2.1 The Bayes Problem

In this section, we define the mathematical model and introduce the notation we will use for the entire book. Let (X, Y) be a pair of random variables taking their respective values from $\mathcal{R}^d$ and $\{0, 1\}$. The random pair (X, Y) may be described in a variety of ways: for example, it is defined by the pair (μ, η), where μ is the probability measure for X and η is the regression of Y on X. More precisely, for a Borel-measurable set $A \subseteq \mathcal{R}^d$,

$$\mu(A) = \mathbf{P}\{X \in A\},$$

and for any $x \in \mathcal{R}^d$,

$$\eta(x) = \mathbf{P}\{Y = 1 | X = x\} = \mathbf{E}\{Y | X = x\}.$$

Thus, $\eta(x)$ is the conditional probability that Y is 1 given $X = x$. To see that this suffices to describe the distribution of (X, Y), observe that for any $C \subseteq \mathcal{R}^d \times \{0, 1\}$, we have

$$C = \big(C \cap (\mathcal{R}^d \times \{0\})\big) \cup \big(C \cap (\mathcal{R}^d \times \{1\})\big) \stackrel{\text{def}}{=} C_0 \times \{0\} \cup C_1 \times \{1\},$$

and

$$\begin{aligned} \mathbf{P}\{(X, Y) \in C\} &= \mathbf{P}\{X \in C_0, Y = 0\} + \mathbf{P}\{X \in C_1, Y = 1\} \\ &= \int_{C_0} (1 - \eta(x))\mu(dx) + \int_{C_1} \eta(x)\mu(dx). \end{aligned}$$

As this is valid for any Borel-measurable set C, the distribution of (X, Y) is determined by (μ, η). The function η is sometimes called the *a posteriori probability*.

Any function $g : \mathcal{R}^d \to \{0, 1\}$ defines a *classifier* or a *decision function*. The error probability of g is $L(g) = \mathbf{P}\{g(X) \neq Y\}$. Of particular interest is the Bayes decision function

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \\ 0 & \text{otherwise.} \end{cases}$$

This decision function minimizes the error probability.

Theorem 2.1. *For any decision function $g : \mathcal{R}^d \to \{0, 1\}$,*

$$\mathbf{P}\{g^*(X) \neq Y\} \leq \mathbf{P}\{g(X) \neq Y\},$$

that is, g^* is the optimal decision.

PROOF. Given $X = x$, the conditional error probability of any decision g may be expressed as

$$\begin{aligned} &\mathbf{P}\{g(X) \neq Y | X = x\} \\ &\quad = 1 - \mathbf{P}\{Y = g(X) | X = x\} \\ &\quad = 1 - (\mathbf{P}\{Y = 1, g(X) = 1 | X = x\} + \mathbf{P}\{Y = 0, g(X) = 0 | X = x\}) \\ &\quad = 1 - \left(I_{\{g(x)=1\}} \mathbf{P}\{Y = 1 | X = x\} + I_{\{g(x)=0\}} \mathbf{P}\{Y = 0 | X = x\}\right) \\ &\quad = 1 - \left(I_{\{g(x)=1\}} \eta(x) + I_{\{g(x)=0\}} (1 - \eta(x))\right), \end{aligned}$$

where I_A denotes the indicator of the set A. Thus, for every $x \in \mathcal{R}^d$,

$$\begin{aligned} &\mathbf{P}\{g(X) \neq Y | X = x\} - \mathbf{P}\{g^*(X) \neq Y | X = x\} \\ &\quad = \eta(x) \left(I_{\{g^*(x)=1\}} - I_{\{g(x)=1\}}\right) + (1 - \eta(x)) \left(I_{\{g^*(x)=0\}} - I_{\{g(x)=0\}}\right) \\ &\quad = (2\eta(x) - 1) \left(I_{\{g^*(x)=1\}} - I_{\{g(x)=1\}}\right) \\ &\quad \geq 0 \end{aligned}$$

by the definition of g^*. The statement now follows by integrating both sides with respect to $\mu(dx)$. □

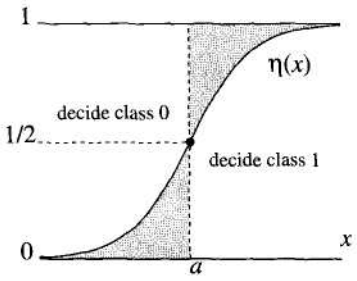

FIGURE 2.1. *The Bayes decision in the example on the left is 1 if $x > a$, and 0 otherwise.*

g^* is called the Bayes decision and $L^* = \mathbf{P}\{g^*(X) \neq Y\}$ is referred to as es probability of error, Bayes error, or Bayes risk. The proof given above that

$$L(g) = 1 - \mathbf{E}\left\{I_{\{g(X)=1\}}\eta(X) + I_{\{g(X)=0\}}(1-\eta(X))\right\},$$

particular,

$$L^* = 1 - \mathbf{E}\left\{I_{\{\eta(X)>1/2\}}\eta(X) + I_{\{\eta(X)\leq 1/2\}}(1-\eta(X))\right\}. \quad \square$$

e observe that the a posteriori probability

$$\eta(x) = \mathbf{P}\{Y = 1|X = x\} = \mathbf{E}\{Y|X = x\}.$$

imizes the squared error when Y is to be predicted by $f(X)$ for some function $\mathcal{R}^d \to \mathcal{R}$:

$$\mathbf{E}\left\{(\eta(X) - Y)^2\right\} \leq \mathbf{E}\left\{(f(X) - Y)^2\right\}.$$

ee why the above inequality is true, observe that for each $x \in \mathcal{R}^d$,

$$\begin{aligned}
&\mathbf{E}\left\{(f(X) - Y)^2|X = x\right\} \\
&\quad = \mathbf{E}\left\{(f(x) - \eta(x) + \eta(x) - Y)^2|X = x\right\} \\
&\quad = (f(x) - \eta(x))^2 + 2(f(x) - \eta(x))\mathbf{E}\{\eta(x) - Y|X = x\} \\
&\qquad + \mathbf{E}\left\{(\eta(X) - Y)^2|X = x\right\} \\
&\quad = (f(x) - \eta(x))^2 + \mathbf{E}\left\{(\eta(X) - Y)^2|X = x\right\}.
\end{aligned}$$

he conditional median, i.e., the function minimizing the absolute error $\mathbf{E}\{|f(X) - |\}$ is even more closely related to the Bayes rule (see Problem 2.12).

2.2 A Simple Example

Let us consider the prediction of a student's performance in a course (pass/fail) when given a number of important factors. First, let $Y = 1$ denote a pass and let $Y = 0$ stand for failure. The sole observation X is the number of hours of study per week. This, in itself, is not a foolproof predictor of a student's performance, because for that we would need more information about the student's quickness of mind, health, and social habits. The regression function $\eta(x) = \mathbf{P}\{Y = 1|X = x\}$ is probably monotonically increasing in x. If it were known to be $\eta(x) = x/(c + x)$, $c > 0$, say, our problem would be solved because the Bayes decision is

$$g^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 \text{ (i.e., } x > c) \\ 0 & \text{otherwise.} \end{cases}$$

The corresponding Bayes error is

$$L^* = L(g^*) = \mathbf{E}\{\min(\eta(X), 1 - \eta(X))\} = \mathbf{E}\left\{\frac{\min(c, X)}{c + X}\right\}.$$

While we could deduce the Bayes decision from η alone, the same cannot be said for the Bayes error L^*—it requires knowledge of the distribution of X. If $X = c$ with probability one (as in an army school, where all students are forced to study c hours per week), then $L^* = 1/2$. If we have a population that is nicely spread out, say, X is uniform on $[0, 4c]$, then the situation improves:

$$L^* = \frac{1}{4c}\int_0^{4c} \frac{\min(c, x)}{c + x} dx = \frac{1}{4}\log\frac{5e}{4} \approx 0.305785.$$

Far away from $x = c$, discrimination is really simple. In general, discrimination is much easier than estimation because of this phenomenon.

2.3 Another Simple Example

Let us work out a second simple example in which $Y = 0$ or $Y = 1$ according to whether a student fails or passes a course. X represents one or more observations regarding the student. The components of X in our example will be denoted by T, B, and E respectively, where T is the average number of hours the students watches TV, B is the average number of beers downed each day, and E is an intangible quantity measuring extra negative factors such as laziness and learning difficulties. In our cooked-up example, we have

$$Y = \begin{cases} 1 & \text{if } T + B + E < 7 \\ 0 & \text{otherwise.} \end{cases}$$

Thus, if T, B, and E are known, Y is known as well. The Bayes classifier decides 1 if $T + B + E < 7$ and 0 otherwise. The corresponding Bayes probability of error is zero. Unfortunately, E is intangible, and not available to the observer. We only have access to T and B. Given T and B, when should we guess that $Y = 1$? To answer this question, one must know the joint distribution of (T, B, E), or, equivalently, the joint distribution of (T, B, Y). So, let us assume that T, B, and E are i.i.d. exponential random variables (thus, they have density e^{-u} on $[0, \infty)$). The Bayes rule compares $\mathbf{P}\{Y = 1|T, B\}$ with $\mathbf{P}\{Y = 0|T, B\}$ and makes a decision consistent with the maximum of these two values. A simple calculation shows that

$$\begin{aligned} \mathbf{P}\{Y = 1|T, B\} &= \mathbf{P}\{T + B + E < 7|T, B\} \\ &= \mathbf{P}\{E < 7 - T - B|T, B\} \\ &= \max\left(0, 1 - e^{-(7-T-B)}\right). \end{aligned}$$

The crossover between two decisions occurs when this value equals $1/2$. Thus, the Bayes classifier is as follows:

$$g^*(T, B) = \begin{cases} 1 & \text{if } T + B < 7 - \log 2 = 6.306852819\ldots \\ 0 & \text{otherwise.} \end{cases}$$

ourse, this classifier is not perfect. The probability of error is

$$\begin{aligned}
&\mathbf{P}\{g^*(T,B) \neq Y\} \\
&= \mathbf{P}\{T+B < 7-\log 2, T+B+E \geq 7\} \\
&\quad + \mathbf{P}\{T+B \geq 7-\log 2, T+B+E < 7\} \\
&= \mathbf{E}\left\{e^{-(7-T-B)} I_{\{T+B<7-\log 2\}}\right\} \\
&\quad + \mathbf{P}\left\{\left(1-e^{-(7-T-B)}\right) I_{\{7>T+B\geq 7-\log 2\}}\right\} \\
&= \int_0^{7-\log 2} x e^{-x} e^{-(7-x)}\, dx + \int_{7-\log 2}^{7} x e^{-x}\left(1-e^{-(7-x)}\right) dx \\
&\quad \text{(since the density of } T+B \text{ is } ue^{-u} \text{ on } [0,\infty)) \\
&= e^{-7}\left(\frac{(7-\log 2)^2}{2} + 2(8-\log 2) - 8 - \frac{7^2}{2} + \frac{(7-\log 2)^2}{2}\right) \\
&\quad \text{(as } \int_x^\infty ue^{-u}du = (1+x)e^{-x}\text{)} \\
&= 0.0199611\ldots .
\end{aligned}$$

we have only access to T, then the Bayes classifier is allowed to use T only. irst, we find

$$\begin{aligned}
\mathbf{P}\{Y=1|T\} &= \mathbf{P}\{E+B<7-T|T\} \\
&= \max\left(0, 1-(1+7-T)e^{-(7-T)}\right) .
\end{aligned}$$

The crossover at $1/2$ occurs at $T = c \stackrel{\text{def}}{=} 5.321653009\ldots$, so that the Bayes classifier is given by

$$g^*(T) = \begin{cases} 1 & \text{if } T < c \\ 0 & \text{otherwise.} \end{cases}$$

The probability of error is

$$\begin{aligned}
&\mathbf{P}\{g^*(T) \neq Y\} \\
&= \mathbf{P}\{T<c, T+B+E \geq 7\} + \mathbf{P}\{T \geq c, T+B+E < 7\} \\
&= \mathbf{E}\left\{(1+7-T)e^{-(7-T)} I_{\{T<c\}}\right\} \\
&\quad + \mathbf{P}\left\{\left(1-(1+7-T)e^{-(7-T)}\right) I_{\{7>T\geq c\}}\right\} \\
&= \int_0^c e^{-x}(1+7-x)e^{-(7-x)}\, dx + \int_c^7 e^{-x}\left(1-(1+7-x)e^{-(7-x)}\right) dx \\
&= e^{-7}\left(\frac{8^2}{2} - \frac{(8-c)^2}{2} + e^{-(c-7)} - 1 - \frac{(8-c)^2}{2} + \frac{1}{2}\right) \\
&= 0.02235309002\ldots .
\end{aligned}$$

The Bayes error has increased slightly, but not by much. Finally, if we do not have access to any of the three variables, T, B, and E, the best we can do is see which

class is most likely. To this end, we compute

$$\mathbf{P}\{Y = 0\} = \mathbf{P}\{T + B + E \geq 7\} = (1 + 7 + 7^2/2)e^{-7} = .02963616388\ldots .$$

If we set $g \equiv 1$ all the time, we make an error with probability $0.02963616388\ldots$.

In practice, Bayes classifiers are unknown simply because the distribution of (X, Y) is unknown. Consider a classifier based upon (T, B). Rosenblatt's perceptron (see Chapter 4) looks for the best linear classifier based upon the data. That is, the decision is of the form

$$g(T, B) = \begin{cases} 1 & \text{if } aT + bB < c \\ 0 & \text{otherwise} \end{cases}$$

for some data-based choices for a, b and c. If we have lots of data at our disposal, then it is possible to pick out a linear classifier that is nearly optimal. As we have seen above, the Bayes classifier happens to be linear. That is a sheer coincidence, of course. If the Bayes classifier had not been linear—for example, if we had $Y = I_{\{T+B^2+E<7\}}$—then even the best perceptron would be suboptimal, regardless of how many data pairs one would have. If we use the 3-nearest neighbor rule (Chapter 5), the asymptotic probability of error is not more than 1.3155 times the Bayes error, which in our example is about 0.02625882705. The example above also shows the need to look at individual components, and to evaluate how many and which components would be most useful for discrimination. This subject is covered in the chapter on feature extraction (Chapter 32).

2.4 Other Formulas for the Bayes Risk

The following forms of the Bayes error are often convenient:

$$\begin{aligned} L^* &= \inf_{g:\mathcal{R}^d \to \{0,1\}} \mathbf{P}\{g(X) \neq Y\} \\ &= \mathbf{E}\{\min\{\eta(X), 1 - \eta(X)\}\} \\ &= \frac{1}{2} - \frac{1}{2}\mathbf{E}\{|2\eta(X) - 1|\}. \end{aligned}$$

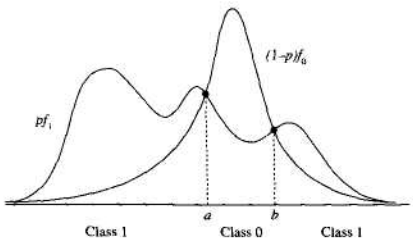

FIGURE 2.2. *The Bayes decision when class conditional densities exist. In the figure on the left, the decision is* 0 *on* $[a, b]$ *and* 1 *elsewhere.*

In special cases, we may obtain other helpful forms. For example, if X has a density f, then

$$\begin{aligned} L^* &= \int \min(\eta(x), 1 - \eta(x)) f(x) dx \\ &= \int \min((1-p) f_0(x), p f_1(x))\, dx, \end{aligned}$$

where $p = \mathbf{P}\{Y = 1\}$, and $f_i(x)$ is the density of X given that $Y = i$. p and $1 - p$ are called the *class probabilities*, and f_0, f_1 are the class-conditional densities. If f_0 and f_1 are nonoverlapping, that is, $\int f_0 f_1 = 0$, then obviously $L^* = 0$. Assume moreover that $p = 1/2$. Then

$$\begin{aligned} L^* &= \frac{1}{2} \int \min(f_0(x), f_1(x))\, dx \\ &= \frac{1}{2} \int f_1(x) - (f_1(x) - f_0(x))_+ \, dx \\ &= \frac{1}{2} - \frac{1}{4} \int |f_1(x) - f_0(x)| dx. \end{aligned}$$

Here g_+ denotes the positive part of a function g. Thus, the Bayes error is directly related to the L_1 distance between the class densities.

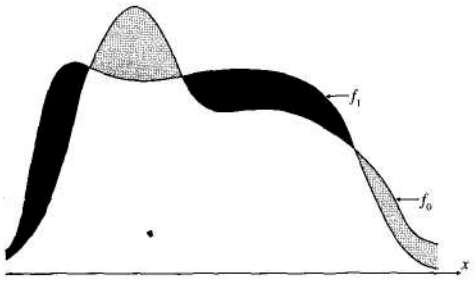

FIGURE 2.3. *The shaded area is the L_1 distance between the class-conditional densities.*

Problems and Exercises

PROBLEM 2.1. Let $T : \mathcal{X} \to \mathcal{X}'$ be an arbitrary measurable function. If L_X^* and $L_{T(X)}^*$ denote the Bayes error probabilities for (X, Y) and $(T(X), Y)$, respectively, then prove that

$$L_{T(X)}^* \geq L_X^*.$$

(This shows that transformations of X destroy information, because the Bayes risk increases.)

PROBLEM 2.2. Let X' be independent of (X, Y). Prove that

$$L_{(X,X')}^* = L_X^*.$$

PROBLEM 2.3. Show that $L^* \leq \min(p, 1-p)$, where $p, 1-p$ are the class probabilities. Show that equality holds if X and Y are independent. Exhibit a distribution where X is not independent of Y, but $L^* = \min(p, 1-p)$.

PROBLEM 2.4. NEYMAN-PEARSON LEMMA. Consider again the decision problem, but with a decision g, we now assign two error probabilities,

$$L^{(0)}(g) = \mathbf{P}\{g(X) = 1 | Y = 0\} \quad \text{and} \quad L^{(1)}(g) = \mathbf{P}\{g(X) = 0 | Y = 1\}.$$

Assume that the class-conditional densities f_0, f_1 exist. For $c > 0$, define the decision

$$g_c(x) = \begin{cases} 1 & \text{if } cf_1(x) > f_0(x) \\ 0 & \text{otherwise.} \end{cases}$$

Prove that for any decision g, if $L^{(0)}(g) \leq L^{(0)}(g_c)$, then $L^{(1)}(g) \geq L^{(1)}(g_c)$. In other words, if $L^{(0)}$ is required to be kept under a certain level, then the decision minimizing $L^{(1)}$ has the form of g_c for some c. Note that g^* is like that.

PROBLEM 2.5. DECISIONS WITH REJECTION. Sometimes in decision problems, one is allowed to say "I don't know," if this does not happen frequently. These decisions are called decisions with a reject option (see, e.g., Forney (1968), Chow (1970)). Formally, a decision $g(x)$ can have three values: 0, 1, and "reject." There are two performance measures: the probability of rejection $\mathbf{P}\{g(X) = \text{"reject"}\}$, and the error probability $\mathbf{P}\{g(X) \neq Y | g(X) \neq \text{"reject"}\}$. For a $0 < c < 1/2$, define the decision

$$g_c(x) = \begin{cases} 1 & \text{if } \eta(x) > 1/2 + c \\ 0 & \text{if } \eta(x) \leq 1/2 - c \\ \text{"reject"} & \text{otherwise.} \end{cases}$$

Show that for any decision g, if

$$\mathbf{P}\{g(X) = \text{"reject"}\} \leq \mathbf{P}\{g_c(X) = \text{"reject"}\},$$

then

$$\mathbf{P}\{g(X) \neq Y | g(X) \neq \text{"reject"}\} \geq \mathbf{P}\{g_c(X) \neq Y | g_c(X) \neq \text{"reject"}\}.$$

Thus, to keep the probability of rejection under a certain level, decisions of the form of g_c are optimal (Györfi, Györfi, and Vajda (1978)).

PROBLEM 2.6. Consider the prediction of a student's failure based upon variables T and B, where $Y = I_{\{T+B+E<7\}}$ and E is an inaccessible variable (see Section 2.3).

(1) Let T, B, and E be independent. Merely by changing the distribution of E, show that the Bayes error for classification based upon (T, B) can be made as close as desired to $1/2$.

(2) Let T and B be independent and exponentially distributed. Find a joint distribution of (T, B, E) such that the Bayes classifier is not a linear classifier.

(3) Let T and B be independent and exponentially distributed. Find a joint distribution of (T, B, E) such that the Bayes classifier is given by

$$g^*(T, B) = \begin{cases} 1 & \text{if } T^2 + B^2 < 10, \\ 0 & \text{otherwise}. \end{cases}$$

(4) Find the Bayes classifier and Bayes error for classification based on (T, B) (with Y as above) if (T, B, E) is uniformly distributed on $[0, 4]^3$.

PROBLEM 2.7. Assume that T, B, and E are independent uniform $[0, 4]$ random variables with interpretations as in Section 2.3. Let $Y = 1$ (0) denote whether a student passes (fails) a course. Assume that $Y = 1$ if and only if $TBE \leq 8$.

(1) Find the Bayes decision if no variable is available, if only T is available, and if only T and B are available.

(2) Determine in all three cases the Bayes error.

(3) Determine the best linear classifier based upon T and B only.

PROBLEM 2.8. Let $\eta', \eta'' : \mathcal{R}^d \to [0, 1]$ be arbitrary measurable functions, and define the corresponding decisions by $g'(x) = I_{\{\eta'(x)>1/2\}}$ and $g''(x) = I_{\{\eta''(x)>1/2\}}$. Prove that

$$|L(g') - L(g'')| \leq \mathbf{P}\{g'(X) \neq g''(X)\}$$

and

$$|L(g') - L(g'')| \leq \mathbf{E}\left\{|2\eta(X) - 1| I_{\{g'(X)\neq g''(X)\}}\right\}.$$

PROBLEM 2.9. Prove Theorem 2.3.

PROBLEM 2.10. Assume that the class-conditional densities f_0 and f_1 exist and are approximated by the densities $\tilde{f}_0$ and $\tilde{f}_1$, respectively. Assume furthermore that the class probabilities $p = \mathbf{P}\{Y = 1\}$ and $1 - p = \mathbf{P}\{Y = 0\}$ are approximated by $\tilde{p}_1$ and $\tilde{p}_0$. Prove that for the error probability of the plug-in decision function

$$g(x) = \begin{cases} 0 & \text{if } \tilde{p}_1 \tilde{f}_1(x) \leq \tilde{p}_0 \tilde{f}_0(x) \\ 1 & \text{otherwise,} \end{cases}$$