION

First, recall the definition of the *expected value* of a scalar function $f(x)$, defined for some density $p(x)$:

$$\mathcal{E}[f(x)] \equiv \int_{-\infty}^{\infty} f(x)p(x)\,dx. \tag{32}$$

If the values of the feature x are restricted to points in a discrete set $\mathcal{D}$, we must sum over all samples as

$$\mathcal{E}[f(x)] = \sum_{x \in \mathcal{D}} f(x)P(x), \tag{33}$$

where $P(x)$ is the probability mass at x. We shall occassionally need to calculate expected values by these and analogous equations defined in higher dimensions (see Appendix Sections A.4.2, A.4.5 and A.4.9).*

2.5.1 Univariate Density

We begin with the continuous univariate normal or Gaussian density,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right], \tag{34}$$

for which the *expected value* of x (an average, here taken over the feature space) is

$$\mu \equiv \mathcal{E}[x] = \int_{-\infty}^{\infty} xp(x)\,dx, \tag{35}$$

VARIANCE

and where the expected squared deviation or *variance* is

$$\sigma^2 \equiv \mathcal{E}[(x-\mu)^2] = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,dx. \tag{36}$$

MEAN

The univariate normal density is completely specified by two parameters: its mean μ and variance σ^2. For simplicity, we often abbreviate Eq. 34 by writing $p(x) \sim N(\mu, \sigma^2)$ to say that x is distributed normally with mean μ and variance σ^2. Samples from normal distributions tend to cluster about the mean, with a spread related to the standard deviation σ (Fig. 2.7).

ENTROPY

There is a deep relationship between the normal distribution and *entropy*. We discuss entropy in greater detail in Appendix Section A.7, but for now we merely state that the entropy of a distribution is given by

$$H(p(x)) = -\int p(x)\,\ln p(x)\,dx, \tag{37}$$

NAT
BIT

and measured in *nats*; if a $\log_2$ is used instead, the unit is the *bit*. The entropy measures the fundamental uncertainty in the values of points selected randomly from a

*We will often use somewhat loose engineering terminology and refer to a single point as a "sample." Statisticians, however, always refer to a sample as a *collection* of points, and they discuss "a sample of size n." When taken in context, there are rarely ambiguities in such usage.

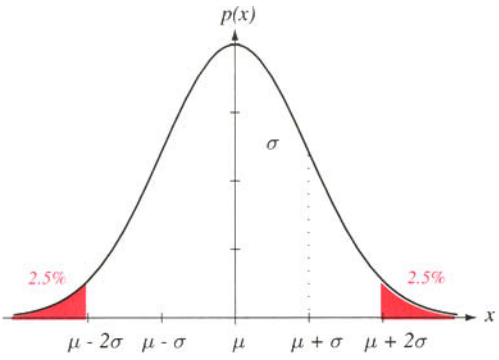


FIGURE 2.7. A univariate normal distribution has roughly 95% of its area in the range $|x - \mu| \leq 2\sigma$, as shown. The peak of the distribution has value $p(\mu) = 1/\sqrt{2\pi}\sigma$.

CENTRAL LIMIT THEOREM

distribution. It can be shown that the normal distribution has the maximum entropy of all distributions having a given mean and variance (Problem 20). Moreover, as stated by the *Central Limit Theorem*, the aggregate effect of the sum of a large number of small, independent random disturbances will lead to a Gaussian distribution (Computer exercise 5). Because many patterns—from fish to handwritten characters to some speech sounds—can be viewed as some ideal or prototype pattern corrupted by a large number of random processes, the Gaussian is often a good model for the actual probability distribution.

2.5.2 Multivariate Density

The general multivariate normal density in d dimensions is written as

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right], \tag{38}$$

COVARIANCE MATRIX

INNER PRODUCT

where $\mathbf{x}$ is a d-component column vector, $\boldsymbol{\mu}$ is the d-component *mean vector*, $\boldsymbol{\Sigma}$ is the d-by-d *covariance matrix*, and $|\boldsymbol{\Sigma}|$ and $\boldsymbol{\Sigma}^{-1}$ are its determinant and inverse, respectively. Further, we let $(\mathbf{x} - \boldsymbol{\mu})^t$ denote the transpose of $\mathbf{x} - \boldsymbol{\mu}$.* Our notation for the *inner product* is

$$\mathbf{a}^t\mathbf{b} = \sum_{i=1}^{d} a_i b_i, \tag{39}$$

which is often called a *dot product*. For simplicity, we often abbreviate Eq. 38 as $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Formally, we have

$$\boldsymbol{\mu} \equiv \mathcal{E}[\mathbf{x}] = \int \mathbf{x} p(\mathbf{x})\, d\mathbf{x} \tag{40}$$

*The mathematical expressions for the multivariate normal density are greatly simplified by employing the concepts and notation of linear algebra. Readers who are unsure of our notation or who would like to review linear algebra should see Appendix Section A.2.

and

$$\boldsymbol{\Sigma} \equiv \mathcal{E}[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t] = \int (\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^t p(\mathbf{x})\, d\mathbf{x}, \tag{41}$$

where the expected value of a vector or a matrix is found by taking the expected values of its components. In other words, if x_i is the ith component of $\mathbf{x}$, μ_i the ith component of $\boldsymbol{\mu}$, and σ_{ij} the ijth component of $\boldsymbol{\Sigma}$, then

$$\mu_i = \mathcal{E}[x_i] \tag{42}$$

and

$$\sigma_{ij} = \mathcal{E}[(x_i - \mu_i)(x_j - \mu_j)]. \tag{43}$$

COVARIANCE

STATISTICAL INDEPENDENCE

The covariance matrix $\boldsymbol{\Sigma}$ is always symmetric and positive semidefinite. We shall restrict our attention to the case in which $\boldsymbol{\Sigma}$ is positive definite, so that the determinant of $\boldsymbol{\Sigma}$ is strictly positive.* The diagonal elements σ_{ii} are the variances of the respective x_i (i.e., σ_i^2), and the off-diagonal elements σ_{ij} are the *covariances* of x_i and x_j. We would expect a positive covariance for the length and weight features of a population of fish, for instance. If x_i and x_j are *statistically independent*, then $\sigma_{ij} = 0$. If all the off-diagonal elements are zero, $p(\mathbf{x})$ reduces to the product of the univariate normal densities for the components of $\mathbf{x}$.

Linear combinations of jointly normally distributed random variables, independent or not, are normally distributed. In particular, if $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{A}$ is a d-by-k matrix and $\mathbf{y} = \mathbf{A}^t\mathbf{x}$ is a k-component vector, then $p(\mathbf{y}) \sim N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$, as illustrated in Fig. 2.8. In the special case where $k = 1$ and $\mathbf{A}$ is a unit-length vector $\mathbf{a}$, $y = \mathbf{a}^t\mathbf{x}$ is a scalar that represents the projection of $\mathbf{x}$ onto a line in the direction of $\mathbf{a}$; in that case $\mathbf{a}^t\boldsymbol{\Sigma}\mathbf{a}$ is the variance of the projection of $\mathbf{x}$ onto $\mathbf{a}$. In general then, knowledge of the covariance matrix allows us to calculate the dispersion of the data in any direction, or in any subspace.

It is sometimes convenient to perform a coordinate transformation that converts an arbitrary multivariate normal distribution into a spherical one—that is, one having a covariance matrix proportional to the identity matrix $\mathbf{I}$. If we define $\boldsymbol{\Phi}$ to be the matrix whose columns are the orthonormal eigenvectors of $\boldsymbol{\Sigma}$, and $\boldsymbol{\Lambda}$ the diagonal matrix of the corresponding eigenvalues, then the transformation

$$\mathbf{A}_w = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2} \tag{44}$$

WHITENING TRANSFORM

applied to the coordinates ensures that the transformed distribution has covariance matrix equal to the identity matrix. In signal processing, $\mathbf{A}_w$ yields a so-called *whitening* transform, because it makes the spectrum of eigenvalues of the transformed distribution uniform.

The multivariate normal density is completely specified by $d + d(d+1)/2$ parameters, namely the elements of the mean vector $\boldsymbol{\mu}$ and the independent elements of the covariance matrix $\boldsymbol{\Sigma}$. Samples drawn from a normal population tend to fall in a single cloud or cluster (Fig. 2.9); the center of the cluster is determined by the mean vector, and the shape of the cluster is determined by the covariance matrix. It

*If sample vectors are drawn from a linear subspace, $|\boldsymbol{\Sigma}| = 0$ and $p(\mathbf{x})$ is degenerate. This occurs, for example, when one component of $\mathbf{x}$ has zero variance, or when two components are identical or multiples of one another.

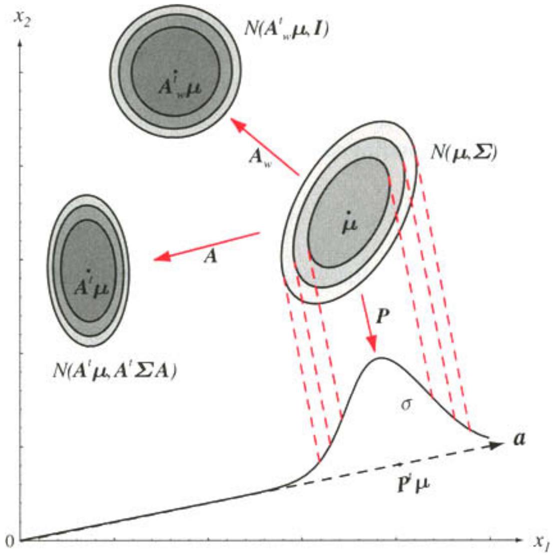


FIGURE 2.8. The action of a linear transformation on the feature space will convert an arbitrary normal distribution into another normal distribution. One transformation, **A**, takes the source distribution into distribution $N(\mathbf{A}^t\boldsymbol{\mu}, \mathbf{A}^t\boldsymbol{\Sigma}\mathbf{A})$. Another linear transformation—a projection **P** onto a line defined by vector **a**—leads to $N(\mu, \sigma^2)$ measured along that line. While the transforms yield distributions in a different space, we show them superimposed on the original x_1x_2-space. A whitening transform, $\mathbf{A}_w$, leads to a circularly symmetric Gaussian, here shown displaced.

follows from Eq. 38 that the loci of points of constant density are hyperellipsoids for which the quadratic form $(\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ is constant. The principal axes of these hyperellipsoids are given by the eigenvectors of $\boldsymbol{\Sigma}$ (described by $\boldsymbol{\Phi}$); the eigenvalues (described by $\boldsymbol{\Lambda}$) determine the lengths of these axes. The quantity

$$r^2 = (\mathbf{x}-\boldsymbol{\mu})^t\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \tag{45}$$

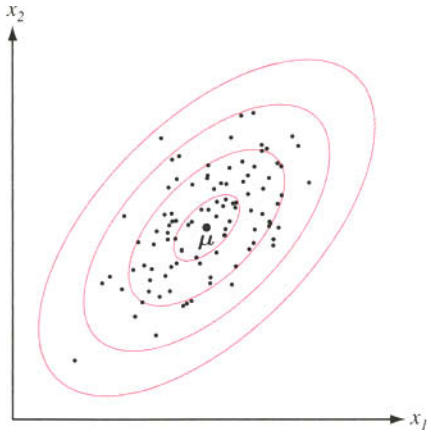


FIGURE 2.9. Samples drawn from a two-dimensional Gaussian lie in a cloud centered on the mean $\boldsymbol{\mu}$. The ellipses show lines of equal probability density of the Gaussian.

MAHALANOBIS DISTANCE

is sometimes called the squared *Mahalanobis distance* from $\mathbf{x}$ to $\boldsymbol{\mu}$. Thus, the contours of constant density are hyperellipsoids of constant Mahalanobis distance to $\boldsymbol{\mu}$ and the volume of these hyperellipsoids measures the scatter of the samples about the mean. It can be shown (Problems 15 and 16) that the volume of the hyperellipsoid corresponding to a Mahalanobis distance r is given by

$$V = V_d |\boldsymbol{\Sigma}|^{1/2} r^d, \tag{46}$$

where V_d is the volume of a d-dimensional unit hypersphere:

$$V_d = \begin{cases} \pi^{d/2}/(d/2)! & d \text{ even} \\ 2^d \pi^{(d-1)/2} (\frac{d-1}{2})!/d! & d \text{ odd.} \end{cases} \tag{47}$$

Thus, for a given dimensionality, the scatter of the samples varies directly with $|\boldsymbol{\Sigma}|^{1/2}$ (Problem 17).

2.6 DISCRIMINANT FUNCTIONS FOR THE NORMAL DENSITY

In Section 2.4.1 we saw that the minimum-error-rate classification can be achieved by use of the discriminant functions

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i). \tag{48}$$

This expression can be readily evaluated if the densities $p(\mathbf{x}|\omega_i)$ are multivariate normal—that is, if $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$. In this case, then, from Eq. 38 we have

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\boldsymbol{\Sigma}_i| + \ln P(\omega_i). \tag{49}$$

Let us examine this discriminant function and resulting classification for a number of special cases.

2.6.1 Case 1: $\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$

The simplest case occurs when the features are statistically independent and when each feature has the same variance, σ^2. In this case the covariance matrix is diagonal, being merely σ^2 times the identity matrix $\mathbf{I}$. Geometrically, this corresponds to the situation in which the samples fall in equal-size hyperspherical clusters, the cluster for the ith class being centered about the mean vector $\boldsymbol{\mu}_i$. The computation of the determinant and the inverse of $\boldsymbol{\Sigma}_i$ is particularly easy: $|\boldsymbol{\Sigma}_i| = \sigma^{2d}$ and $\boldsymbol{\Sigma}_i^{-1} = (1/\sigma^2)\mathbf{I}$. Because both $|\boldsymbol{\Sigma}_i|$ and the $(d/2)$ ln 2π term in Eq. 49 are independent of i, they are unimportant additive constants that can be ignored. Thus we obtain the simple discriminant functions

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i), \tag{50}$$

EUCLIDEAN NORM

where $\|\cdot\|$ denotes the *Euclidean norm*, that is,

$$\|\mathbf{x} - \boldsymbol{\mu}_i\|^2 = (\mathbf{x} - \boldsymbol{\mu}_i)^t (\mathbf{x} - \boldsymbol{\mu}_i). \tag{51}$$

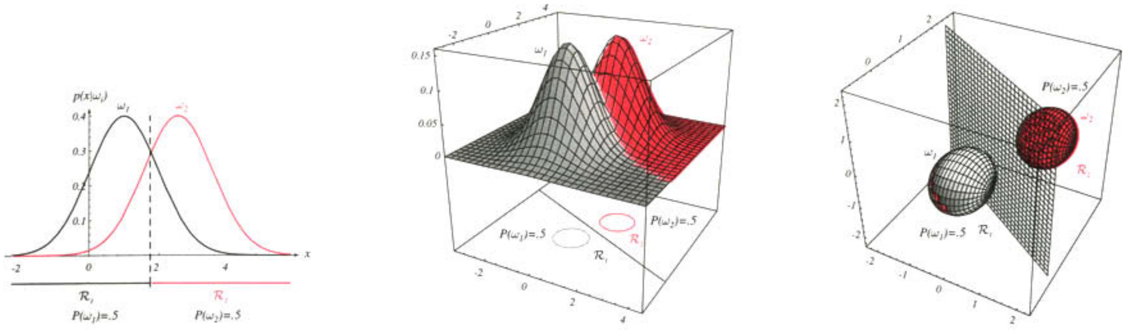


FIGURE 2.10. If the covariance matrices for two distributions are equal and proportional to the identity matrix, then the distributions are spherical in d dimensions, and the boundary is a generalized hyperplane of $d-1$ dimensions, perpendicular to the line separating the means. In these one-, two-, and three-dimensional examples, we indicate $p(\mathbf{x}|\omega_i)$ and the boundaries for the case $P(\omega_1) = P(\omega_2)$. In the three-dimensional case, the grid plane separates $\mathcal{R}_1$ from $\mathcal{R}_2$.

If the prior probabilities are not equal, then Eq. 50 shows that the squared distance $\|\mathbf{x}-\boldsymbol{\mu}\|^2$ must be normalized by the variance σ^2 and offset by adding $\ln P(\omega_i)$; thus, if $\mathbf{x}$ is equally near two different mean vectors, the optimal decision will favor the a priori more likely category.

Regardless of whether the prior probabilities are equal or not, it is not actually necessary to compute distances. Expansion of the quadratic form $(\mathbf{x}-\boldsymbol{\mu}_i)^t(\mathbf{x}-\boldsymbol{\mu}_i)$ yields

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}[\mathbf{x}^t\mathbf{x} - 2\boldsymbol{\mu}_i^t\mathbf{x} + \boldsymbol{\mu}_i^t\boldsymbol{\mu}_i] + \ln P(\omega_i), \qquad (52)$$

LINEAR DISCRIMINANT

which appears to be a quadratic function of $\mathbf{x}$. However, the quadratic term $\mathbf{x}^t\mathbf{x}$ is the same for all i, making it an ignorable additive constant. Thus, we obtain the equivalent *linear discriminant functions*

$$g_i(\mathbf{x}) = \mathbf{w}_i^t\mathbf{x} + w_{i0}, \qquad (53)$$

where

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i \qquad (54)$$

and

$$w_{i0} = \frac{-1}{2\sigma^2}\boldsymbol{\mu}_i^t\boldsymbol{\mu}_i + \ln P(\omega_i). \qquad (55)$$

THRESHOLD BIAS

We call w_{i0} the *threshold* or *bias* for the ith category.

LINEAR MACHINE

A classifier that uses linear discriminant functions is called a *linear machine*. This kind of classifier has many interesting theoretical properties, some of which will be discussed in Chapter 5. At this point we merely note that the decision surfaces for a linear machine are pieces of hyperplanes defined by the linear equations $g_i(\mathbf{x}) = g_j(\mathbf{x})$ for the two categories with the highest posterior probabilities. For our particular case, this equation can be written as

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0) = 0, \qquad (56)$$

where

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j \tag{57}$$

and

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j). \tag{58}$$

These equations define a hyperplane through the point $\mathbf{x}_0$ and orthogonal to the vector $\mathbf{w}$. Because $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$, the hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ is orthogonal to the line linking the means. If $P(\omega_i) = P(\omega_j)$, the second term on the right of Eq. 58 vanishes, and thus the point $\mathbf{x}_0$ is halfway between the means, and the hy-

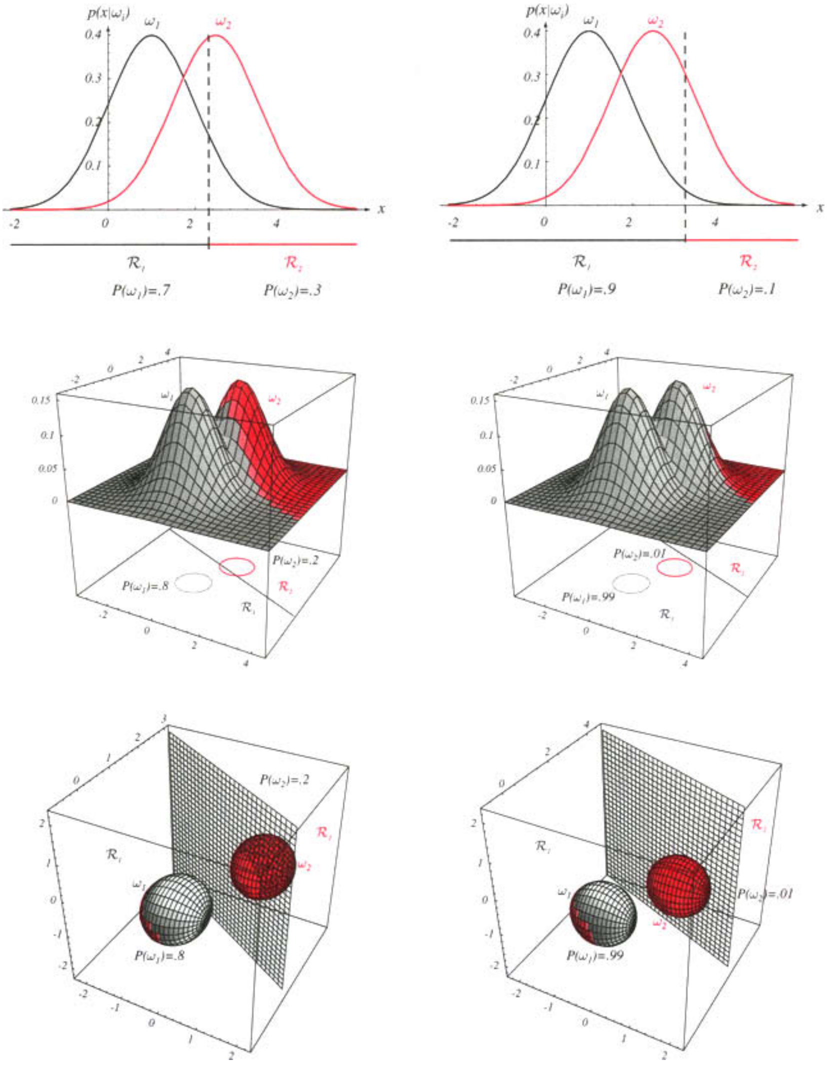


FIGURE 2.11. As the priors are changed, the decision boundary shifts; for sufficiently disparate priors the boundary will not lie between the means of these one-, two- and three-dimensional spherical Gaussian distributions.

perplane is the perpendicular bisector of the line between the means (Fig. 2.11). If $P(\omega_i) \neq P(\omega_j)$, the point $\mathbf{x}_0$ shifts away from the more likely mean. Note, however, that if the variance σ^2 is small relative to the squared distance $\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2$, then the position of the decision boundary is relatively insensitive to the exact values of the prior probabilities.

MINIMUM-DISTANCE CLASSIFIER

TEMPLATE-MATCHING

If the prior probabilities $P(\omega_i)$ are the same for all c classes, then the ln $P(\omega_i)$ term becomes another unimportant additive constant that can be ignored. When this happens, the optimum decision rule can be stated very simply: To classify a feature vector $\mathbf{x}$, measure the Euclidean distance $\|\mathbf{x} - \boldsymbol{\mu}_i\|$ from each $\mathbf{x}$ to each of the c mean vectors, and assign $\mathbf{x}$ to the category of the nearest mean. Such a classifier is called a *minimum-distance classifier*. If each mean vector is thought of as being an ideal prototype or template for patterns in its class, then this is essentially a *template-matching* procedure (Fig. 2.10), a technique we will consider again in Chapter 4 on the nearest-neighbor algorithm.

2.6.2 Case 2: $\Sigma_i = \Sigma$

Another simple case arises when the covariance matrices for all of the classes are identical but otherwise arbitrary. Geometrically, this corresponds to the situation in which the samples fall in hyperellipsoidal clusters of equal size and shape, the cluster for the ith class being centered about the mean vector $\boldsymbol{\mu}_i$. Because both $|\Sigma_i|$ and the $(d/2)$ ln 2π term in Eq. 49 are independent of i, they can be ignored as superfluous additive constants. This simplification leads to the discriminant functions

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i). \tag{59}$$

If the prior probabilities $P(\omega_i)$ are the same for all c classes, then the ln $P(\omega_i)$ term can be ignored. In this case, the optimal decision rule can once again be stated very simply: To classify a feature vector $\mathbf{x}$, measure the squared Mahalanobis distance $(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$ from $\mathbf{x}$ to each of the c mean vectors, and assign $\mathbf{x}$ to the category of the nearest mean. As before, unequal prior probabilities bias the decision in favor of the a priori more likely category.

Expansion of the quadratic form $(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_i)$ results in a sum involving a quadratic term $\mathbf{x}^t \Sigma^{-1}\mathbf{x}$ which here is independent of i. After this term is dropped from Eq. 59, the resulting discriminant functions are again linear:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0}, \tag{60}$$

where

$$\mathbf{w}_i = \Sigma^{-1}\boldsymbol{\mu}_i \tag{61}$$

and

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t \Sigma^{-1}\boldsymbol{\mu}_i + \ln P(\omega_i). \tag{62}$$

Because the discriminants are linear, the resulting decision boundaries are again hyperplanes (Fig. 2.10). If $\mathcal{R}_i$ and $\mathcal{R}_j$ are contiguous, the boundary between them has the equation

$$\mathbf{w}^t(\mathbf{x}-\mathbf{x}_0)=0, \tag{63}$$

where

$$\mathbf{w}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j) \tag{64}$$

and

$$\mathbf{x}_0=\frac{1}{2}(\boldsymbol{\mu}_i+\boldsymbol{\mu}_j)-\frac{\ln[P(\omega_i)/P(\omega_j)]}{(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)^t\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j). \tag{65}$$

Because $\mathbf{w}=\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i-\boldsymbol{\mu}_j)$ is generally not in the direction of $\boldsymbol{\mu}_i-\boldsymbol{\mu}_j$, the hyperplane separating $\mathcal{R}_i$ and $\mathcal{R}_j$ is generally not orthogonal to the line between the means. However, it does intersect that line at the point $\mathbf{x}_0$; if the prior probabilities are equal then $\mathbf{x}_0$ is halfway between the means. If the prior probabilities are not equal, the optimal boundary hyperplane is shifted away from the more likely mean (Fig. 2.12). As before, with sufficient bias the decision plane need not lie between the two mean vectors.

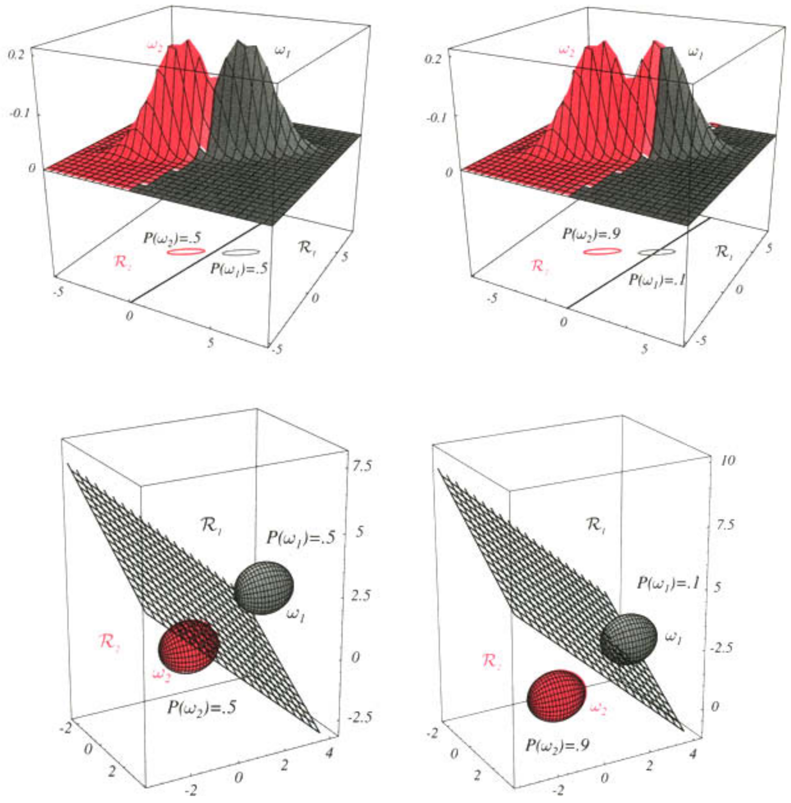


FIGURE 2.12. Probability densities (indicated by the surfaces in two dimensions and ellipsoidal surfaces in three dimensions) and decision regions for equal but asymmetric Gaussian distributions. The decision hyperplanes need not be perpendicular to the line connecting the means.

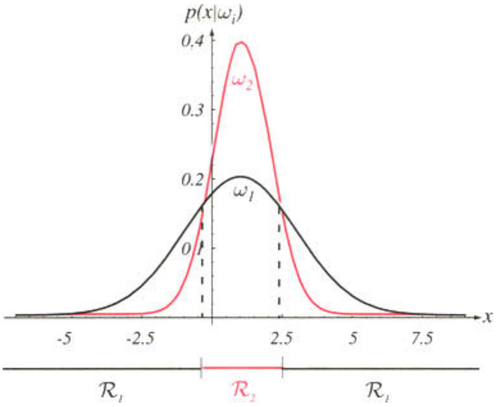


FIGURE 2.13. Non-simply connected decision regions can arise in one dimensions for Gaussians having unequal variance, as shown in this case with $P(\omega_1) = P(\omega_2)$.

2.6.3 Case 3: $\mathbf{\Sigma}_i$ = arbitrary

In the general multivariate normal case, the covariance matrices are different for each category. The only term that can be dropped from Eq. 49 is the $(d/2)\ \ln 2\pi$ term, and the resulting discriminant functions are inherently quadratic:

$$g_i(\mathbf{x}) = \mathbf{x}^t\mathbf{W}_i\mathbf{x} + \mathbf{w}_i^t\mathbf{x} + w_{i0}, \qquad (66)$$

where

$$\mathbf{W}_i = -\frac{1}{2}\mathbf{\Sigma}_i^{-1}, \qquad (67)$$

$$\mathbf{w}_i = \mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i \qquad (68)$$

and

$$w_{i0} = -\frac{1}{2}\boldsymbol{\mu}_i^t\mathbf{\Sigma}_i^{-1}\boldsymbol{\mu}_i - \frac{1}{2}\ln|\mathbf{\Sigma}_i| + \ln P(\omega_i). \qquad (69)$$

HYPERQUADRIC

In the two-category case, the decision surfaces are *hyperquadrics*, and they can assume any of the general forms: hyperplanes, pairs of hyperplanes, hyperspheres, hyperellipsoids, hyperparaboloids, and hyperhyperboloids of various types (Problem 30). Even in one dimension, for arbitrary variance the decision regions need not be simply connected (Fig. 2.13). The two- and three-dimensional examples in Figs. 2.14 and 2.15 indicate how these different forms can arise.

The extension of these results to more than two categories is straightforward though here we need to keep clear which two of the total c categories are responsible for any boundary segment. Figure 2.16 shows the decision surfaces for a four-category case made up of Gaussian distributions. Of course, if the distributions are more complicated, the decision regions can be even more complex, though the same underlying theory holds there too.

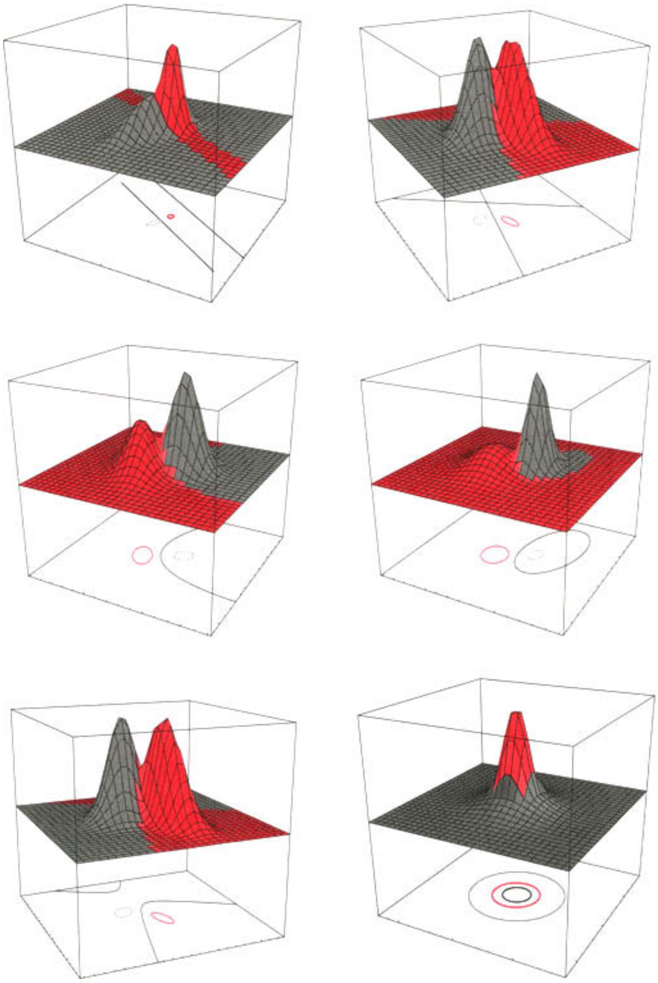


FIGURE 2.14. Arbitrary Gaussian distributions lead to Bayes decision boundaries that are general hyperquadrics. Conversely, given any hyperquadric, one can find two Gaussian distributions whose Bayes decision boundary is that hyperquadric. These variances are indicated by the contours of constant probability density.

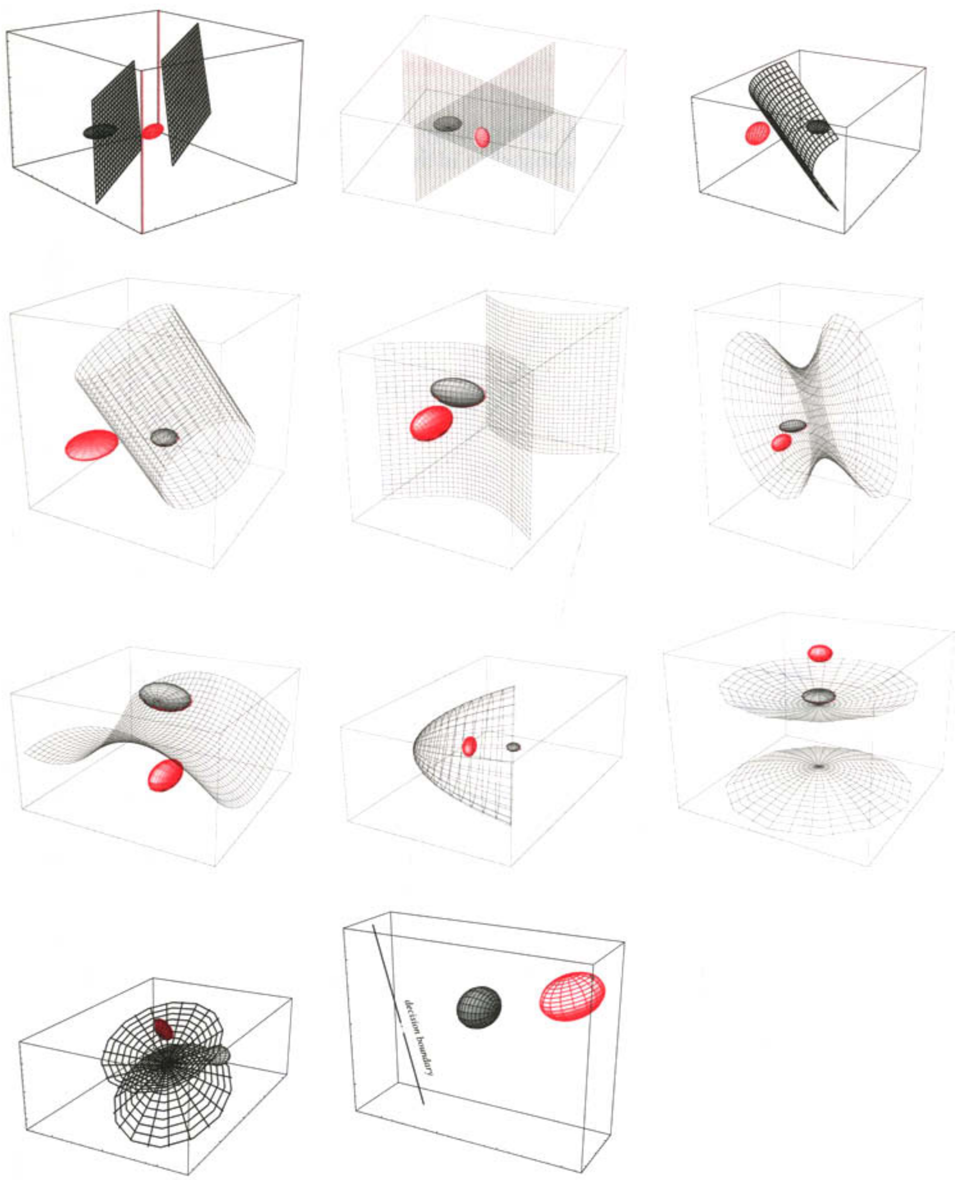
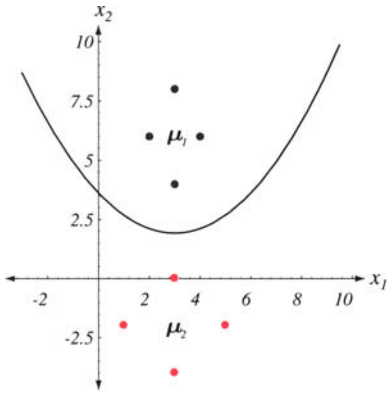


FIGURE 2.15. Arbitrary three-dimensional Gaussian distributions yield Bayes decision boundaries that are two-dimensional hyperquadrics. There are even degenerate cases in which the decision boundary is a line.

EXAMPLE 1 Decision Regions for Two-Dimensional Gaussian Data

To clarify these ideas, we explicitly calculate the decision boundary for the two-category two-dimensional data in the Example figure.



The computed Bayes decision boundary for two Gaussian distributions, each based on four data points.

Let ω_1 be the set of the four black points, and ω_2 the red points. Although we will spend much of the next chapter understanding how to estimate the parameters of our distributions, for now we simply assume that we need merely calculate the means and covariances by the discrete versions of Eqs. 40 and 41; they are found to be:

$$\boldsymbol{\mu}_1 = \begin{bmatrix} 3 \\ 6 \end{bmatrix}; \quad \boldsymbol{\Sigma}_1 = \begin{pmatrix} 1/2 & 0 \\ 0 & 2 \end{pmatrix} \text{ and } \boldsymbol{\mu}_2 = \begin{bmatrix} 3 \\ -2 \end{bmatrix}; \quad \boldsymbol{\Sigma}_2 = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

The inverse matrices are then,

$$\boldsymbol{\Sigma}_1^{-1} = \begin{pmatrix} 2 & 0 \\ 0 & 1/2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\Sigma}_2^{-1} = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}.$$

We assume equal prior probabilities, $P(\omega_1) = P(\omega_2) = 0.5$, and substitute these into the general form for a discriminant, Eqs. 66–69, setting $g_1(\mathbf{x}) = g_2(\mathbf{x})$ to obtain the decision boundary:

$$\mathbf{x}_2 = 3.514 - 1.125\mathbf{x}_1 + 0.1875\mathbf{x}_1^2.$$

This equation describes a parabola with vertex at $\binom{3}{1.83}$. Note that despite the fact that the variance in the data along the $\mathbf{x}_2$ direction for both distributions is the same, the decision boundary does not pass through the point $\binom{3}{2}$, midway between the means, as we might have naively guessed. This is because for the ω_1 distribution, the probability distribution is "squeezed" in the $\mathbf{x}_1$-direction more so than for the ω_2 distribution. The ω_1 distribution is increased along the $\mathbf{x}_2$ direction (relative to that for the ω_2 distribution). Thus the decision boundary lies slightly lower than the point midway between the two means, as can be seen in the decision boundary.

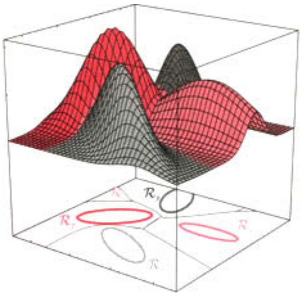


FIGURE 2.16. The decision regions for four normal distributions. Even with such a low number of categories, the shapes of the boundary regions can be rather complex.

*2.7 ERROR PROBABILITIES AND INTEGRALS

We can obtain additional insight into the operation of a general classifier—Bayes or otherwise—if we consider the sources of its error. Consider first the two-category case, and suppose the dichotomizer has divided the space into two regions $\mathcal{R}_1$ and $\mathcal{R}_2$ in a possibly nonoptimal way. There are two ways in which a classification error can occur; either an observation $\mathbf{x}$ falls in $\mathcal{R}_2$ and the true state of nature is ω_1, or $\mathbf{x}$ falls in $\mathcal{R}_1$ and the true state of nature is ω_2. Because these events are mutually exclusive and exhaustive, the probability of error is

$$\begin{aligned} P(error) &= P(\mathbf{x} \in \mathcal{R}_2, \omega_1) + P(\mathbf{x} \in \mathcal{R}_1, \omega_2) \\ &= P(\mathbf{x} \in \mathcal{R}_2|\omega_1)P(\omega_1) + P(\mathbf{x} \in \mathcal{R}_1|\omega_2)P(\omega_2) \\ &= \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)P(\omega_1)\, d\mathbf{x} + \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)P(\omega_2)\, d\mathbf{x}. \end{aligned} \tag{70}$$

This result is illustrated in the one-dimensional case in Fig. 2.17. The two integrals in Eq. 70 represent the pink and the gray areas in the tails of the functions $p(\mathbf{x}|\omega_i)P(\omega_i)$. Because the decision point x^* (and hence the regions $\mathcal{R}_1$ and $\mathcal{R}_2$) were chosen arbitrarily for that figure, the probability of error is not as small as it might be. In particular, the triangular area marked "reducible error" can be eliminated if the decision boundary is moved to x_B. This is the Bayes optimal decision boundary and gives the lowest probability of error. In general, if $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$, it is advantageous to classify $\mathbf{x}$ as in $\mathcal{R}_1$ so that the smaller quantity will contribute to the error integral; this is exactly what the Bayes decision rule achieves.

In the multicategory case, there are more ways to be wrong than to be right, and it is simpler to compute the probability of being correct. Clearly,

$$\begin{aligned} P(correct) &= \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i, \omega_i) \\ &= \sum_{i=1}^{c} P(\mathbf{x} \in \mathcal{R}_i|\omega_i)P(\omega_i) \\ &= \sum_{i=1}^{c} \int_{\mathcal{R}_i} p(\mathbf{x}|\omega_i)P(\omega_i)\, d\mathbf{x}. \end{aligned} \tag{71}$$

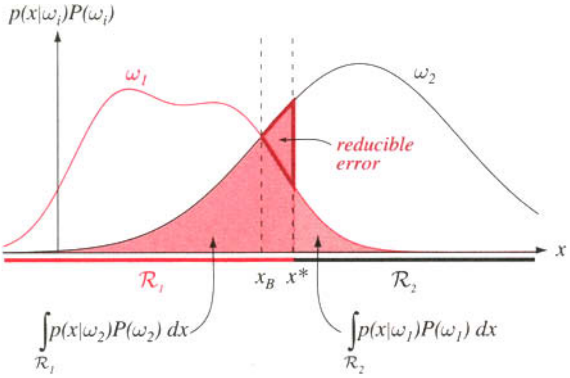


FIGURE 2.17. Components of the probability of error for equal priors and (nonoptimal) decision point x^*. The pink area corresponds to the probability of errors for deciding ω_1 when the state of nature is in fact ω_2; the gray area represents the converse, as given in Eq. 70. If the decision boundary is instead at the point of equal posterior probabilities, x_B, then this reducible error is eliminated and the total shaded area is the minimum possible; this is the Bayes decision and gives the Bayes error rate.

The general result of Eq. 71 depends neither on how the feature space is partitioned into decision regions nor on the form of the underlying distributions. The Bayes classifier maximizes this probability by choosing the regions so that the integrand is maximal for all **x**; no other partitioning can yield a smaller probability of error.

*2.8 ERROR BOUNDS FOR NORMAL DENSITIES

The Bayes decision rule guarantees the lowest average error rate, and we have seen how to calculate the decision boundaries for normal densities. However, these results do not tell us what the probability of error actually *is*. The full calculation of the error for the Gaussian case would be quite difficult, especially in high dimensions, because of the discontinuous nature of the decision regions in the integral in Eq. 71. However, in the two-category case the general error integral of Eq. 5 can be approximated analytically to give us an upper bound on the error.

2.8.1 Chernoff Bound

To derive a bound for the error, we need the following inequality:

$$\min[a, b] \leq a^{\beta} b^{1-\beta} \quad \text{for } a, b \geq 0 \text{ and } 0 \leq \beta \leq 1. \tag{72}$$

To understand this inequality we can, without loss of generality, assume $a \geq b$. Thus we need only show that $b \leq a^{\beta} b^{1-\beta} = (a/b)^{\beta} b$. But this inequality is manifestly valid, because $(a/b)^{\beta} \geq 1$. Using Eqs. 7 and 1, we apply this inequality to the vector form of Eq. 5 and get the bound:

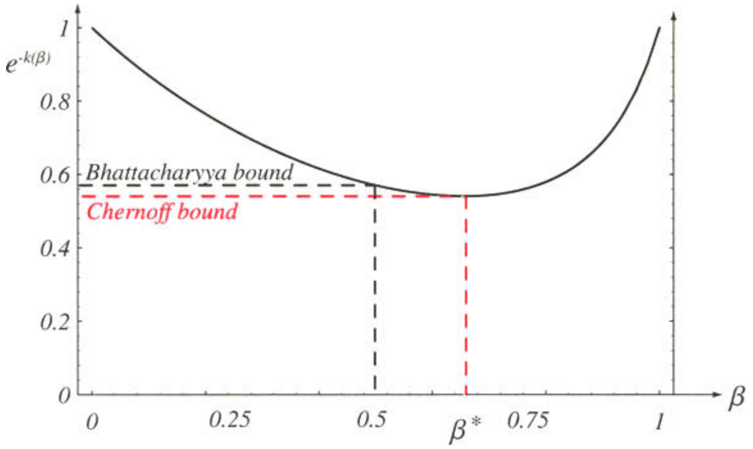


FIGURE 2.18. The Chernoff error bound is never looser than the Bhattacharyya bound. For this example, the Chernoff bound happens to be at $\beta^* = 0.66$, and is slightly tighter than the Bhattacharyya bound ($\beta = 0.5$).

$$P(error) \leq P^{\beta}(\omega_1)P^{1-\beta}(\omega_2)\int p^{\beta}(\mathbf{x}|\omega_1)p^{1-\beta}(\mathbf{x}|\omega_2)\,d\mathbf{x} \qquad \text{for } 0 \leq \beta \leq 1. \tag{73}$$

Note especially that this integral is over *all* feature space—we do not need to impose integration limits corresponding to decision boundaries.

If the conditional probabilities are normal, the integral in Eq. 73 can be evaluated analytically (Problem 36), yielding

$$\int p^{\beta}(\mathbf{x}|\omega_1)p^{1-\beta}(\mathbf{x}|\omega_2)\,d\mathbf{x} = e^{-k(\beta)} \tag{74}$$

where

$$\begin{aligned} k(\beta) = {} & \frac{\beta(1-\beta)}{2}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^t[(1-\beta)\boldsymbol{\Sigma}_1 + \beta\boldsymbol{\Sigma}_2]^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \\ & + \frac{1}{2}\ln\frac{(1-\beta)\boldsymbol{\Sigma}_1 + \beta\boldsymbol{\Sigma}_2}{|\boldsymbol{\Sigma}_1|^{1-\beta}|\boldsymbol{\Sigma}_2|^{\beta}}. \end{aligned} \tag{75}$$

The graph in Fig. 2.18 shows a typical example of how $e^{-k(\beta)}$ varies with β. The *Chernoff bound* on $P(error)$ is found by analytically or numerically finding the value of β that minimizes $P^{\beta}(\omega_1)P^{1-\beta}(\omega_2)e^{-k(\beta)}$ and then substituting this β into Eq. 73. The key benefit here is that this optimization is in the one-dimensional β space, despite the fact that the distributions themselves might be in a space of arbitrarily high dimension.

2.8.2 Bhattacharyya Bound

The general dependence of the Chernoff bound upon β shown in Fig. 2.18 is typical of a wide range of problems; The bound is loose for extreme values (i.e., $\beta \to 1$ and $\beta \to 0$), and it is tighter for intermediate ones. While the precise value of the optimal β depends upon the parameters of the distributions and the prior probabilities, a computationally simpler but slightly less tight bound can be derived by simply setting $\beta = 1/2$. This gives the so-called *Bhattacharyya bound* on the error, where Eq. 73 then has the form

$$P(error) \leq \sqrt{P(\omega_1)P(\omega_2)} \int \sqrt{p(\mathbf{x}|\omega_1)p(\mathbf{x}|\omega_2)}\, d\mathbf{x}$$
$$= \sqrt{P(\omega_1)P(\omega_2)}e^{-k(1/2)}, \qquad (76)$$

where by Eq. 75 we have for the Gaussian case:

$$k(1/2) = 1/8(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)^t \left[\frac{\boldsymbol{\Sigma}_1 + \boldsymbol{\Sigma}_2}{2}\right]^{-1} (\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1) + \frac{1}{2} \ln \frac{\left|\frac{\boldsymbol{\Sigma}_1+\boldsymbol{\Sigma}_2}{2}\right|}{\sqrt{|\boldsymbol{\Sigma}_1||\boldsymbol{\Sigma}_2|}}. \qquad (77)$$

The Chernoff and Bhatacharyya bounds may still be used even if the underlying distributions are not Gaussian. However, for distributions that deviate markedly from a Gaussian, the bounds will not be informative (Problem 34).

EXAMPLE 2 Error Bounds for Gaussian Distributions

It is a straightforward matter to calculate the Bhattacharyya bound for the two-dimensional data sets of Example 1. Substituting the means and covariances of Example 1 into Eq. 77, we find $k(1/2) = 4.11157$, and thus by Eqs. 76 and 77 the Bhattacharyya bound on the error is $P(error) \leq 0.008191$.

A slightly tighter bound on the error can be approximated by searching numerically for the Chernoff bound of Eq. 75, which for this problem gives 0.008190. Numerical integration of Eq. 5 gives an error rate of 0.0021; thus the bounds here are not particularly tight. Such numerical integration is often impractical for Gaussians in higher than two or three dimensions.

2.8.3 Signal Detection Theory and Operating Characteristics

Another measure of distance between two Gaussian distributions has found great use in experimental psychology, radar detection and other fields. Suppose we are interested in detecting a single weak pulse, such as a dim flash of light or a weak radar reflection. Our model is, then, that at some point in the detector there is an internal signal (such as a voltage) x, whose value has mean μ_2 when the external signal (pulse) is present, and mean μ_1 when it is not present. Because of random noise—within and outside the detector itself—the actual value is a random variable. We assume the distributions are normal with different means but the same variance—that is, $p(x|\omega_i) \sim N(\mu_i, \sigma^2)$—as shown in Fig. 2.19.

The detector (classifier) employs a threshold value x^* for determining whether the external pulse is present, but suppose we, as experimenters, do not have access to this value (nor to the means and standard deviations of the distributions). We seek to find some measure of the ease of discriminating whether the pulse is present or not, in a form independent of the choice of x^*. Such a measure is the *discriminability*, which describes the inherent and unchangeable properties due to noise and the strength of the external signal, but not on the decision strategy (i.e., the actual choice of x^*). This discriminability is defined as

DISCRIMINABILITY

$$d' = \frac{|\mu_2 - \mu_1|}{\sigma}. \qquad (78)$$

A high d' is of course desirable.

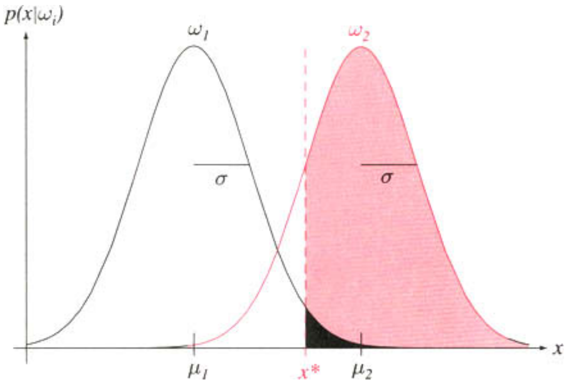


FIGURE 2.19. During any instant when no external pulse is present, the probability density for an internal signal is normal, that is, $p(x|\omega_1) \sim N(\mu_1, \sigma^2)$; when the external signal is present, the density is $p(x|\omega_2) \sim N(\mu_2, \sigma^2)$. Any decision threshold x^* will determine the probability of a hit (the pink area under the ω_2 curve, above x^*) and of a false alarm (the black area under the ω_1 curve, above x^*).

While we do not know μ_1, μ_2, σ or x^*, we assume here that we know the state of nature and the decision of the system. Such information allows us to find d'. To this end, we consider the following four probabilities:

- $P(x > x^*|x \in \omega_2)$: a *hit*—the probability that the internal signal is above x^* given that the external signal is present
- $P(x > x^*|x \in \omega_1)$: a *false alarm*—the probability that the internal signal is above x^* despite there being no external signal present
- $P(x < x^*|x \in \omega_2)$: a *miss*—the probability that the internal signal is below x^* given that the external signal is present
- $P(x < x^*|x \in \omega_1)$: a *correct rejection*—the probability that the internal signal is below x^* given that the external signal is not present.

RECEIVER OPERATING CHARACTERISTIC

If we have a large number of trials (and we can assume x^* is fixed, albeit at an unknown value), we can determine these probabilities experimentally, in particular the hit and false alarm rates. We plot a point representing these rates on a two-dimensional graph. If the densities are fixed but the threshold x^* is changed, then our hit and false alarm rates will also change. Thus we see that for a given discriminability d', our point will move along a smooth curve—a *receiver operating characteristic* or ROC curve (Fig. 2.20).

The great benefit of this signal detection framework is that we can distinguish operationally between *discriminability* and *decision bias*: While the former is an inherent property of the detector system, the latter is due to the receiver's implied but changeable loss matrix. Through any pair of hit and false alarm rates passes one and only one ROC curve; thus, so long as neither rate is exactly 0 or 1, we can determine the discriminability from these rates (Problem 39). Moreover, if the Gaussian assumption holds, a determination of the discriminability (from an arbitrary x^*) allows us to calculate the Bayes error rate—the most important property of any classifier. If the actual error rate differs from the Bayes rate inferred in this way, we should alter the threshold x^* accordingly.

It is a simple matter to generalize the above discussion and apply it to two categories having arbitrary multidimensional distributions, Gaussian or not. Suppose we

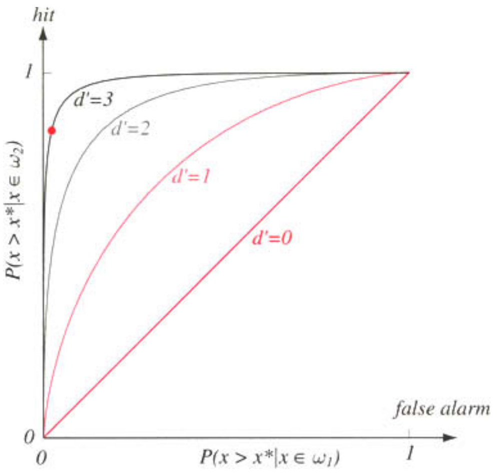


FIGURE 2.20. In a receiver operating characteristic (ROC) curve, the abscissa is the probability of false alarm, $P(x > x^*|x \in \omega_1)$, and the ordinate is the probability of hit, $P(x > x^*|x \in \omega_2)$. From the measured hit and false alarm rates (here corresponding to x^* in Fig. 2.19 and shown as the red dot), we can deduce that $d' = 3$.

have two distributions $p(\mathbf{x}|\omega_1)$ and $p(\mathbf{x}|\omega_2)$ which overlap, and thus have nonzero Bayes classification error. Just as we saw above, any pattern actually from ω_2 could be properly classified as ω_2 (a "hit") or misclassified as ω_1 (a "false alarm"). Unlike the one-dimensional case above, however, there may be *many* decision boundaries that correspond to a particular hit rate, each with a different false alarm rate. Clearly here we cannot determine a fundamental measure of discriminability without knowing more about the underlying decision rule than just the hit and false alarm rates.

In a rarely attainable ideal, we can imagine that our measured hit and false alarm rates are *optimal*—for example, that of all the decision rules giving the measured hit rate, the rule that is actually used is the one having the minimum false alarm rate. If we constructed a multidimensional classifier—regardless of the distributions used—we might try to characterize the problem in this way, though it would probably require great computational resources to search for such optimal hit and false alarm rates.

OPERATING CHARACTERISTIC

In practice, instead we forgo optimality, and simply vary a single control parameter for the decision rule and plot the resulting hit and false alarm rates—a curve called merely an *operating characteristic*. It is traditional to use a control parameter that can yield, at extreme values, either a vanishing false alarm or a vanishing hit rate, just as can be achieved with a very large or a very small x^* in an ROC curve. We should note that since the distributions can be arbitrary, the operating characteristic need not be symmetric (Fig. 2.21); in rare cases it need not even be concave down at all points.

Classifier operating curves are of value for problems where the loss matrix λ_{ij} might be changed. If the operating characteristic has been determined as a function of the control parameter ahead of time, it is a simple matter, when faced with a new loss function, to deduce the control parameter setting that will minimize the expected risk (Problem 39).

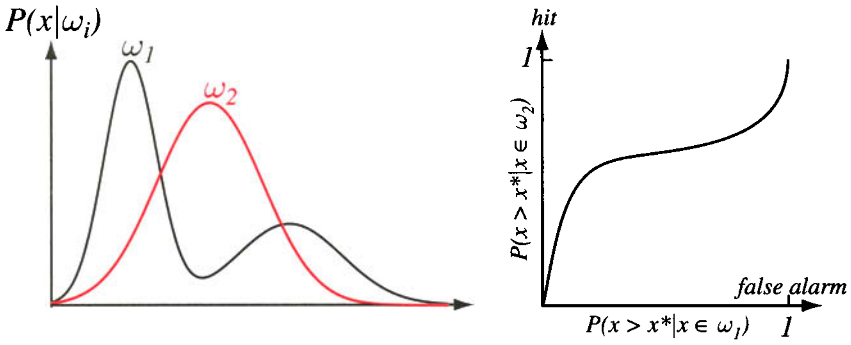


FIGURE 2.21. In a general operating characteristic curve, the abscissa is the probability of false alarm, $P(x \in \mathcal{R}_2|x \in \omega_1)$, and the ordinate is the probability of hit, $P(x \in \mathcal{R}_2|x \in \omega_2)$. As illustrated here, operating characteristic curves are generally not symmetric, as shown at the right.

2.9 BAYES DECISION THEORY—DISCRETE FEATURES

Until now we have assumed that the feature vector $\mathbf{x}$ could be any point in a d-dimensional Euclidean space, $\mathbf{R}^d$. However, in many practical applications the components of $\mathbf{x}$ are binary-, ternary-, or higher-integer-valued, so that $\mathbf{x}$ can assume only one of m discrete values $\mathbf{v}_1, \ldots, \mathbf{v}_m$. In such cases, the probability density function $p(\mathbf{x}|\omega_j)$ becomes singular; integrals of the form

$$\int p(\mathbf{x}|\omega_j)\, d\mathbf{x} \tag{79}$$

must then be replaced by corresponding sums, such as

$$\sum_{\mathbf{x}} P(\mathbf{x}|\omega_j), \tag{80}$$

where we understand that the summation is over all values of $\mathbf{x}$ in the discrete distribution.* Bayes formula then involves probabilities, rather than probability densities:

$$P(\omega_j|\mathbf{x}) = \frac{P(\mathbf{x}|\omega_j)P(\omega_j)}{P(\mathbf{x})}, \tag{81}$$

where

$$P(\mathbf{x}) = \sum_{j=1}^{c} P(\mathbf{x}|\omega_j)P(\omega_j). \tag{82}$$

The definition of the conditional risk $R(\alpha|\mathbf{x})$ is unchanged, and the fundamental Bayes decision rule remains the same: To minimize the overall risk, select the action α_i for which $R(\alpha_i|\mathbf{x})$ is minimum, or stated formally,

$$\alpha^* = \arg\min_i R(\alpha_i|\mathbf{x}). \tag{83}$$

*Technically speaking, Eq. 80 should be written as $\sum_k P(\mathbf{v}_k|\omega_j)$ where $P(\mathbf{v}_k|\omega_j)$ is the conditional probability that $\mathbf{x} = \mathbf{v}_k$, given that the state of nature is ω_j.

The basic rule to minimize the error rate by maximizing the posterior probability is also unchanged as are the discriminant functions of Eqs. 26–28, given the obvious replacement of densities $p(\cdot)$ by probabilities $P(\cdot)$.

2.9.1 Independent Binary Features

As an example of a classification involving discrete features, consider the two-category problem in which the components of the feature vector are binary-valued and conditionally independent. To be more specific we let $\mathbf{x} = (x_1, \ldots, x_d)^t$, where the components x_i are either 0 or 1, with probabilities

$$p_i = \Pr[x_i = 1|\omega_1] \tag{84}$$

and

$$q_i = \Pr[x_i = 1|\omega_2]. \tag{85}$$

This is a model of a classification problem in which each feature gives us a yes/no answer about the pattern. If $p_i > q_i$, we expect the ith feature to give a "yes" answer more frequently when the state of nature is ω_1 than when when it is ω_2. (As an example, consider two factories each making the same automobile, each of whose d components could be functional or defective. If it were known how the factories differed in their reliabilities for making each component, then this model could be used to judge which factory manufactured a given automobile based on the knowledge of which features are functional and which defective.) By assuming conditional independence we can write $P(\mathbf{x}|\omega_i)$ as the product of the probabilities for the components of $\mathbf{x}$. Given this assumption, a particularly convenient way of writing the class-conditional probabilities is as follows:

$$P(\mathbf{x}|\omega_1) = \prod_{i=1}^{d} p_i^{x_i}(1 - p_i)^{1-x_i} \tag{86}$$

and

$$P(\mathbf{x}|\omega_2) = \prod_{i=1}^{d} q_i^{x_i}(1 - q_i)^{1-x_i}. \tag{87}$$

Then the likelihood ratio is given by

$$\frac{P(\mathbf{x}|\omega_1)}{P(\mathbf{x}|\omega_2)} = \prod_{i=1}^{d} \left(\frac{p_i}{q_i}\right)^{x_i} \left(\frac{1 - p_i}{1 - q_i}\right)^{1-x_i} \tag{88}$$

and consequently Eq. 31 yields the discriminant function

$$g(\mathbf{x}) = \sum_{i=1}^{d} \left[x_i \ln \frac{p_i}{q_i} + (1 - x_i) \ln \frac{1 - p_i}{1 - q_i} \right] + \ln \frac{P(\omega_1)}{P(\omega_2)}. \tag{89}$$

We note especially that this discriminant function is linear in the x_i and thus we can write

$$g(\mathbf{x}) = \sum_{i=1}^{d} w_i x_i + w_0, \tag{90}$$

where

$$w_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} \qquad i = 1, \ldots, d \tag{91}$$

and

$$w_0 = \sum_{i=1}^{d} \ln \frac{1-p_i}{1-q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)}. \tag{92}$$

Let us examine these results to see what insight they can give. Recall first that we decide ω_1 if $g(\mathbf{x}) > 0$ and ω_2 if $g(\mathbf{x}) \leq 0$. We have seen that $g(\mathbf{x})$ is a weighted combination of the components of $\mathbf{x}$. The magnitude of the weight w_i indicates the relevance of a "yes" answer for x_i in determining the classification. If $p_i = q_i$, x_i gives us no information about the state of nature, and $w_i = 0$, just as we might expect. If $p_i > q_i$, then $1 - p_i < 1 - q_i$ and w_i is positive. Thus in this case a "yes" answer for x_i contributes w_i votes for ω_1. Furthermore, for any fixed $q_i < 1$, w_i gets larger as p_i gets larger. On the other hand, if $p_i < q_i$, w_i is negative and a "yes" answer contributes $|w_i|$ votes for ω_2.

The condition of feature independence leads to a very simple (linear) classifier; of course if the features were not independent, a more complicated classifier would be needed. We shall come across this again for systems with continuous features but note here that the more independent we can make the features, the simpler the classifier can be.

The prior probabilities $P(\omega_i)$ appear in the discriminant only through the threshold weight w_0. Increasing $P(\omega_1)$ increases w_0 and biases the decision in favor of ω_1, whereas decreasing $P(\omega_1)$ has the opposite effect. Geometrically, the possible values for $\mathbf{x}$ appear as the vertices of a d-dimensional hypercube; the decision surface defined by $g(\mathbf{x}) = 0$ is a hyperplane that separates ω_1 vertices from ω_2 vertices.

EXAMPLE 3 Bayesian Decisions for Three-Dimensional Binary Data

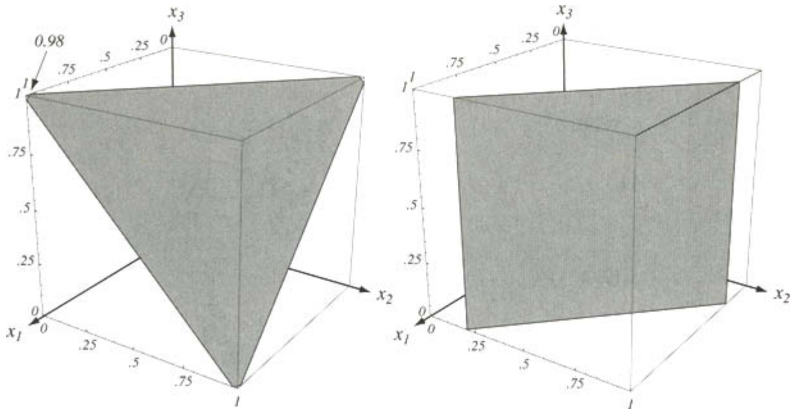
Consider a two-class problem having three independent binary features with known feature probabilities. Let us construct the Bayesian decision boundary if $P(\omega_1) = P(\omega_2) = 0.5$ and the individual components obey $p_i = 0.8$ and $q_i = 0.5$ for $i = 1, 2, 3$. By Eqs. 91 and 92 we have that the weights are

$$w_i = \ln \frac{.8(1-.5)}{.5(1-.8)} = 1.3863$$

and the bias value is

$$w_0 = \sum_{i=1}^{3} \ln \frac{1-.8}{1-.5} + \ln \frac{.5}{.5} = -2.75.$$

The surface $g(\mathbf{x}) = 0$ from Eq. 90 is shown on the left of the figure. Indeed, as we might have expected, the boundary places points with two or more "yes" answers into category ω_1, because that category has a higher probability of having any feature take value 1.



The decision boundary for the example involving three-dimensional binary features. On the left we show the case $p_i = .8$ and $q_i = .5$. On the right we use the same values except $p_3 = q_3$, which leads to $w_3 = 0$ and a decision surface parallel to the x_3 axis.

Suppose instead that while the prior probabilities remained the same, our individual components obeyed $p_1 = p_2 = 0.8$, $p_3 = 0.5$ and $q_1 = q_2 = q_3 = 0.5$. In this case feature x_3 gives us no predictive information about the categories, and hence the decision boundary is parallel to the x_3 axis. Note that in this discrete case there is a large range in positions of the decision boundary that leaves the categorization unchanged, as is particularly clear in the figure on the right.

*2.10 MISSING AND NOISY FEATURES

If we know the full probability structure of a problem, we can construct the (optimal) Bayes decision rule. Suppose we develop a Bayes classifier using uncorrupted data, but our input (test) data are then corrupted in particular known ways. How can we classify such corrupted inputs to obtain a minimum error now?

There are two analytically solvable cases of particular interest: when some of the features are *missing*, and when they are corrupted by a *noise source* with known properties. In each case our basic approach is to recover as much information about the underlying distribution as possible and use the Bayes decision rule.

2.10.1 Missing Features

Suppose we have a Bayesian (or other) recognizer for a problem using two features, but that for a particular pattern to be classified, one of the features is missing. For example, we can easily imagine that the lightness can be measured from a portion of a fish, but the width cannot because of occlusion by another fish.

We can illustrate with four categories a somewhat more general case (Fig. 2.22). Suppose that for a particular test pattern the feature x_1 is missing, and the measured value of x_2 is $\hat{x}_2$. Clearly if we assume that the missing value is the *mean* of all the x_1 values (i.e., $\bar{x}_1$), we will classify the pattern as ω_3. However, if the priors are equal, ω_2 would be a better decision, because the figure implies that $p(\hat{x}_2|\omega_2)$ is the largest of the four likelihoods.

To clarify our derivation we let $\mathbf{x} = [\mathbf{x}_g, \mathbf{x}_b]$, where $\mathbf{x}_g$ represents the known or "good" features and $\mathbf{x}_b$ represents the "bad" ones—that is, either unknown or missing. We seek the Bayes rule given the good features, and for that the posterior

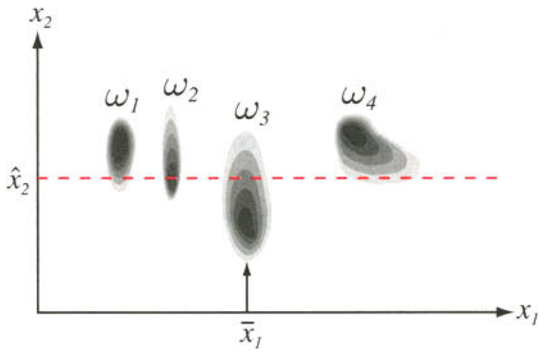


FIGURE 2.22. Four categories have equal priors and the class-conditional distributions shown. If a test point is presented in which one feature is missing (here, x_1) and the other is measured to have value $\hat{x}_2$ (red dashed line), we want our classifier to classify the pattern as category ω_2, because $p(\hat{x}_2|\omega_2)$ is the largest of the four likelihoods.

probabilities are needed. In terms of the good features the posteriors are

$$\begin{aligned} P(\omega_i|\mathbf{x}_g) &= \frac{p(\omega_i, \mathbf{x}_g)}{p(\mathbf{x}_g)} = \frac{\int p(\omega_i, \mathbf{x}_g, \mathbf{x}_b)\, d\mathbf{x}_b}{p(\mathbf{x}_g)} \\ &= \frac{\int P(\omega_i|\mathbf{x}_g, \mathbf{x}_b) p(\mathbf{x}_g, \mathbf{x}_b)\, d\mathbf{x}_b}{p(\mathbf{x}_g)} \\ &= \frac{\int g_i(\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}_b}{\int p(\mathbf{x})\, d\mathbf{x}_b}, \end{aligned} \tag{93}$$

where $g_i(\mathbf{x}) = g_i(\mathbf{x}_g, \mathbf{x}_b) = P(\omega_i|\mathbf{x}_g, \mathbf{x}_b)$ is one form of our discriminant function.

MARGINAL

We refer to $\int p(\omega_i, \mathbf{x}_g, \mathbf{x}_b)\, d\mathbf{x}_b$, as a *marginal distribution*; we say the full joint distribution is marginalized over the variable $\mathbf{x}_b$. In short, Eq. 93 shows that we must integrate (marginalize) the posterior probability over the bad features. Finally we use the Bayes decision rule on the resulting posterior probabilities, that is, choose ω_i if $P(\omega_i|\mathbf{x}_g) > P(\omega_j|\mathbf{x}_g)$ for all i and j. In Chapter 3 we shall consider the Expectation-Maximization (EM) algorithm, which addresses a related problem involving missing features.

2.10.2 Noisy Features

It is a simple matter to generalize the results of Eq. 93 to the case where a particular feature has been corrupted by statistically independent noise. For instance, in our fish classification example, we might have a reliable measurement of the length, while variability of the light source might degrade the measurement of the lightness. We assume we have uncorrupted (good) features $\mathbf{x}_g$, as before, and a *noise model*, expressed as $p(\mathbf{x}_b|\mathbf{x}_t)$. Here we let $\mathbf{x}_t$ denote the true value of the observed $\mathbf{x}_b$ features, that is, without the noise present; in short, the $\mathbf{x}_b$ are observed instead of the true $\mathbf{x}_t$. We assume that if $\mathbf{x}_t$ were known, $\mathbf{x}_b$ would be independent of ω_i and $\mathbf{x}_g$. From such an assumption we get:

$$P(\omega_i|\mathbf{x}_g, \mathbf{x}_b) = \frac{\int p(\omega_i, \mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)\, d\mathbf{x}_t}{p(\mathbf{x}_g, \mathbf{x}_b)}. \tag{94}$$

Now $p(\omega_i, \mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) = P(\omega_i|\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t)$, but by our independence assumption, if we know $\mathbf{x}_t$, then $\mathbf{x}_b$ does not provide any additional information about ω_i. Thus we have $P(\omega_i|\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) = P(\omega_i|\mathbf{x}_g, \mathbf{x}_t)$. Similarly, we have $p(\mathbf{x}_g, \mathbf{x}_b, \mathbf{x}_t) = p(\mathbf{x}_b|\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_t)$ and $p(\mathbf{x}_b|\mathbf{x}_g, \mathbf{x}_t) = p(\mathbf{x}_b|\mathbf{x}_t)$. We put these together and thereby obtain

$$\begin{aligned} P(\omega_i|\mathbf{x}_g, \mathbf{x}_b) &= \frac{\int P(\omega_i|\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_b|\mathbf{x}_t)\,d\mathbf{x}_t}{\int p(\mathbf{x}_g, \mathbf{x}_t)p(\mathbf{x}_b|\mathbf{x}_t)\,d\mathbf{x}_t} \\ &= \frac{\int g_i(\mathbf{x})p(\mathbf{x})p(\mathbf{x}_b|\mathbf{x}_t)\,d\mathbf{x}_t}{\int p(\mathbf{x})p(\mathbf{x}_b|\mathbf{x}_t)\,d\mathbf{x}_t}, \end{aligned} \tag{95}$$

which we use as discriminant functions for classification in the manner dictated by Bayes.

Equation 95 differs from Eq. 93 solely by the fact that the integral is weighted by the noise model. In the extreme case where $p(\mathbf{x}_b|\mathbf{x}_t)$ is uniform over the entire space (and hence provides no predictive information for categorization), the equation reduces to the case of missing features—a satisfying result.

*2.11 BAYESIAN BELIEF NETWORKS

The methods we have described up to now are fairly general—all that we assumed, at base, was that we could parameterize the probability distributions by a vector $\boldsymbol{\theta}$. If we had prior information about $\boldsymbol{\theta}$, this too could be used. Sometimes our knowledge about a distribution is not directly expressed by a parameter vector, but instead about the statistical dependencies (or independencies) or the causal relationships among the component variables. (Recall that for some multidimensional distribution $p(\mathbf{x})$, if for two features we have $p(x_i, x_j) = p(x_i)p(x_j)$, we say those variables are statistically independent, as illustrated in Fig. 2.23.)

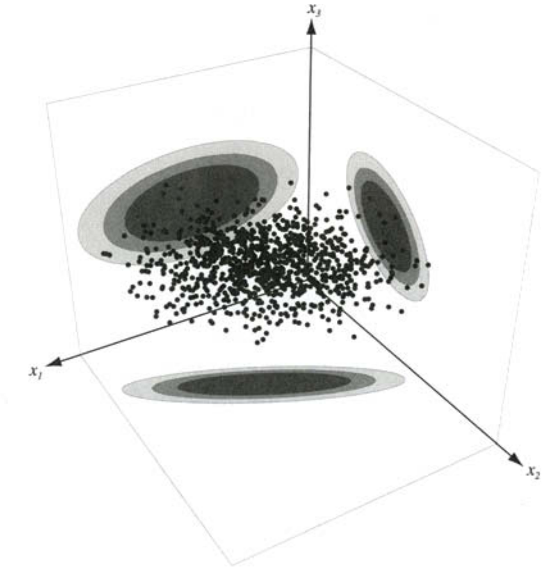


FIGURE 2.23. A three-dimensional distribution which obeys $p(x_1, x_3) = p(x_1)p(x_3)$; thus here x_1 and x_3 are statistically independent but the other feature pairs are not.

There are many such cases where we know—or can safely assume—which variables are or are not causally related, even if it may be more difficult to specify the precise probabilistic relationships among those variables. Suppose, for instance, we are describing the state of an automobile: temperature of the engine, pressure of the brake fluid, pressure of the air in the tires, voltages in the wires, and so on. Our basic knowledge of cars includes the fact that the oil pressure in the engine and the air pressure in a tire are *not* causally related while the engine temperature and oil temperature *are* causally related. Furthermore, we may know *several* variables that might influence another: The coolant temperature is affected by the engine temperature, the speed of the radiator fan (which blows air over the coolant-filled radiator), and so on. We shall now exploit this structural information when reasoning about the system and its variables.

We represent these causal dependencies graphically by means of *Bayesian belief nets*, also called *causal networks*, or simply *belief nets*. While these nets can represent continuous multidimensional distributions over their variables, they have enjoyed greatest application and success for discrete variables. For this reason, and because the calculations are simpler, we shall concentrate on the discrete case.

NODE

Each *node* (or unit) represents one of the system components, and here it takes on discrete values. We label nodes **A**, **B**, ... and their variables by the corresponding lowercase letter. Thus, while there are a discrete number of possible values of node **A**—for instance two, a_1 and a_2—there may be continuous-valued *probabilities* on these discrete states. For example, if node **A** represents the automobile ignition switch—$a_1 = on$, $a_2 = off$—we might have $P(a_1) = 0.739$, $P(a_2) = 0.261$, or indeed any other probabilities. Each link in the net is directional and joins two nodes; the link represents the causal influence of one node upon another. Thus in the net in Fig. 2.24 **A** directly influences **D**. While **B** also influences **D**, such influence is indirect, through **C**. In considering a single node in a net, it is useful to distinguish the set of nodes immediately *before* that node—called its *parents*—and the set of those immediately *after* it—called its *children*. Thus in Fig. 2.24 the parents of **D** are **A** and **C** while the child of **D** is **E**. Suppose we have a belief net, complete with causal dependencies indicated by the topology of the links. Through a direct application of Bayes rule, we can determine the probability of any configuration of variables in

PARENT

CHILD

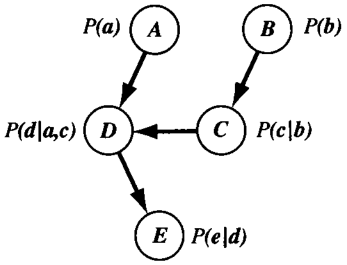


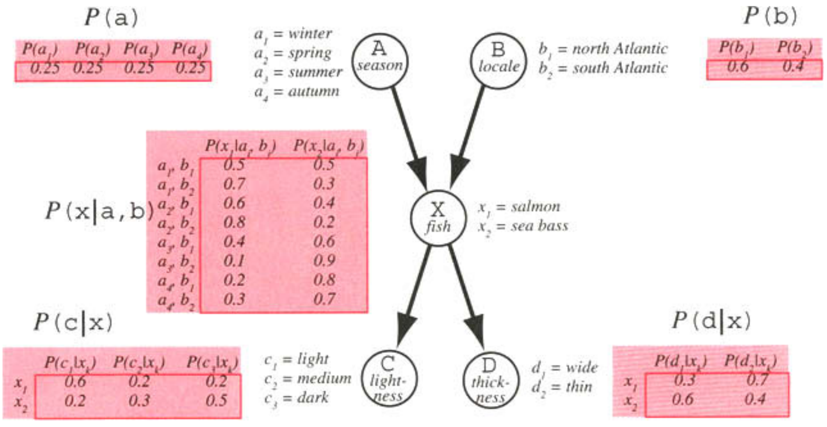
FIGURE 2.24. A belief network consists of nodes (labeled with uppercase bold letters) and their associated discrete states (in lowercase). Thus node **A** has states $\{a_1, a_2, \ldots\}$, which collectively are denoted simply **a**; node **B** has states $\{b_1, b_2, \ldots\}$, denoted **b**, and so forth. The links between nodes represent direct causal influence. For example the link from **A** to **D** represents the direct influence of **A** upon **D**. In this network, the variables at **B** may influence those at **D**, but only indirectly through their effect on **C**. Simple probabilities are denoted $P(\mathbf{a})$ and $P(\mathbf{b})$, and conditional probabilities $P(\mathbf{c}|\mathbf{b})$, $P(\mathbf{d}|\mathbf{a}, \mathbf{c})$ and $P(\mathbf{e}|\mathbf{d})$.

CONDITIONAL PROBABILITY TABLE

the joint distribution. To proceed, though, we also need the *conditional probability tables*, which give the probability of any variable at a node for each conditioning event—that is, for the values of the variables in the parent nodes. Each row in a conditional probability table sums to 1, as its entries describe all possible cases for the variable. If a node has no parents, then the table just contains the prior probabilities of the variables. (There are sophisticated algorithms for learning the entries in such a table based on a data set of variable values. We shall not address such learning here as our main concern is how to represent and reason about this probabilistic information.) Since the network and conditional probability tables contain all the information of the problem domain, we can use them to calculate any entry in the joint probability distribution, as illustrated in Example 4.

EXAMPLE 4 Belief Network for Fish

Consider again the problem of classifying fish, but now we want to incorporate more information than the measurements of the lightness and width. Imagine that a human expert has constructed the simple belief network in the figure, where node **A** represents the time of year and can have four values: a_1 = *winter*, a_2 = *spring*, a_3 = *summer* and a_4 = *autumn*. Node **B** represents the locale where the fish was caught: b_1 = *north Atlantic* and b_2 = *south Atlantic*. Node **X**, which represents the fish, has just two possible values: x_1 = *salmon* and x_2 = *sea bass*. **A** and **B** are the parents of the **X**. Similarly, our expert tells us that the children nodes of **X** represent lightness, **C**, with c_1 = *dark*, c_2 = *medium* and c_3 = *light*, as well as thickness, **D**, with d_1 = *thick* and d_2 = *thin*. Thus the season and the locale determine directly what kind of fish is likely to be caught; the season and locale also determine the fish's lightness and thickness, but only indirectly through their effect on **X**.



P(a)

$P(a_1)$	$P(a_2)$	$P(a_3)$	$P(a_4)$
0.25	0.25	0.25	0.25

P(b)

$P(b_1)$	$P(b_2)$
0.6	0.4

P(x|a,b)

	$P(x_1\|a_i, b_j)$	$P(x_2\|a_i, b_j)$
a_1, b_1	0.5	0.5
a_1, b_2	0.7	0.3
a_2, b_1	0.6	0.4
a_2, b_2	0.8	0.2
a_3, b_1	0.4	0.6
a_3, b_2	0.1	0.9
a_4, b_1	0.2	0.8
a_4, b_2	0.3	0.7

P(c|x)

	$P(c_1\|x_i)$	$P(c_2\|x_i)$	$P(c_3\|x_i)$
x_1	0.6	0.2	0.2
x_2	0.2	0.3	0.5

P(d|x)

	$P(d_1\|x_i)$	$P(d_2\|x_i)$
x_1	0.3	0.7
x_2	0.6	0.4

A simple belief net for the fish example. The season and the locale (and many other stochastic variables) determine directly the probability of the two different types of fish caught. The type of fish directly affects the lightness and thickness measured. The conditional probability tables quantifying these relationships are shown in pink.

Imagine that fishing boats go out throughout the year; then the probability distribution on the variables at **A** is uniform. Imagine, too, that boats generally spend more time in the north than the south Atlantic areas, specifically the probabilities that any fish came from those areas are 0.6 and 0.4, respectively. The other conditional probabilities are similarly given in the tables.

Now we can determine the value of any entry in the joint probability, for instance the probability that the fish was caught in the summer in the north Atlantic and is a sea bass that is dark and thin:

$$\begin{aligned} P(a_3, b_1, x_2, c_3, d_2) &= P(a_3)P(b_1)P(x_2|a_3, b_1)P(c_3|x_2)P(d_2|x_2) \\ &= 0.25 \times 0.6 \times 0.6 \times 0.5 \times 0.4 \\ &= 0.018. \end{aligned}$$

Note how the topology of the net is captured by the probabilities in the expression. Specifically, since **X** is the only node to have two parents, only the $P(x_2|\cdot, \cdot)$ term has two conditioning variables; the other conditional probabilities have just one each. The product of these probabilities corresponds to the assumption of statistical independence.

We now illustrate more fully how to exploit the causal structure in a Bayes belief net when determining the probability of its variables. Suppose we wish to determine the probability distribution over the variables $d_1, d_2, \ldots$ at **D** in the left network of Fig. 2.25 using the conditional probability tables and the network topology. We evaluate this by summing the full joint distribution, $P(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d})$, over all the variables other than **d**:

$$\begin{aligned} P(\mathbf{d}) &= \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a}, \mathbf{b}, \mathbf{c}, \mathbf{d}) \\ &= \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a})P(\mathbf{b}|\mathbf{a})P(\mathbf{c}|\mathbf{b})P(\mathbf{d}|\mathbf{c}) \\ &= \underbrace{\sum_{\mathbf{c}} P(\mathbf{d}|\mathbf{c}) \underbrace{\sum_{\mathbf{b}} P(\mathbf{c}|\mathbf{b}) \underbrace{\sum_{\mathbf{a}} P(\mathbf{b}|\mathbf{a})P(\mathbf{a})}_{P(\mathbf{b})}}_{P(\mathbf{c})}}_{P(\mathbf{d})}. \end{aligned} \tag{96}$$

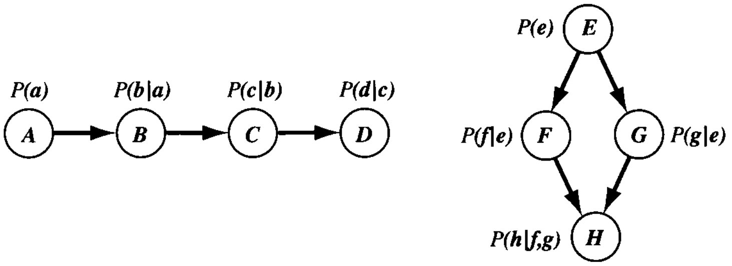


FIGURE 2.25. Two simple belief networks. The one on the left is a simple linear chain, the one on the right a simple loop. The conditional probability tables are indicated, for instance, as $P(\mathbf{h}|\mathbf{f}, \mathbf{g})$.

In Eq. 96 the summation variables can be split simply, and the intermediate terms have simple interpretations, as indicated. If we wanted the probability of a *particular* value of **D**, for instance d_2, we would compute

$$P(d_2) = \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a}, \mathbf{b}, \mathbf{c}, d_2), \tag{97}$$

and proceed as above. In either case, the conditional probabilities are simple because of the simple linear topology of the network.

Now consider computing the probabilities of the variables at **H** in the network with the loop on the right of Fig. 2.25. Here we find

$$\begin{aligned} P(\mathbf{h}) &= \sum_{\mathbf{e},\mathbf{f},\mathbf{g}} P(\mathbf{e}, \mathbf{f}, \mathbf{g}, \mathbf{h}) \\ &= \sum_{\mathbf{e},\mathbf{f},\mathbf{g}} P(\mathbf{e})P(\mathbf{f}|\mathbf{e})P(\mathbf{g}|\mathbf{e})P(\mathbf{h}|\mathbf{f}, \mathbf{g}) \\ &= \sum_{\mathbf{e}} P(\mathbf{e})P(\mathbf{f}|\mathbf{e})P(\mathbf{g}|\mathbf{e}) \sum_{\mathbf{f},\mathbf{g}} P(\mathbf{h}|\mathbf{f}, \mathbf{g}). \end{aligned} \tag{98}$$

Note particularly that the expansion of the full sum differs somewhat from that in Eq. 96 because of the $P(\mathbf{h}|\mathbf{f}, \mathbf{g})$ term, which itself arises from the loop topology of the network.

EVIDENCE

Bayes belief nets are most useful in the case where are given the values of some of the variables—the *evidence*—and we seek to determine some particular configuration of other variables. Thus in our fish example we might seek to determine the probability that a fish came from the north Atlantic, given that it is springtime, and that the fish is a light salmon. (Notice that even here we may not be given the values of some variables such as the width of the fish.) In that case, the probability we seek is $P(b_1|a_2, x_1, c_1)$. In practice, we determine the values of several query variables (denoted collectively **x**) given the evidence of all other variables (denoted **e**) by

$$P(\mathbf{x}|\mathbf{e}) = \frac{P(\mathbf{x}, \mathbf{e})}{P(\mathbf{e})} = \alpha P(\mathbf{x}, \mathbf{e}), \tag{99}$$

where α is a constant of proportionality.

As an example, suppose we are given the Bayes belief net and conditional probability tables in Example 4. Suppose we know that a fish is light (c_1) and caught in the south Atlantic (b_2), but we do not know what time of year the fish was caught nor it thickness. How shall we classify the fish for minimum expected classification error? Of course we must compute the probability it is a salmon, and also the probability it is sea bass. We focus first on the relative probability the fish is a salmon given this evidence:

$$\begin{aligned} P(x_1|c_1, b_2) &= \frac{P(x_1, c_1, b_2)}{P(c_1, b_2)} \\ &= \alpha \sum_{\mathbf{a},\mathbf{d}} P(x_1, \mathbf{a}, b_2, c_1, \mathbf{d}) \end{aligned} \tag{100}$$

$$= \alpha \sum_{\mathbf{a},\mathbf{d}} P(\mathbf{a})P(b_2)P(x_1|\mathbf{a},b_2)P(c_1|x_1)P(\mathbf{d}|x_1)$$

$$= \alpha P(b_2)P(c_1|x_1)$$

$$\times \left[\sum_{\mathbf{a}} P(\mathbf{a})P(x_1|\mathbf{a},b_2)\right]\left[\sum_{\mathbf{d}} P(\mathbf{d}|x_1)\right]$$

$$= \alpha P(b_2)P(c_1|x_1)$$

$$\times[P(a_1)P(x_1|a_1,b_2) + P(a_2)P(x_1|a_2,b_2)$$

$$+ P(a_3)P(x_1|a_3,b_2) + P(a_4)P(x_1|a_4,b_2)]$$

$$\times \underbrace{[P(d_1|x_1) + P(d_2|x_1)]}_{=1}$$

$$= \alpha(0.4)(0.6)[(0.25)(0.7) + (0.25)(0.8) + (0.25)(0.1)$$

$$+ (0.25)(0.3)]1.0$$

$$= \alpha\, 0.114.$$

Note that in this case,

$$\sum_{\mathbf{d}} P(\mathbf{d}|x_1) = 1, \tag{101}$$

that is, if we do not measure information corresponding to node **D**, the conditional probability table at **D** does not affect our results. A computation similar to that in Eq. 100 shows $P(x_2|c_1, b_2) = \alpha\ 0.042$. We normalize these probabilities (and hence eliminate α) and find $P(x_1|c_1, b_2) = 0.73$ and $P(x_2|c_1, b_2) = 0.27$. Thus given this evidence, we should classify this fish as a salmon.

When the dependency relationships among the features used by a classifier are unknown, we generally proceed by taking the simplest assumption, namely, that the features are conditionally independent given the category, that is,

$$P(\mathbf{a}, \mathbf{b}|\mathbf{x}) = P(\mathbf{a}|\mathbf{x})P(\mathbf{b}|\mathbf{x}). \tag{102}$$

NAIVE BAYES' RULE

In practice, this so-called *naive Bayes' rule* or *idiot Bayes' rule* often works quite well in practice, despite its manifest simplicity. Other approaches are to assume some functional form of conditional probability tables.

Belief nets have found increasing use in complicated problems such as medical diagnosis. Here the uppermost nodes (ones without their own parents) represent a fundamental biological agent such as the presence of a virus or bacteria. Intermediate nodes then describe diseases, such as flu or emphysema, and the lowermost nodes describe the symptoms, such as high temperature or coughing. A physician enters measured values into the net and finds the most likely disease or cause.

*2.12 COMPOUND BAYESIAN DECISION THEORY AND CONTEXT

Let us reconsider our introductory example of designing a classifier to sort two types of fish. Our original assumption was that the sequence of types of fish was so unpredictable that the state of nature looked like a random variable. Without abandoning this attitude, let us consider the possibility that the consecutive states of nature might not be statistically independent. We should be able to exploit such statistical dependence to gain improved performance. This is one example of the use of *context* to aid decision making.

The way in which we exploit such context information is somewhat different when we can wait for n fish to emerge and then make all n decisions jointly than when we must decide as each fish emerges. The first problem is a *compound decision problem*, and the second is a *sequential compound decision problem*. The former case is conceptually simpler and is the one we shall examine here.

To state the general problem, let $\boldsymbol{\omega} = (\omega(1), \ldots, \omega(n))^t$ be a vector denoting the n states of nature, with $\omega(i)$ taking on one of the c values $\omega_1, \ldots, \omega_c$. Let $P(\boldsymbol{\omega})$ be the prior probability for the n states of nature. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ be a matrix giving the n observed feature vectors, with $\mathbf{x}_i$ being the feature vector obtained when the state of nature was $\omega(i)$. Finally, let $p(\mathbf{X}|\boldsymbol{\omega})$ be the conditional probability density function for $\mathbf{X}$ given the true set of states of nature $\boldsymbol{\omega}$. Using this notation we see that the posterior probability of $\boldsymbol{\omega}$ is given by

$$P(\boldsymbol{\omega}|\mathbf{X}) = \frac{p(\mathbf{X}|\boldsymbol{\omega})P(\boldsymbol{\omega})}{p(\mathbf{X})} = \frac{p(\mathbf{X}|\boldsymbol{\omega})P(\boldsymbol{\omega})}{\sum_{\boldsymbol{\omega}} p(\mathbf{X}|\boldsymbol{\omega})P(\boldsymbol{\omega})}. \tag{103}$$

In general, one can define a loss matrix for the compound decision problem and seek a decision rule that minimizes the compound risk. The development of this theory parallels our discussion for the simple decision problem, and concludes that the optimal procedure is to minimize the compound conditional risk. In particular, if there is no loss for being correct and if all errors are equally costly, then the procedure reduces to computing $P(\boldsymbol{\omega}|\mathbf{X})$ for all $\boldsymbol{\omega}$ and selecting the $\boldsymbol{\omega}$ for which this posterior probability is maximum.

While this provides the theoretical solution, in practice the computation of $P(\boldsymbol{\omega}|\mathbf{X})$ can easily prove to be an enormous task. If each component $\omega(i)$ can have one of c values, there are c^n possible values of $\boldsymbol{\omega}$ to consider. Some simplification can be obtained if the distribution of the feature vector $\mathbf{x}_i$ depends only on the corresponding state of nature $\omega(i)$, not on the values of the other feature vectors or the other states of nature. In this case the joint density $p(\mathbf{X}|\boldsymbol{\omega})$ is merely the product of the component densities $p(\mathbf{x}_i|\omega(i))$:

$$p(\mathbf{X}|\boldsymbol{\omega}) = \prod_{i=1}^{n} p(\mathbf{x}_i|\omega(i)). \tag{104}$$

While this simplifies the problem of computing $p(\mathbf{X}|\boldsymbol{\omega})$, there is still the problem of computing the prior probabilities $P(\boldsymbol{\omega})$. This joint probability is central to the compound Bayes decision problem, because it reflects the interdependence of the states of nature. Thus it is unacceptable to simplify the problem of calculating $P(\boldsymbol{\omega})$ by assuming that the states of nature are independent. In addition, practical applications usually require some method of avoiding the computation of $P(\boldsymbol{\omega}|\mathbf{X})$ for all c^n possible values of $\boldsymbol{\omega}$. We shall find some solutions to this problem in Chapter 3.

SUMMARY

The basic ideas underlying Bayes decision theory are very simple. To minimize the overall risk, one should always choose the action that minimizes the conditional risk $R(\alpha|\mathbf{x})$. In particular, to minimize the probability of error in a classification problem, one should always choose the state of nature that maximizes the posterior probability $P(\omega_j|\mathbf{x})$. Bayes formula allows us to calculate such probabilities from the prior probabilities $P(\omega_j)$ and the conditional densities $p(\mathbf{x}|\omega_j)$. If there are different penalties for misclassifying patterns from ω_i as if from ω_j, the posteriors must be first weighted according to such penalties before taking action.

If the underlying distributions are multivariate Gaussian, the decision boundaries will be hyperquadrics, whose form and position depends upon the prior probabilities, means and covariances of the distributions in question. The true expected error can be bounded above by the Chernoff and computationally simpler Bhattacharyya bounds. If an input (test) pattern has missing or corrupted features, we should form the marginal distributions by integrating over such features and then using Bayes decision procedure on the resulting distributions. Receiver operating characteristic curves describe the inherent and unchangeable properties of a classifier and can be used, for example, to determine the Bayes rate.

Bayesian belief nets allow the designer to specify, by means of connection topology, the functional dependencies and independencies among model variables. When any subset of variables is clamped to some known values, each node comes to a probability of its value through a Bayesian inference calculation. Parameters representing conditional dependencies can be set by an expert.

For many pattern classification applications, the chief problem in applying these results is that the conditional densities $p(\mathbf{x}|\omega_j)$ are not known. In some cases we may know the form these densities assume, but we may not know characterizing parameter values. The classic case occurs when the densities are known to be, or can assumed to be, multivariate normal, but the values of the mean vectors and the covariance matrices are not known. More commonly, even less is known about the conditional densities, and procedures that are less sensitive to specific assumptions about the densities must be used. Most of the remainder of this book will be devoted to various procedures that have been developed to attack such problems.

BIBLIOGRAPHICAL AND HISTORICAL REMARKS

The power, coherence, and elegance of Bayesian theory in pattern recognition make it among the most beautiful formalisms in science. Its foundations go back to the Reverend Bayes himself, of course [3], but he stated his theorem (Eq. 1) for the case of uniform priors. It was Laplace [29] who first stated it for the more general (but discrete) case. There are several modern and clear descriptions of the ideas—in pattern recognition and general decision theory—that can be recommended [6, 7, 15, 17, 30, 31]. Because Bayesian theory rests on an axiomatic foundation, it is guaranteed to have quantitative coherence; some other classification methods do not. Wald presents a non-Bayesian perspective on these topics that can be highly recommended [41], and the philosophical foundations of Bayesian and non-Bayesian methods are explored in reference [18]. Neyman and Pearson provided some of the most important pioneering work in hypothesis testing, and they used the probability of error as the criterion [32]; Wald extended this work by introducing the notions of loss and risk [40]. Certain conceptual problems have always attended the use of loss functions and prior probabilities. In fact, the Bayesian approach is avoided by many statisticians, partly because there are problems for which a decision is made only once, and partly because there may be no reasonable way to determine the prior probabilities. Neither of these difficulties seems to present a serious drawback in typical pattern recognition applications: For nearly all important pattern recognition problems we will have training data and we will use our recognizer more than once. For these reasons, the Bayesian approach will continue to be of great use in pattern recognition. The single most important drawback of the Bayesian approach is the difficulty of determining and computing the conditional density functions. The multivariate Gaussian model may provide an adequate approximation to the true density, but there are many problems for which the densitites are far from Gaussian. Even when the Gaussian model is satisfactory, we shall see in the next chapter that estimating the unknown parameters from data may not be a trivial task. Subsequent chapters will investigate what can be done when the Gaussian model is not adequate.

Chow was among the earliest to use Bayesian decision theory for pattern recognition [12], and he later established fundamental relations between error and reject rate [13]. Error rates for Gaussians have been explored in reference [20], and the Chernoff and Bhattacharyya bounds were first presented in references [11] and [8], respectively; the bounds are explored in a number of statistics texts, such as reference [19]. Computational approximations for bounding integrals for Bayesian probability of error (the source for one of the homework problems) appears in reference [2]. Neyman and Pearson also worked on classification given constraints [32], and the analysis of minimax estimators for multivariate normals is presented in references [4, 5] and [16]. Signal detection theory and receiver operating characteristics are fully explored in reference [22]; a brief overview, targetting experimental psychologists, is presented in reference [39]. Our discussion of the missing feature problem follows closely the work of Ahmad and Tresp [1], while the definitive book on missing features, including a great deal beyond our discussion here, is reference [35].

The origins of Bayesian belief nets can be traced back to reference [43], and a thorough literature review can be found in reference [10]; excellent modern books [27, 33] and tutorials [9] can be recommended. An important dissertation on the theory of belief nets, with an application to medical diagnosis, is reference [25], and a summary of work on diagnosis of machine faults is given in reference [24]. While we have focussed on directed acyclic graphs, belief nets are of broader use, and they even allow loops or arbitrary topologies—a topic that would lead us far afield here but which is treated in reference [27].

Entropy was the central concept in the foundation of information theory [36], and the relation of Gaussians to entropy is explored in reference [38]. Readers requiring a review of information theory [14], linear algebra [28], calculus and continuous mathematics [37, 44], probability [34], calculus of variations and Lagrange multipliers [21], should consult these texts and those listed in our Appendix.

PROBLEMS

Section 2.1

1. In the two-category case, under the Bayes decision rule the conditional error is given by Eq. 7. Even if the posterior densities are continuous, this form of the conditional error virtually always leads to a discontinuous integrand when calculating the full error by Eq. 5.

(a) Show that for arbitrary densities, we can replace Eq. 7 by $P(error|x) = 2P(\omega_1|x)P(\omega_2|x)$ in the integral and get an upper bound on the full error.

(b) Show that if we use $P(error|x) = \alpha P(\omega_1|x)P(\omega_2|x)$ for $\alpha < 2$, then we are not guaranteed that the integral gives an upper bound on the error.

(c) Analogously, show that we can use instead $P(error|x) = P(\omega_1|x)P(\omega_2|x)$ and get a lower bound on the full error.

(d) Show that if we use $P(error|x) = \beta P(\omega_1|x)P(\omega_2|x)$ for $\beta > 1$, then we are not guaranteed that the integral gives an lower bound on the error.

Section 2.2

2. Suppose two equally probable one-dimensional densities are of the form $p(x|\omega_i) \propto e^{-|x-a_i|/b_i}$ for $i = 1, 2$ and $0 < b_i$.

(a) Write an analytic expression for each density, that is, normalize each function for arbitrary a_i and positive b_i.

(b) Calculate the likelihood ratio as a function of your four variables.

(c) Sketch a graph of the likelihood ratio $p(x|\omega_1)/p(x|\omega_2)$ for the case $a_1 = 0$, $b_1 = 1$, $a_2 = 1$ and $b_2 = 2$.

Section 2.3

3. Consider minimax criterion for the zero-one loss function, that is, $\lambda_{11} = \lambda_{22} = 0$ and $\lambda_{12} = \lambda_{21} = 1$.

(a) Prove that in this case the decision regions will satisfy

$$\int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)\, d\mathbf{x} = \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)\, d\mathbf{x}.$$

(b) Is this solution always unique? If not, construct a simple counterexample.

4. Consider the minimax criterion for a two-category classification problem.

(a) Fill in the steps of the derivation of Eq. 23.

(b) Explain why the overall Bayes risk must be concave down as a function of the prior $P(\omega_1)$, as shown in Fig. 2.4.

(c) Assume we have one-dimensional Gaussian distributions $p(x|\omega_i) \sim N(\mu_i, \sigma_i^2)$, $i = 1, 2$, but completely unknown prior probabilities. Use the minimax criterion to find the optimal decision point x^* in terms of μ_i and σ_i under a zero-one risk.

(d) For the decision point x^* you found in (c), what is the overall minimax risk? Express this risk in terms of an error function $\mathrm{erf}(\cdot)$.

(e) Assume $p(x|\omega_1) \sim N(0, 1)$ and $p(x|\omega_2) \sim N(1/2, 1/4)$, under a zero-one loss. Find x^* and the overall minimax loss.

(f) Assume $p(x|\omega_1) \sim N(5, 1)$ and $p(x|\omega_2) \sim N(6, 1)$. Without performing any explicit calculations, determine x^* for the minimax criterion. Explain your reasoning.

5. Generalize the minimax decision rule in order to classify patterns from three categories having triangle densities as follows:

$$p(x|\omega_i) = T(\mu_i, \delta_i) \equiv \begin{cases} (\delta_i - |x - \mu_i|)/\delta_i^2 & \text{for } |x - \mu_i| < \delta_i \\ 0 & \text{otherwise,} \end{cases}$$

where $\delta_i > 0$ is the half-width of a distribution ($i = 1, 2, 3$). Assume for convenience that $\mu_1 < \mu_2 < \mu_3$, and make some minor simplifying assumptions about the δ_i's as needed, to answer the following:

(a) In terms of the priors $P(\omega_i)$, means and half-widths, find the optimal decision points x_1^* and x_2^* under a zero-one (categorization) loss.

(b) Generalize the minimax decision rule to *two* decision points, x_1^* and x_2^*, for such triangular distributions.

(c) Let $\{\mu_i, \delta_i\} = \{0, 1\}, \{.5, .5\}$, and $\{1, 1\}$. Find the minimax decision rule (i.e., x_1^* and x_2^*) for this case.

(d) What is the minimax risk for part (c)?

6. Consider the Neyman-Pearson criterion for two univariate normal distributions: $p(x|\omega_i) \sim N(\mu_i, \sigma_i^2)$ and $P(\omega_i) = 1/2$ for $i = 1, 2$. Assume a zero-one error loss, and for convenience let $\mu_2 > \mu_1$.

(a) Suppose the maximum acceptable error rate for classifying a pattern that is actually in ω_1 as if it were in ω_2 is E_1. Determine the single-point decision boundary in terms of the variables given.

(b) For this boundary, what is the error rate for classifying ω_2 as ω_1?

(c) What is the overall error rate under zero-one loss?

(d) Apply your results to the specific case $p(x|\omega_1) \sim N(-1, 1)$ and $p(x|\omega_2) \sim N(1, 1)$ and $E_1 = 0.05$.

(e) Compare your result to the Bayes error rate (i.e., without the Neyman-Pearson conditions).

7. Consider Neyman-Pearson criteria for two Cauchy distributions in one dimension:

$$p(x|\omega_i) = \frac{1}{\pi b} \cdot \frac{1}{1 + \left(\frac{x-a_i}{b}\right)^2}, \qquad i = 1, 2.$$

Assume a zero-one error loss, and for simplicity $a_2 > a_1$, the same "width" b, and equal priors.

(a) Suppose the maximum acceptable error rate for classifying a pattern that is actually in ω_1 as if it were in ω_2 is E_1. Determine the single-point decision boundary in terms of the variables given.

(b) For this boundary, what is the error rate for classifying ω_2 as ω_1?

(c) What is the overall error rate under zero-one loss?

(d) Apply your results to the specific case $b = 1$ and $a_1 = -1$, $a_2 = 1$ and $E_1 = 0.1$.

(e) Compare your result to the Bayes error rate (i.e., without the Neyman-Pearson conditions).

8. Let the conditional densities for a two-category one-dimensional problem be given by the Cauchy distribution described in Problem 7.

(a) By explicit integration, check that the distributions are indeed normalized.

(b) Assuming $P(\omega_1) = P(\omega_2)$, show that $P(\omega_1|x) = P(\omega_2|x)$ if $x = (a_1 + a_2)/2$, that is, the minimum error decision boundary is a point midway between the peaks of the two distributions, regardless of b.

(c) Plot $P(\omega_1|x)$ for the case $a_1 = 3$, $a_2 = 5$ and $b = 1$.

(d) How do $P(\omega_1|x)$ and $P(\omega_2|x)$ behave as $x \rightarrow -\infty$? $x \rightarrow +\infty$? Explain.

9. Use the conditional densities given in Problem 7, and assume equal prior probabilities for the categories.

(a) Show that the minimum probability of error is given by

$$P(error) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1}\left|\frac{a_2 - a_1}{2b}\right|.$$

(b) Plot this as a function of $|a_2 - a_1|/b$.

(c) What is the maximum value of $P(error)$ and under which conditions can this occur? Explain.

10. Consider the following decision rule for a two-category one-dimensional problem: Decide ω_1 if $x > \theta$; otherwise decide ω_2.

(a) Show that the probability of error for this rule is given by

$$P(error) = P(\omega_1)\int_{-\infty}^{\theta} p(x|\omega_1)\,dx + P(\omega_2)\int_{\theta}^{\infty} p(x|\omega_2)\,dx.$$

(b) By differentiating, show that a necessary condition to minimize $P(error)$ is that θ satisfies

$$p(\theta|\omega_1)P(\omega_1) = p(\theta|\omega_2)P(\omega_2).$$

(c) Does this equation define θ uniquely?

(d) Give an example where a value of θ satisfying the equation actually *maximizes* the probability of error.

11. Suppose that we replace the deterministic decision function $\alpha(\mathbf{x})$ with a *randomized rule*, namely, one giving the probability $P(\alpha_i|\mathbf{x})$ of taking action α_i upon observing $\mathbf{x}$.

(a) Show that the resulting risk is given by

$$R = \int \left[\sum_{i=1}^{a} R(\alpha_i|\mathbf{x}) P(\alpha_i|\mathbf{x}) \right] p(\mathbf{x})\, d\mathbf{x}.$$

(b) In addition, show that R is minimized by choosing $P(\alpha_i|\mathbf{x}) = 1$ for the action α_i associated with the minimum conditional risk $R(\alpha_i|\mathbf{x})$, thereby showing that no benefit can be gained from randomizing the best decision rule.

(c) Can we benefit from randomizing a suboptimal rule? Explain.

12. Let $\omega_{max}(\mathbf{x})$ be the state of nature for which $P(\omega_{max}|\mathbf{x}) \geq P(\omega_i|\mathbf{x})$ for all i, $i = 1, \ldots, c$.

(a) Show that $P(\omega_{max}|\mathbf{x}) \geq 1/c$.

(b) Show that for the minimum-error-rate decision rule the average probability of error is given by

$$P(error) = 1 - \int P(\omega_{max}|\mathbf{x}) p(\mathbf{x})\, d\mathbf{x}.$$

(c) Use these two results to show that $P(error) \leq (c-1)/c$.

(d) Describe a situation for which $P(error) = (c-1)/c$.

Section 2.4

13. In many pattern classification problems one has the option either to assign the pattern to one of c classes, or to *reject* it as being unrecognizable. If the cost for rejects is not too high, rejection may be a desirable action. Let

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \quad i, j = 1, \ldots, c \\ \lambda_r & i = c+1 \\ \lambda_s & \text{otherwise,} \end{cases}$$

where λ_r is the loss incurred for choosing the $(c+1)$th action, rejection, and λ_s is the loss incurred for making any substitution error. Show that the minimum risk is obtained if we decide ω_i if $P(\omega_i|\mathbf{x}) \geq P(\omega_j|\mathbf{x})$ for all j and if $P(\omega_i|\mathbf{x}) \geq 1 - \lambda_r/\lambda_s$, and reject otherwise. What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$?

14. Consider the classification problem with rejection option.

(a) Use the results of Problem 13 to show that the following discriminant functions are optimal for such problems:

$$g_i(\mathbf{x}) = \begin{cases} p(\mathbf{x}|\omega_i)P(\omega_i) & i = 1, \ldots, c \\ \frac{\lambda_s - \lambda_r}{\lambda_s} \sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j) & i = c+1. \end{cases}$$

(b) Plot these discriminant functions and the decision regions for the two-category one-dimensional case having

- $p(x|\omega_1) \sim N(1, 1)$,
- $p(x|\omega_2) \sim N(-1, 1)$,
- $P(\omega_1) = P(\omega_2) = 1/2$, and
- $\lambda_r/\lambda_s = 1/4$.

(c) Describe qualitatively what happens as λ_r/λ_s is increased from 0 to 1.

(d) Repeat for the case having

- $p(x|\omega_1) \sim N(1, 1)$,
- $p(x|\omega_2) \sim N(0, 1/4)$,
- $P(\omega_1) = 1/3$, $P(\omega_2) = 2/3$, and
- $\lambda_r/\lambda_s = 1/2$.

Section 2.5

15. Confirm Eq. 47 for the volume of a d-dimensional hypersphere as follows:

(a) Verify that the equation is correct for a line segment ($d = 1$).

(b) Verify that the equation is correct for a disk ($d = 2$).

(c) Integrate the volume of a line over appropriate limits to obtain the volume of a disk.

(d) Consider a general d-dimensional hypersphere. Integrate its volume to obtain a formula (involving the ratio of gamma functions, $\Gamma(\cdot)$) for the volume of a $(d + 1)$-dimensional hypersphere.

(e) Apply your formula to find the volume of a hypersphere in an odd-dimensional space by integrating the volume of a hypersphere in the lower even-dimensional space, and thereby confirm Eq. 47 for odd dimensions.

(f) Repeat the above but for finding the volume of a hypersphere in even dimensions.

16. Derive the formula for the volume of a d-dimensional hypersphere in Eq. 47 as follows:

(a) State by inspection the formula for V_1.

(b) Follow the general procedure outlined in Problem 15 and integrate twice to find V_{d+2} as a function of V_d.

(c) Assume that the functional form of V_d is the same for all odd dimensions (and likewise for all even dimensions). Use your integration results to determine the formula for V_d for d odd.

(d) Use your intermediate integration results to determine V_d for d even.

(e) Explain why we should expect the functional form of V_d to be different in even and in odd dimensions.

17. Derive the formula (Eq. 46) for the volume V of a hyperellipsoid of constant Mahalanobis distance r (Eq. 45) for a Gaussian distribution having covariance $\mathbf{\Sigma}$.

18. Consider two normal distributions in one dimension: $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$. Imagine that we choose two random samples x_1 and x_2, one from each of the normal distributions and calculate their sum $x_3 = x_1 + x_2$. Suppose we do this repeatedly.

(a) Consider the resulting distribution of the values of x_3. Show that x_3 possesses the requisite statistical properties and thus its distribution is normal.

(b) What is the mean, μ_3, of your new distribution?

(c) What is the variance, σ_3^2?

(d) Repeat the above with two distributions in a multi-dimensional space, i.e., $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$.

19. Starting from the definition of entropy (Eq. 37), derive the general equation for the maximum-entropy distribution given constraints expressed in the general form

$$\int b_k(x)p(x)\,dx = a_k, \qquad k = 1, 2, \ldots, q$$

as follows:

(a) Use Lagrange undetermined multipliers $\lambda_1, \lambda_2, \ldots, \lambda_q$ and derive the synthetic function:

$$H_s = -\int p(x)\left[\ln p(x) - \sum_{k=0}^{q} \lambda_k b_k(x)\right] dx - \sum_{k=0}^{q} \lambda_k a_k.$$

State why we know $a_0 = 1$ and $b_0(x) = 1$ for all x.

(b) Take the derivative of H_s with respect to $p(x)$. Set the integrand to zero, and thereby prove that the minimum-entropy distribution obeys

$$p(x) = \exp\left[\sum_{k=0}^{q} \lambda_k b_k(x) - 1\right],$$

where the $q + 1$ parameters are determined by the constraint equation.

20. Use the final result from Problem 19 for the following.

(a) Suppose we know solely that a distribution is nonzero only in the range $x_l \le x \le x_u$. Prove that the maximum entropy distribution is uniform in that range, that is,

$$p(x) \sim U(x_l, x_u) = \begin{cases} 1/|x_u - x_l| & x_l \le x \le x_u \\ 0 & \text{otherwise.} \end{cases}$$

(b) Suppose we know solely that a distribution is nonzero only for $x \ge 0$ and that its mean is μ. Prove that the maximum entropy distribution is

$$p(x) = \begin{cases} \frac{1}{\mu}e^{-x/\mu} & \text{for } x \ge 0 \\ 0 & \text{otherwise.} \end{cases}$$

(c) Now suppose we know solely that the distribution is normalized, has mean μ, and standard deviation σ, and thus from Problem 19 our maximum entropy distribution must be of the form

$$p(x) = \exp[\lambda_0 - 1 + \lambda_1 x + \lambda_2 x^2].$$

Write out the three constraints and solve for λ_0, λ_1, and λ_2 and thereby prove that the maximum entropy solution is a Gaussian, that is,

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right].$$

21. Three distributions—a Gaussian, a uniform distribution, and a triangle distribution (cf. Problem 5)—each have mean zero and standard deviation σ. Use Eq. 37 to calculate and compare their entropies.

22. Calculate the entropy of a multidimensional Gaussian $p(\mathbf{x}) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

23. Consider the three-dimensional normal distribution $p(\mathbf{x}|\omega) \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \\ 2 \end{pmatrix} \text{ and } \boldsymbol{\Sigma} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 2 \\ 0 & 2 & 5 \end{pmatrix}.$$

(a) Find the probability density at the point $\mathbf{x}_0 = (.5, 0, 1)^t$.

(b) Construct the whitening transformation $\mathbf{A}_w$ (Eq. 44). Compute the matrices representing eigenvectors and eignvalues, $\boldsymbol{\Phi}$ and $\boldsymbol{\Lambda}$. Next, convert the distribution to one centered on the origin with covariance matrix equal to the identity matrix, $p(\mathbf{x}|\omega) \sim N(\mathbf{0}, \mathbf{I})$.

(c) Apply the same overall transformation to $\mathbf{x}_0$ to yield a transformed point $\mathbf{x}_w$.

(d) By explicit calculation, confirm that the Mahalanobis distance from $\mathbf{x}_0$ to the mean $\boldsymbol{\mu}$ in the original distribution is the same as for $\mathbf{x}_w$ to $\mathbf{0}$ in the transformed distribution.

(e) Does the probability density remain unchanged under a general linear transformation? In other words, is $p(\mathbf{x}_0|N(\boldsymbol{\mu}, \boldsymbol{\Sigma})) = p(\mathbf{T}^t\mathbf{x}_0|N(\mathbf{T}^t\boldsymbol{\mu}, \mathbf{T}^t\boldsymbol{\Sigma}\mathbf{T}))$ for some linear transform $\mathbf{T}$? Explain.

(f) Prove that a general whitening transform $\mathbf{A}_w = \boldsymbol{\Phi}\boldsymbol{\Lambda}^{-1/2}$ when applied to a Gaussian distribution ensures that the final distribution has covariance proportional to the identity matrix $\mathbf{I}$. Check whether normalization is preserved by the transformation.

24. Consider the multivariate normal density with mean $\boldsymbol{\mu}$, $\sigma_{ij} = 0$ and $\sigma_{ii} = \sigma_i^2$, that is, the covariance matrix is diagonal: $\boldsymbol{\Sigma} = diag(\sigma_1^2, \sigma_2^2, \ldots, \sigma_d^2)$.

(a) Show that the evidence is

$$p(\mathbf{x}) = \frac{1}{\prod\limits_{i=1}^{d}\sqrt{2\pi}\sigma_i}\exp\left[-\frac{1}{2}\sum_{i=1}^{d}\left(\frac{x_i-\mu_i}{\sigma_i}\right)^2\right].$$

(b) Plot and describe the contours of constant density.

(c) Write an expression for the Mahalanobis distance from $\mathbf{x}$ to $\boldsymbol{\mu}$.

Section 2.6

25. Fill in the steps in the derivation from Eq. 59 to Eqs. 60–65.

26. Let $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$ for $i = 1, 2$ in a two-category d-dimensional problem with the same covariances but arbitrary means and prior probabilities. Consider the squared Mahalanobis distance

$$r_i^2 = (\mathbf{x} - \boldsymbol{\mu}_i)^t \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i).$$

(a) Show that the gradient of r_i^2 is given by

$$\nabla r_i^2 = 2\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i).$$

(b) Show that at any position on a given line through $\boldsymbol{\mu}_i$ the gradient ∇r_i^2 points in the same direction. Must this direction be parallel to that line?

(c) Show that ∇r_1^2 and ∇r_2^2 point in opposite directions along the line from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$.

(d) Show that the optimal separating hyperplane is tangent to the constant probability density hyperellipsoids at the point that the separating hyperplane cuts the line from $\boldsymbol{\mu}_1$ to $\boldsymbol{\mu}_2$.

(e) True or False: For a two-category problem involving normal densities with arbitrary means and covariances, and $P(\omega_1) = P(\omega_2) = 1/2$, the Bayes decision boundary consists of the set of points of equal Mahalanobis distance from the respective sample means. Explain.

27. Suppose we have two normal distributions with the same covariances but different means: $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$. In terms of their prior probabilities $P(\omega_1)$ and $P(\omega_2)$, state the condition that the Bayes decision boundary *not* pass between the two means.

28. Two random variables $\mathbf{x}$ and $\mathbf{y}$ are called statistically independent if $p(\mathbf{x}, \mathbf{y}|\omega) = p(\mathbf{x}|\omega)p(\mathbf{y}|\omega)$.

(a) Prove that if $x_i - \mu_i$ and $x_j - \mu_j$ are statistically independent (for $i \neq j$), then σ_{ij} as defined in Eq. 43 is 0.

(b) Prove that the converse is true for the Gaussian case.

(c) Show by counterexample that this converse is *not* true in the general case.

29. Figure 2.15 shows that it is possible for a decision boundary for two three-dimensional Gaussians to be a line segment. Explain how this can arise by analyzing a simpler one-dimensional case as follows.

(a) Consider two one-dimensional Gaussians whose means differ and whose variances differ. Explain why for this case we can always find prior probabilities such that the decision boundary is a single point.

(b) Use your result to explain how the three-dimensional two-Gaussian case can yield a line segment decision boundary.

30. Consider the Bayes decision boundary for two-category classification in d dimensions.

(a) Prove that for any arbitrary hyperquadric in d dimensions, there exist normal distributions $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$ and priors $P(\omega_i)$, $i = 1, 2$, that possess this hyperquadric as their Bayes decision boundary.

(b) Is your answer to part (a) true if the priors are held fixed and nonzero, e.g., $P(\omega_1) = P(\omega_2) = 1/2$?

Section 2.7

31. Let $p(x|\omega_i) \sim N(\mu_i, \sigma^2)$ for a two-category one-dimensional problem with $P(\omega_1) = P(\omega_2) = 1/2$.

(a) Show that the minimum probability of error is given by

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

where $a = |\mu_2 - \mu_1|/(2\sigma)$.

(b) Use the inequality

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^\infty e^{-t^2/2}\, dt \le \frac{1}{\sqrt{2\pi}a} e^{-a^2/2}$$

to show that P_e goes to zero as $|\mu_2 - \mu_1|/\sigma$ goes to infinity.

32. Let $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \sigma^2\mathbf{I})$ for a two-category d-dimensional problem with $P(\omega_1) = P(\omega_2) = 1/2$.

(a) Show that the minimum probability of error is given by

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

where $a = \|\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1\|/(2\sigma)$.

(b) Let $\boldsymbol{\mu}_1 = \mathbf{0}$ and $\boldsymbol{\mu}_2 = (\mu_1, \ldots, \mu_d)^t \neq \mathbf{0}$. Use the inequality from Problem 31 to show that P_e approaches zero as the dimension d approaches infinity.

(c) Express the meaning of this result in words.

33. Suppose we know exactly two arbitrary distributions $p(\mathbf{x}|\omega_i)$ and priors $P(\omega_i)$ in a d-dimensional feature space.

(a) Prove that the true error cannot decrease if we first project the distributions to a lower-dimensional space and then classify them.

(b) Despite this fact, suggest why in an actual pattern recognition application we might not want to include an arbitrarily high number of feature dimensions.

Section 2.8

34. Show that if the densities in a two-category classification problem differ significantly from Gaussian, the Chernoff and Bhattacharyya bounds are not likely to be informative by considering the following one-dimensional examples. Consider a number of problems in which the mean and variance are the same (and thus the Chernoff bound and the Bhattacharyya bound remain the same), but nevertheless have a wide range in Bayes error. For definiteness, assume $P(\omega_1) = P(\omega_2) = 0.5$ and the distributions have means at $\mu_1 = -\mu$ and $\mu_2 = +\mu$, and $\sigma_1^2 = \sigma_2^2 = \mu^2$.

(a) Use the equations in the text to calculate the Chernoff and the Bhattacharyya bounds on the error.

(b) Suppose the distributions are both Gaussian. Calculate explicitly the Bayes error. Express it in terms of an error function erf(·) and as a numerical value.

(c) Now consider another case, in which half the density for ω_1 is concentrated at a point $x = -2\mu$ and half at $x = 0$; likewise (symmetrically) the density for ω_2 has half its mass at $x = +2\mu$ and half at $x = 0$. Show that the means and variance remain as desired, but that now the Bayes error is 0.25.

(d) Now consider yet another case, in which half the density for ω_1 is concentrated near $x = -2$ and half at $x = -\epsilon$, where ϵ is an infinitessimally small positive distance; likewise (symmetrically) the density for ω_2 has half its mass near $x = +2\mu$ and half at $+\epsilon$. Show that by making ϵ sufficiently small, the means and variances can be made arbitrarily close to μ and μ^2, respectively. Show, too, that now the Bayes error is zero.

(e) Compare your errors in (b), (c), and (d) to your Chernoff and Bhattacharyya bounds of (a) and explain in words why those bounds are unlikely to be of much use if the distributions differ markedly from Gaussians.

35. Show for nonpathological cases that if we include more feature dimensions in a Bayesian classifier for multidimensional Gaussian distributions then the Bhattacharyya bound decreases. Do this as follows: Let $P_d(P(\omega_1), \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, P(\omega_2), \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2)$, or simply P_d, be the Bhattacharyya bound if we consider the distributions restricted to d dimensions.

(a) Using general properties of a covariance matrix, prove that $k(1/2)$ of Eq. 77 must increase as we increase from d to $d+1$ dimensions, and hence the error bound must decrease.

(b) Explain why this general result does or does not depend upon *which* dimension is added.

(c) What is a "pathological" case in which the error bound does *not* decrease, that is, for which $P_{d+1} = P_d$?

(d) Is it ever possible that the *true* error—that is, not just the *bound*—could *increase* as we go to higher dimension?

(e) Prove that as $d \to \infty$, $P_d \to 0$ for nonpathological distributions. Describe pathological distributions for which this infinite limit does not hold.

(f) Given that the Bhattacharyya bound decreases for the inclusion of a particular dimension, does this guarantee that the *true* error will decrease? Explain.

36. Derive Eqs. 74 and 75 from Eq. 73 by the following steps:

(a) Substitute the normal distributions into the integral and gather the terms dependent upon $\mathbf{x}$ and those that are not dependent upon $\mathbf{x}$.

(b) Factor the term independent of $\mathbf{x}$ from the integral.

(c) Integrate explicitly the term dependent upon $\mathbf{x}$.

37. Consider a two-category classification problem in two dimensions with

$$p(\mathbf{x}|\omega_1) \sim N(\mathbf{0}, \mathbf{I}),\ p(\mathbf{x}|\omega_2) \sim N\left(\begin{pmatrix}1\\1\end{pmatrix}, \mathbf{I}\right),\ \text{and } P(\omega_1) = P(\omega_2) = 1/2.$$

(a) Calculate the Bayes decision boundary.

(b) Calculate the Bhattacharyya error bound.

(c) Repeat the above for the same prior probabilities, but

$$p(\mathbf{x}|\omega_1) \sim N\left(\mathbf{0}, \begin{pmatrix}2 & .5\\ .5 & 2\end{pmatrix}\right) \text{ and } p(\mathbf{x}|\omega_2) \sim N\left(\begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}5 & 4\\ 4 & 5\end{pmatrix}\right).$$

38. Derive the Bhattacharyya error bound without the need for first examining the Chernoff bound. Do this as follows:

(a) If a and b are nonnegative numbers, show directly that min[a,b] $\leq \sqrt{ab}$.

(b) Use this to show that the error rate for a two-category Bayes classifier must satisfy

$$P(error) \leq \sqrt{P(\omega_1)P(\omega_2)}\ \rho \leq \rho/2,$$

where ρ is the so-called *Bhattacharyya coefficient*

$$\rho = \int \sqrt{p(\mathbf{x}|\omega_1)\ p(\mathbf{x}|\omega_2)}\ d\mathbf{x}.$$

39. Use signal detection theory, as well as the notation and basic Gaussian assumptions described in the text, to address the following.

(a) Prove that $P(x > x^*|x \in \omega_2)$ and $P(x > x^*|x \in \omega_1)$, taken together, uniquely determine the discriminability d'.

(b) Use error functions erf($\cdot$) to express d' in terms of the hit and false alarm rates. Estimate d' if $P(x > x^*|x \in \omega_1) = 0.8$ and $P(x > x^*|x \in \omega_2) = 0.3$. Repeat for $P(x > x^*|x \in \omega_1) = 0.7$ and $P(x > x^*|x \in \omega_1) = 0.4$.

(c) Given that the Gaussian assumption is valid, calculate the Bayes error for both the cases in (b).

(d) Using a trivial one-line computation determine which case has the higher d':

Case A: $P(x > x^*|x \in \omega_1) = 0.8$, $P(x > x^*|x \in \omega_2) = 0.3$ or

Case B: $P(x > x^*|x \in \omega_1) = 0.3$, $P(x > x^*|x \in \omega_2) = 0.7$.

Explain your logic.

40. Suppose in our signal detection framework we had two Gaussians, but with different variances (cf. Fig. 2.20)—that is, $p(x|\omega_1) \sim N(\mu_1, \sigma_1^2)$ and $p(x|\omega_2) \sim N(\mu_2, \sigma_2^2)$ for $\mu_2 > \mu_1$ and $\sigma_2^2 \neq \sigma_1^2$. In that case the resulting ROC curve would no longer be symmetric.

(a) Suppose in this asymmetric case we modified the definition of the discriminability to be $d_a' = |\mu_2 - \mu_1|/\sqrt{\sigma_1\sigma_2}$. Show by nontrivial counterexample or analysis that one cannot determine d_a' uniquely based on a single pair of hit and false alarm rates.

(b) Assume we measure the hit and false alarm rates for two different, but unknown, values of the threshold x^*. Derive a formula for d_a' based on such measurements.

(c) State and explain all pathological values for which your formula does not give a meaningful value for d_a'.

(d) Plot several ROC curves for the case $p(x|\omega_1) \sim N(0, 1)$ and $p(x|\omega_2) \sim N(1, 2)$.

41. Consider two one-dimensional triangle distributions having different means, but the same width:

$$p(x|\omega_i) = T(\mu_i, \delta) = \begin{cases} (\delta - |x - \mu_i|)/\delta^2 & \text{for } |x - \mu_i| < \delta \\ 0 & \text{otherwise,} \end{cases}$$

with $\mu_2 > \mu_1$. We define a new discriminability here as $d_T' = (\mu_2 - \mu_1)/\delta$.

(a) Write an analytic function, parameterized by d'_T, for the operating characteristic curves.

(b) Plot these novel operating characteristic curves for $d'_T = \{.1, .2, \ldots, 1.0\}$. Interpret your answer for the case $d'_T = 1.0$ and 2.0.

(c) Suppose we measure $P(x > x^*|x \in \omega_2) = 0.4$ and $P(x > x^*|x \in \omega_1) = 0.7$. What is d'_T? What is the Bayes error rate?

(d) Infer the decision rule employed in part (c). That is, express x^* in terms of the variables given in the problem.

(e) Suppose we measure $P(x > x^*|x \in \omega_2) = 0.3$ and $(x > x^*|x \in \omega_1) = 0.9$. What is d'_T? What is the Bayes error rate?

(f) Infer the decision rule employed in part (e). That is, express x^* in terms of the variables given in the problem.

42. Equation 72 can be used to obtain an upper bound on the error. One can also derive tighter analytic bounds in the two-category case—both upper and lower bounds—analogous to Eq. 73 for general distributions. If we let $p \equiv p(x|\omega_1)$, then we seek tighter bounds on $\min[p, 1-p]$ (which has discontinuous derivative).

(a) Prove that

$$b_L(p) = \frac{1}{\beta} \ln \left[\frac{1 + e^{-\beta}}{e^{-\beta p} + e^{-\beta(1-p)}} \right]$$

for any $\beta > 0$ is a lower bound on $\min[p, 1-p]$.

(b) Prove that one can choose β in (a) to give an arbitrarily tight lower bound.

(c) Repeat (a) and (b) for the upper bound given by

$$b_U(p) = b_L(p) + [1 - 2b_L(0.5)]b_G(p)$$

where $b_G(p)$ is any upper bound that obeys

$$\begin{aligned} b_G(p) &\geq \min[p, 1-p] \\ b_G(p) &= b_G(1-p) \\ b_G(0) &= b_G(1) = 0 \\ b_G(0.5) &= 0.5. \end{aligned}$$

(d) Confirm that $b_G(p) = 1/2 \sin[\pi p]$ obeys the conditions in (c).

(e) Let $b_G(p) = 1/2 \sin[\pi p]$, and plot your upper and lower bounds as a function of p, for $0 \leq p \leq 1$ and $\beta = 1$, 10, and 50.

Section 2.9

43. Let the components of the vector $\mathbf{x} = (x_1, \ldots, x_d)^t$ be binary-valued (0 or 1), and let $P(\omega_j)$ be the prior probability for the state of nature ω_j and $j = 1, \ldots, c$. Now define

$$p_{ij} = \Pr[x_i = 1|\omega_j] \qquad \begin{array}{l} i = 1, \ldots, d \\ j = 1, \ldots, c, \end{array}$$

with the components of x_i being statistically independent for all $\mathbf{x}$ in ω_j.

(a) Interpret in words the meaning of p_{ij}.

(b) Show that the minimum probability of error is achieved by the following decision rule: Decide ω_k if $g_k(\mathbf{x}) \geq g_j(\mathbf{x})$ for all j and k, where

$$g_j(\mathbf{x}) = \sum_{i=1}^{d} x_i \ln \frac{p_{ij}}{1 - p_{ij}} + \sum_{i=1}^{d} \ln(1 - p_{ij}) + \ln P(\omega_j).$$

44. Let the components of the vector $\mathbf{x} = (x_1, \ldots, x_d)^t$ be ternary valued (1, 0 or -1), with

$$\begin{aligned} p_{ij} &= \Pr[x_i = \ 1\ |\omega_j] \\ q_{ij} &= \Pr[x_i = \ 0\ |\omega_j] \\ r_{ij} &= \Pr[x_i = -1|\omega_j], \end{aligned}$$

and with the components of x_i being statistically independent for all $\mathbf{x}$ in ω_j.

(a) Show that a minimum probability of error decision rule can be derived that involves discriminant functions $g_j(\mathbf{x})$ that are quadratic function of the components x_i.

(b) Suggest a generalization to more categories of your answers to this and Problem 43.

45. Let $\mathbf{x}$ be distributed as in Problem 43 with $c = 2$, d odd, and

$$\begin{aligned} p_{i1} &= p > 1/2 \qquad i = 1, \ldots, d \\ p_{i2} &= 1 - p \qquad\quad i = 1, \ldots, d, \end{aligned}$$

and $P(\omega_1) = P(\omega_2) = 1/2$.

(a) Show that the minimum-error-rate decision rule becomes

$$\text{Decide } \omega_1 \text{ if } \sum_{i=1}^{d} x_i > d/2 \text{ and } \omega_2 \text{ otherwise.}$$

(b) Show that the minimum probability of error is given by

$$P_e(d, p) = \sum_{k=0}^{(d-1)/2} \binom{d}{k} p^k (1 - p)^{d-k}.$$

where $\binom{d}{k} = d!/(k!(d-k)!)$ is the binomial coefficient.

(c) What is the limiting value of $P_e(d, p)$ as $p \to 1/2$? Explain.

(d) Show that $P_e(d, p)$ approaches zero as $d \to \infty$. Explain.

46. Under the natural assumption concerning losses, i.e., that $\lambda_{21} > \lambda_{11}$ and $\lambda_{12} > \lambda_{22}$, show that the general minimum risk discriminant function for the independent binary case described in Section 2.9.1 is given by $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x} + w_0$, where $\mathbf{w}$ is unchanged, and

$$w_0 = \sum_{i=1}^{d} \ln \frac{1 - p_i}{1 - q_i} + \ln \frac{P(\omega_1)}{P(\omega_2)} + \ln \frac{\lambda_{21} - \lambda_{11}}{\lambda_{12} - \lambda_{22}}.$$

47. The Poisson distribution for a discrete variable $x = 0, 1, 2, \ldots$ and real parameter λ is

$$P(x|\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}.$$

(a) Prove that the mean of such a distribution is $\mathcal{E}[x] = \lambda$.

(b) Prove that the variance of such a distribution is $\mathcal{E}[x - \bar{x}] = \lambda$.

(c) The *mode* of a distribution is the value of x that has the maximum probability. Prove that the mode of a Poisson distribution is the greatest integer that does not exceed λ. That is, prove that the mode is $\lfloor\lambda\rfloor$, read "floor of lambda." (If λ is an integer, then both λ and $\lambda - 1$ are modes.)

(d) Consider two equally probable categories having Poisson distributions but with differing parameters; assume for definiteness $\lambda_1 > \lambda_2$. What is the Bayes classification decision?

(e) What is the Bayes error rate?

Section 2.10

48. Suppose we have three categories in two dimensions with the following underlying distributions:

- $p(\mathbf{x}|\omega_1) \sim N(\mathbf{0}, \mathbf{I})$
- $p(\mathbf{x}|\omega_2) \sim N\left(\begin{pmatrix}1\\1\end{pmatrix}, \mathbf{I}\right)$
- $p(\mathbf{x}|\omega_3) \sim \frac{1}{2}N\left(\begin{pmatrix}.5\\.5\end{pmatrix}, \mathbf{I}\right) + \frac{1}{2}N\left(\begin{pmatrix}-.5\\.5\end{pmatrix}, \mathbf{I}\right)$

with $P(\omega_i) = 1/3, i = 1, 2, 3$.

(a) By explicit calculation of posterior probabilities, classify the point $\mathbf{x} = \binom{.3}{.3}$ for minimum probability of error.

(b) Suppose that for a particular test point the first feature is missing. That is, classify $\mathbf{x} = \binom{*}{.3}$.

(c) Suppose that for a particular test point the second feature is missing. That is, classify $\mathbf{x} = \binom{.3}{*}$.

(d) Repeat all of the above for $\mathbf{x} = \binom{.2}{.6}$.

49. Show that Eq. 95 reduces to Bayes rule when the true feature is $\boldsymbol{\mu}_i$ and $p(\mathbf{x}_b|\mathbf{x}_t) \sim N(\mathbf{x}_t, \boldsymbol{\Sigma})$. Interpret this answer in words.

Section 2.11

50. Use the conditional probability matrices in Example 4 to answer the following separate problems.

(a) Suppose it is December 20—the end of autumn and the beginning of winter—and thus let $P(a_1) = P(a_4) = 0.5$. Furthermore, it is known that the fish was caught in the north Atlantic, that is, $P(b_1) = 1$. Suppose the lightness has not been measured but it is known that the fish is thin, that is, $P(d_2) = 1$. Classify the fish as salmon or sea bass. What is the expected error rate?

(b) Suppose all we know is that a fish is thin and medium lightness. What season is it now, most likely? What is your probability of being correct?

(c) Suppose we know a fish is thin and medium lightness and that it was caught in the north Atlantic. What season is it, most likely? What is the probability of being correct?

51. Consider a Bayesian belief net with several nodes having unspecified values. Suppose that one such node is selected at random, with the probabilities of its nodes computed by the formulas described in the text. Next, another such node is chosen at random (possibly even a node already visited), and the probabilities are similarly updated. Prove that this procedure will converge to the desired probabilities throughout the full network.

Section 2.12

52. Suppose we have three categories with $P(\omega_1) = 1/2$, $P(\omega_2) = P(\omega_3) = 1/4$ and the following distributions

- $p(x|\omega_1) \sim N(0, 1)$
- $p(x|\omega_2) \sim N(.5, 1)$
- $p(x|\omega_3) \sim N(1, 1)$,

and that we sample the following four points: $x = 0.6, 0.1, 0.9, 1.1$.

(a) Calculate explicitly the probability that the sequence actually came from $\omega_1, \omega_3, \omega_3, \omega_2$. Be careful to consider normalization.

(b) Repeat for the sequence $\omega_1, \omega_2, \omega_2, \omega_3$.

(c) Find the sequence having the maximum probability.

COMPUTER EXERCISES

Several of the computer exercises will rely on the following data.

	ω_1			ω_2			ω_3		
sample	x_1	x_2	x_3	x_1	x_2	x_3	x_1	x_2	x_3
1	−5.01	−8.12	−3.68	−0.91	−0.18	−0.05	5.35	2.26	8.13
2	−5.43	−3.48	−3.54	1.30	−2.06	−3.53	5.12	3.22	−2.66
3	1.08	−5.52	1.66	−7.75	−4.54	−0.95	−1.34	−5.31	−9.87
4	0.86	−3.78	−4.11	−5.47	0.50	3.92	4.48	3.42	5.19
5	−2.67	0.63	7.39	6.14	5.72	−4.85	7.11	2.39	9.21
6	4.94	3.29	2.08	3.60	1.26	4.36	7.17	4.33	−0.98
7	−2.51	2.09	−2.59	5.37	−4.63	−3.65	5.75	3.97	6.65
8	−2.25	−2.13	−6.94	7.18	1.46	−6.66	0.77	0.27	2.41
9	5.56	2.86	−2.26	−7.39	1.17	6.30	0.90	−0.43	−8.71
10	1.03	−3.33	4.33	−7.50	−6.32	−0.31	3.52	−0.36	6.43

Section 2.5

1. You may need the following procedures for several exercises below.

(a) Write a procedure to generate random samples according to a normal distribution $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ in d dimensions.

(b) Write a procedure to calculate the discriminant function (of the form given in Eq. 49) for a given normal distribution and prior probability $P(\omega_i)$.

(c) Write a procedure to calculate the Euclidean distance between two arbitrary points.

(d) Write a procedure to calculate the Mahalanobis distance between the mean $\boldsymbol{\mu}$ and an arbitrary point $\mathbf{x}$, given the covariance matrix $\boldsymbol{\Sigma}$.

2. Refer to Computer exercise 1 (b) and consider the problem of classifying 10 samples from the table above. Assume that the underlying distributions are normal.

(a) Assume that the prior probabilities for the first two categories are equal ($P(\omega_1) = P(\omega_2) = 1/2$ and $P(\omega_3) = 0$) and design a dichotomizer for those two categories using only the x_1 feature value.

(b) Determine the empirical training error on your samples, that is, the percentage of points misclassified.

(c) Use the Bhattacharyya bound to bound the error you will get on novel patterns drawn from the distributions.

(d) Repeat all of the above, but now use *two* feature values, x_1 and x_2.

(e) Repeat, but use all *three* feature values.

(f) Discuss your results. In particular, is it ever possible for a finite set of data that the empirical error might be *larger* for more data dimensions?

3. Repeat Computer exercise 2 but for categories ω_1 and ω_3.

4. Consider the three categories in Computer exercise 2, and assume $P(\omega_i) = 1/3$.

(a) What is the Mahalanobis distance between each of the following test points and each of the category means in Computer exercise 2: $(1, 2, 1)^t$, $(5, 3, 2)^t$, $(0, 0, 0)^t$, $(1, 0, 0)^t$.

(b) Classify those points.

(c) Assume instead that $P(\omega_1) = 0.8$, and $P(\omega_2) = P(\omega_3) = 0.1$ and classify the test points again.

5. Illustrate the fact that the average of a large number of independent random variables will approximate a Gaussian by the following:

(a) Write a program to generate n random integers from a uniform distribution $U(x_l, x_u)$. (Some computer systems include this as a single, compiled function call.)

(b) Now write a routine to choose x_l and x_u randomly, in the range $-100 \leq x_l < x_u \leq +100$, and n (the number of samples) randomly in the range $0 < n \leq 1000$.

(c) Generate and plot a histogram of the accumulation of 10^4 points sampled as just described.

(d) Calculate the mean and standard deviation of your histogram, and plot it

(e) Repeat the above for 10^5 and for 10^6. Discuss your results.

Section 2.8

6. Explore how the empirical error does or does not approach the Bhattacharyya bound as follows:

(a) Write a procedure to generate sample points in d dimensions with a normal distribution having mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$.

(b) Consider the normal distributions

$$p(\mathbf{x}|\omega_1) \sim N\left(\begin{pmatrix}1\\0\end{pmatrix}, \mathbf{I}\right) \text{ and } p(\mathbf{x}|\omega_2) \sim N\left(\begin{pmatrix}-1\\0\end{pmatrix}, \mathbf{I}\right)$$

with $P(\omega_1) = P(\omega_2) = 1/2$. By inspection, state the Bayes decision boundary.

(c) Generate $n = 100$ points (50 for ω_1 and 50 for ω_2) and calculate the empirical error.

(d) Repeat for increasing values of n, $100 \leq n \leq 1000$, in steps of 100 and plot your empirical error.

(e) Discuss your results. In particular, is it ever possible that the empirical error is greater than the Bhattacharyya or Chernoff bound?

7. Consider two one-dimensional normal distributions $p(x|\omega_1) \sim N(-.5, 1)$ and $p(x|\omega_2) \sim N(+.5, 1)$ and $P(\omega_1) = P(\omega_2) = 0.5$.

(a) Calculate the Bhattacharyya bound for the error of a Bayesian classifier.

(b) Express the true error rate in terms of an error function, erf(·).

(c) Evaluate this true error to four significant figures by numerical integration (or other routine).

(d) Generate 10 points each for the two categories and determine the empirical error using your Bayesian classifier. (You should recalculate the decision boundary for each of your data sets.)

(e) Plot the empirical error as a function of the number of points from either distribution by repeating the previous part for 50, 100, 200, 500 and 1000 sample points from each distribution. Compare your asymptotic empirical error to the true error and the Bhattacharyya error bound.

8. Repeat Computer exercise 7 with the following conditions:

(a) $p(x|\omega_1) \sim N(-.5, 2)$ and $p(x|\omega_2) \sim N(.5, 2)$, $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$.

(b) $p(x|\omega_1) \sim N(-.5, 2)$ and $p(x|\omega_2) \sim N(.5, 2)$ and $P(\omega_1) = P(\omega_2) = 1/2$.

(c) $p(x|\omega_1) \sim N(-.5, 3)$ and $p(x|\omega_2) \sim N(.5, 1)$ and $P(\omega_1) = P(\omega_2) = 1/2$.

Section 2.11

9. Write a program to evaluate the Bayesian belief net for fish in Example 3, including the information in $P(x_i|a_j)$, $P(x_i|b_j)$, $P(c_i|x_j)$, and $P(d_i|x_j)$. Test your program on the calculation given in the Example. Apply your program to the following cases, and state any assumptions you need to make.

(a) A dark, thin fish is caught in the north Atlantic in summer. What is the probability it is a salmon?

(b) A thin, medium fish is caught in the north Atlantic. What is the probability it is winter? spring? summer? autumn?

(c) A light, wide fish is caught in the autumn. What is the probability it came from the north Atlantic?

BIBLIOGRAPHY

[1] Subutai Ahmad and Volker Tresp. Some solutions to the missing feature problem in vision. In Stephen J. Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 393–400, Morgan Kaufmann San Mateo, CA, 1993.

[2] Hadar Avi-Itzhak and Thanh Diep. Arbitrarily tight upper and lower bounds on the Bayesian probability of error. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-18(1):89–91, 1996.

[3] Thomas Bayes. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society (London)*, 53:370–418, 1763.

[4] James O. Berger. Minimax estimation of a multivariate normal mean under arbitrary quadratic loss. *Journal of Multivariate Analysis*, 6(2):256–264, 1976.

[5] James O. Berger. Selecting a minimax estimator of a multivariate normal mean. *Annals of Statistics*, 10(1):81–92, 1982.

[6] James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer-Verlag, New York, second edition, 1985.

[7] José M. Bernardo and Adrian F. M. Smith. *Bayesian Theory*. Wiley, New York, 1996.

[8] Anil Bhattacharyya. On a measure of divergence between two statistical populations defined by their probability distributions. *Bulletin of the Calcutta Mathematical Society*, 35:99–110, 1943.

[9] Wray L. Buntine. Operations for learning with graphical models. *Journal of Artificial Intelligence Research*, 2:159–225, 1994.

[10] Wray L. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8(2):195–210, 1996.

[11] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[12] Chao K. Chow. An optimum character recognition system using decision functions. *IRE Transactions*, pages 247–254, 1957.

[13] Chao K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, IT-16:41–46, 1970.

[14] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 1991.

[15] Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, New York, 1970.

[16] Bradley Efron and Carl Morris. Families of minimax estimators of the mean of a multivariate normal distribution. *Annals of Statistics*, 4:11–21, 1976.

[17] Thomas S. Ferguson. *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York, 1967.

[18] Simon French. *Decision Theory: An Introduction to the Mathematics of Rationality*. Halsted Press, New York, 1986.

[19] Keinosuke Fukunaga. *Introduction to Statistical Pattern Recognition*. Academic Press, New York, second edition, 1990.

[20] Keinosuke Fukunaga and Thomas F. Krile. Calculation of Bayes recognition error for two multivariate Gaussian distributions. *IEEE Transactions on Computers*, C-18:220–229, 1969.

[21] Izrail M. Gelfand and Sergei Vasilevich Fomin. *Calculus of Variations*. Prentice-Hall, Englewood Cliffs, NJ, translated from the Russian by Richard A. Silverman, 1963.

[22] David M. Green and John A. Swets. *Signal Detection Theory and Psychophysics*. Wiley, New York, 1974.

[23] David J. Hand. *Construction and Assessment of Classification Rules*. Wiley, New York, 1997.

[24] Peter E. Hart and Jamey Graham. Query-free information retrieval. *IEEE Expert: Intelligent Systems and Their Application*, 12(5):32–37, 1997.

[25] David Heckerman. *Probabilistic Similarity Networks*. ACM Doctoral Dissertation Award Series. MIT Press, Cambridge, MA, 1991.

[26] Anil K. Jain. On an estimate of the Bhattacharyya distance. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-16(11):763–766, 1976.

[27] Michael I. Jordan, editor. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

[28] Bernard Kolman. *Elementary Linear Algebra*. Macmillan, New York, fifth edition, 1991.

[29] Pierre Simon Laplace. *Théorie Analytique des Probabiltiés*. Courcier, Paris, France, 1812.

[30] Peter M Lee. *Bayesian Statistics: An Introduction*. Edward Arnold, London, 1989.

[31] Dennis V. Lindley. *Making Decisions*. Wiley, New York, 1991.

[32] Jerzy Neyman and Egon S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society, London*, 231:289–337, 1928.

[33] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, San Mateo, CA, 1988.

[34] Sheldon M. Ross. *Introduction to Probability and Statistics for Engineers*. Wiley, New York, 1987.

[35] Donald B. Rubin and Roderick J. A. Little. *Statistical Analysis with Missing Data*. Wiley, New York, 1987.

[36] Claude E. Shannon. A mathematical theory of communication. *Bell Systems Technical Journal*, 27:379–423, 623–656, 1948.

[37] George B. Thomas, Jr. and Ross L. Finney. *Calculus and Analytic Geometry*. Addison-Wesley, New York, ninth edition, 1996.

[38] Julius T. Tou and Rafael C. Gonzalez. *Pattern Recognition Principles*. Addison-Wesley, New York, 1974.

[39] William R. Uttal. *The Psychobiology of Sensory Coding*. HarperCollins, New York, 1973.

[40] Abraham Wald. Contributions to the theory of statistical estimation and testing of hypotheses. *Annals of Mathematical Statistics*, 10:299–326, 1939.

[41] Abraham Wald. *Statistical Decision Functions*. Wiley, New York, 1950.

[42] Charles T. Wolverton and Terry J. Wagner. Asymptotically optimal discriminant functions for pattern classifiers. *IEEE Transactions on Information Theory*, IT-15(2):258–265, 1969.

[43] Sewal Wright. Correlation and causation. *Journal of Agricultural Research*, 20(7):557–585, 1921.

[44] C. Ray Wylie and Louis C. Barrett. *Advanced Engineering Mathematics*. McGraw-Hill, New York, sixth edition, 1995.