1 Introduction

Recall that a Bayes classifier is a classifier whose risk \( R(h) \) is minimal among all possible classifiers, and the minimum risk \( R^* \) is called the Bayes risk. Assume \( \mathcal{Y} = \{0, 1\} \) and define

\[
\eta(x) := \Pr(Y = 1 | X = x),
\]

the posterior probability of the class being one, and sometimes called the regression function because \( \eta(x) = \mathbb{E}[Y | X = x] \) when \( \mathcal{Y} = \{0, 1\} \). Also define

\[
h^*(x) := \begin{cases} 1 & \text{if } \eta(x) \geq \frac{1}{2} \\ 0 & \text{otherwise.} \end{cases}
\]

2 Properties of the Bayes Risk

**Theorem 1.** (a) \( R(h^*) = R^* \), i.e., \( h^* \) is a Bayes classifier.

(b) For any \( h \), \( R(h) - R^* = 2\mathbb{E}_X \left[ \eta(X) - \frac{1}{2} 1_{\{h(X) \neq h^*(X)\}} \right] \).

(c) \( R^* = \mathbb{E}_X \left[ \min(\eta(X), 1 - \eta(X)) \right] \).

**Proof.** We know that for any \( h \),

\[
R(h) = \mathbb{E}_{XY} \left[ 1_{\{h(X) \neq Y\}} \right] \\
= \mathbb{E}_X \mathbb{E}_{Y|X} \left[ 1_{\{h(X) \neq Y\}} \right] \\
= \mathbb{E}_X [\eta(X) 1_{\{h(X) = 1\}} + (1 - \eta(X)) 1_{\{h(X) = 0\}}].
\]

To minimize \( R(h) \), it suffices to for \( h(x) \) to be such that \( \forall x, \)

\[
\eta(x) 1_{\{h(x) = 0\}} + (1 - \eta(x)) 1_{\{h(x) = 1\}}
\]

is minimized. We also note that the indicators here are mutually exclusive, so it suffices to take

\[
h(x) = \begin{cases} 1 & \text{if } \eta(x) \geq 1 - \eta(x) \\ 0 & \text{otherwise} \end{cases}
\]

Therefore \( R(h^*) = R^* \). This proves part (a).
To prove (b), notice
\[ R(h) - R^* = R(h) - R(h^*) = \mathbb{E}_X [\eta(X)1_{h(X)=0} + (1 - \eta(X))1_{h(X)=1} - \eta(X)1_{h^*(X)=0} - (1 - \eta(X))1_{h^*(X)=1}] = \mathbb{E}_X \left[ 2\eta(X) - 1 \right] 1_{\{h(X)\neq h^*(X)\}} = 2\mathbb{E}_X \left[ \eta(X) - \frac{1}{2} \right] 1_{\{h(X)\neq h^*(X)\}}, \]
where the third equality holds by considering the cases in Table 1.

Table 1: This table shows the possible combinations of values of the argument to the expectation above given the possible values of \( h(x) \) and \( h^*(x) \). From this, we can simplify the expression for the expectation.

Finally, (c) follows from the definition of \( h^* \):
\[ R(h^*) = \mathbb{E}_X [\eta(X)1_{h^*(X)=0} + (1 - \eta(X))1_{h^*(X)=1}] = \mathbb{E}_X \left[ \min(\eta(X), 1 - \eta(X)) \right]. \]

Remark. By (b), \( h^* \) can be redefined arbitrarily for any \( x \) such that \( \eta(x) = \frac{1}{2} \) and still be a Bayes classifier. People often refer to \( h^* \) as the Bayes classifier.

Remark. From (c), we see that \( \eta \) determines the difficulty of the classification problem. Figure 1 shows a setting where the Bayes risk is small, and Figure 2 shows a case where it is large.

Remark. As a final remark, we note that the Bayes classifier can be expressed in different equivalent forms. Assume that there exist class-conditional densities \( p_0, p_1 \). Let \( \pi_y = P_Y(Y = y) \), the prior probability of class \( y \). By Bayes’ rule,
\[ \eta(x) = \frac{\pi_1 p_1(x)}{\pi_1 p_1(x) + \pi_0 p_0(x)} = \frac{1}{1 + \frac{\pi_0 p_0(x)}{\pi_1 p_1(x)}}. \]

This is equivalent to the likelihood ratio test
\[ \frac{p_1(x)}{p_0(x)} \geq \frac{\pi_0}{\pi_1} \iff \eta(x) \geq \frac{1}{2}. \]
Figure 1: An easy classification problem. In the case where $X \sim \text{unif}[0,1]$, the area of the shaded region equals the Bayes risk.

Figure 2: A hard classification problem. In the case where $X \sim \text{unif}[0,1]$, the area of the shaded region equals the Bayes risk.
3 Plug-in Classifiers

A plug-in classifier is based on an estimate of $\eta$. This estimate is then plugged in to the formula for $h^*$. Thus, suppose that $\hat{\eta}_n$ is an estimate of $\eta$ based on $(X_i, Y_i)$, $i = 1, \ldots, n$. We define $\hat{h}_n(x)$ as

$$
\hat{h}_n(x) = \begin{cases} 
1 & \text{if } \hat{\eta}_n(x) \geq \frac{1}{2} \\
0 & \text{otherwise.}
\end{cases}
$$

The following result follows from Theorem 1. The proof is left as an exercise.

**Corollary 1.**

$$
R(\hat{h}_n) - R^* \leq 2E_X[|\eta(X) - \hat{\eta}_n(X)|]
$$

Therefore, if $E_X[|\eta(X) - \hat{\eta}_n(X)|]$ approaches zero (in probability/almost surely) then the classifier $\hat{h}_n$ is (weakly/strongly) consistent. However, if classification is the goal, then the plug-in approach may be unwise because estimating $\eta$ is potentially much harder than estimating $h$. Section 3 shows an example of an $\eta(x)$ which would be harder to accurately estimate than the $h^*(X)$ derived from it. It should be noted, however, that sometimes estimation of $\eta$ is also of interest, a problem known as class probability estimation. One popular method for solving this problem is logistic regression.

![Figure 3: Estimating $\eta$ could be much harder than estimating the “level set” $\{x : \eta(x) \geq \frac{1}{2}\}$.

Exercises

1. Extend Theorem 1 to the multiclass case, $Y = \{1, 2, \ldots, M\}$. Part (b) may or may not have a nice generalization.

2. Let $\alpha \in (0, 1)$. Define the $\alpha$-cost-sensitive risk of a classifier $h$ to be

$$
R_\alpha(h) := E_{X,Y}[(1 - \alpha)1_{\{Y=1,h(X)=0\}} + \alpha 1_{\{Y=0,h(X)=1\}}].
$$

Determine the Bayes classifier and prove an analogue of Theorem 1 for this risk.

3. Prove Corollary 1.