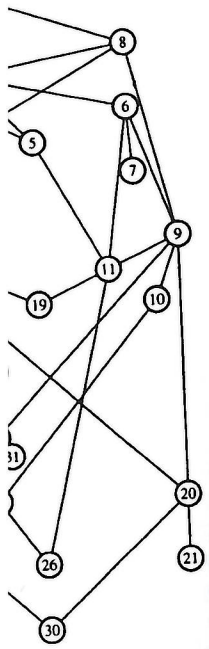

2

The Bayes Error

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

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

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 $\mu(dx)$. □

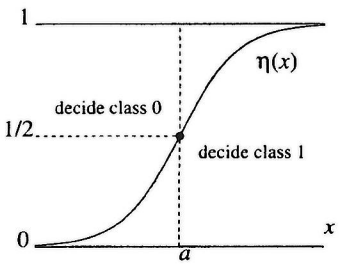

FIGURE 2.1. *The Bayes decision in the example on the left is 1 if $x > a$, and 0 otherwise.*

REMARK. g^* is called the Bay
the Bayes probability of erro
reveals that

$$L(g) = 1 - \mathbf{E}$$

and in particular,

$$L^* = 1 - \mathbf{E}\{I_{\{\eta(}$$

We observe that the a post

$$\eta(x) =$$

minimizes the squared error
$f : \mathcal{R}^d \to \mathcal{R}$:

$$\mathbf{E}\{(\eta(X$$

To see why the above inequa

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

The conditional median, i.e.,
$Y|\}$ is even more closely rela

2.2 A Simple Exan

Let us consider the predictio
when given a number of imp
$Y = 0$ stand for failure. The
per week. This, in itself, is n
because for that we would ne
mind, health, and social habit
probably monotonically incr
$c > 0$, say, our problem woul

$$g^*(x) = \{$$

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

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

and in particular,

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

We observe that the a posteriori probability

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

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

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

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

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

The conditional median, i.e., the function minimizing the absolute error $\mathbf{E}\{|f(X) - Y|\}$ 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

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

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:

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