

# Balanced error rate

Daniel Hsu

November 5, 2025

Suppose  $(X, Y, A)$  is a triple of random variables, where  $X$  is a feature vector,  $Y$  is a  $\{0, 1\}$ -valued label, and  $A$  is a  $\{0, 1\}$ -valued random variable indicating subgroup membership. (Here, it is possible that  $A$  depends on  $(X, Y)$ .) Your training and test data are sampled from a source distribution in which

$$\Pr_{\text{src}}(A = 0) \ll \Pr_{\text{src}}(A = 1),$$

but in the target distribution,

$$\Pr_{\text{tgt}}(A = 0) = \Pr_{\text{tgt}}(A = 1) = \frac{1}{2}.$$

A common approach to dealing with this sort of distribution shift is to use *importance weights*. That is, for any example with  $A = a$ , we assign it an importance weight of

$$i_a = \frac{1/2}{\Pr_{\text{src}}(A = a)}.$$

The importance weight is then used to scale any quantity of interest in an (empirical) expectation computation. (It functions as a “change of measure”.)

For example, suppose test data  $(\tilde{X}^{(1)}, \tilde{Y}^{(1)}, \tilde{A}^{(1)}), \dots, (\tilde{X}^{(m)}, \tilde{Y}^{(m)}, \tilde{A}^{(m)})$  are drawn iid from the source distribution. The test error rate of a classifier  $f$  is

$$\frac{1}{m} \sum_{i=1}^m \mathbb{1}\{f(\tilde{X}^{(i)}) \neq \tilde{Y}^{(i)}\}.$$

This quantity estimates the error rate of  $f$  under the source distribution. The importance-weighted test error rate of a classifier  $f$  is

$$\frac{1}{m} \sum_{i=1}^m i_{\tilde{A}^{(i)}} \cdot \mathbb{1}\{f(\tilde{X}^{(i)}) \neq \tilde{Y}^{(i)}\}.$$

What does this quantity estimate? The expectation of the importance-weighted test error rate is

$$\begin{aligned}
& \mathbb{E}_{\text{src}} \left[ \frac{1}{m} \sum_{i=1}^m i_{\tilde{A}^{(i)}} \cdot \mathbf{1}\{f(\tilde{X}^{(i)}) \neq \tilde{Y}^{(i)}\} \right] \\
&= \mathbb{E}_{\text{src}} \left[ i_{\tilde{A}^{(1)}} \cdot \mathbf{1}\{f(\tilde{X}^{(1)}) \neq \tilde{Y}^{(1)}\} \right] \\
&= \sum_{a \in \{0,1\}} \Pr_{\text{src}}(\tilde{A}^{(1)} = a) \cdot \mathbb{E}_{\text{src}} \left[ \frac{1/2}{\Pr_{\text{src}}(\tilde{A}^{(1)} = a)} \cdot \mathbf{1}\{f(\tilde{X}^{(1)}) \neq \tilde{Y}^{(1)}\} \mid \tilde{A}^{(1)} = a \right] \\
&= \sum_{a \in \{0,1\}} \frac{1}{2} \cdot \mathbb{E}_{\text{src}} \left[ \mathbf{1}\{f(\tilde{X}^{(1)}) \neq \tilde{Y}^{(1)}\} \mid \tilde{A}^{(1)} = a \right].
\end{aligned}$$

If  $A = Y$ , then this is called the *balanced error rate* of  $f$  under the source distribution. If we additionally assume that the conditional distributions of  $(X, Y)$  given  $A = a$  under the source distribution is the same as that under the target distribution (for each  $a \in \{0, 1\}$ ), then we can further conclude that

$$\begin{aligned}
& \sum_{a \in \{0,1\}} \frac{1}{2} \cdot \mathbb{E}_{\text{src}} \left[ \mathbf{1}\{f(\tilde{X}^{(1)}) \neq \tilde{Y}^{(1)}\} \mid \tilde{A}^{(1)} = a \right] \\
&= \sum_{a \in \{0,1\}} \Pr_{\text{tgt}}(\tilde{A}^{(1)} = a) \cdot \mathbb{E}_{\text{tgt}} \left[ \mathbf{1}\{f(\tilde{X}^{(1)}) \neq \tilde{Y}^{(1)}\} \mid \tilde{A}^{(1)} = a \right] \\
&= \mathbb{E}_{\text{tgt}} \left[ \mathbf{1}\{f(\tilde{X}^{(1)}) \neq \tilde{Y}^{(1)}\} \right],
\end{aligned}$$

which is the error rate of  $f$  under the target distribution.