

## Problem

Let  $(X, Y)$  be a pair of random variables taking values in  $\mathbb{R}^d \times \{0, 1\}$ , and let  $g$  be a classifier taking value 1 on the Borel subset  $G$  of  $\mathbb{R}^d$  and 0 elsewhere. In other words, for all  $x \in \mathbb{R}^d$ ,

$$g(x) = \mathbb{1}_{[x \in G]}.$$

Throughout,  $\mu$  is the distribution of  $X$  and  $\eta$  is the regression function

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

We also let  $g^*$  be the Bayes rule associated with  $(X, Y)$  and  $L^*$  be the Bayes risk, that is,

$$L^* = \mathbb{P}(g^*(X) \neq Y).$$

1. Prove that

$$g^*(x) = \mathbb{1}_{[x \in G^*]}, \quad x \in \mathbb{R}^d,$$

where  $G^*$  is some measurable set.

2. Show that

$$\mathbb{P}(g(X) \neq Y) - L^* = \int_{\mathbb{R}^d} |2\eta(x) - 1| \mathbb{1}_{[g(x) \neq g^*(x)]} \mu(dx).$$

3. Let

$$d(G, G^*) = \mathbb{P}(g(X) \neq Y) - L^*.$$

Conclude from the above that

$$d(G, G^*) = \int_{G \Delta G^*} |2\eta(x) - 1| \mu(dx),$$

where  $\Delta$  is the symmetric difference operator<sup>1</sup>.

---

<sup>1</sup>For two sets  $A$  and  $B$ ,  $A \Delta B = (A \cap B^c) \cup (A^c \cap B)$ .

4. What can be said when  $d(G, G^*) = 0$ ?

From now on, we let

$$d_{\Delta}(G, G^*) = \mu(G \Delta G^*).$$

We also denote by **H** the following assumption:

(**H**) There exist  $\kappa \geq 1$ ,  $c_0 > 0$ , and  $\varepsilon_0 \in (0, 1]$  such that

$$d(G, G^*) \geq c_0 d_{\Delta}^{\kappa}(G, G^*)$$

as soon as  $G$  satisfies  $d_{\Delta}(G, G^*) \leq \varepsilon_0$ .

5. Prove that  $d(G, G^*) \leq d_{\Delta}(G, G^*) \leq 1$ .
6. Assume now that, for all  $t \in (0, t^*]$  (where  $0 < t^* \leq 1/2$ ), one has

$$\mathbb{P}(|\eta(X) - 1/2| \leq t) \leq C_{\eta} t^{\alpha}, \quad (1)$$

where  $C_{\eta}$  and  $\alpha$  are two positive constants. Give an interpretation of this assumption by a careful examination of the cases  $\alpha \rightarrow 0$  and  $\alpha \rightarrow \infty$ .

7. **An example.** Assume that  $d = 1$  and that  $X$  has a bounded probability density. Assume in addition that, in a neighborhood of 0,  $\eta(x) = 1/2 + x^{1/\alpha}$  for  $x \geq 0$  and  $\eta(x) = 1/2 - (-x)^{1/\alpha}$  for  $x < 0$ , and that  $\eta(x)$  is away from  $1/2$  everywhere else. Prove that assumption (1) is satisfied.
8. Prove that, under assumption (1), one has, for all  $t \in (0, t^*]$ ,

$$d(G, G^*) \geq 2t [d_{\Delta}(G, G^*) - C_{\eta} t^{\alpha}].$$

9. Deduce that assumption (1) implies assumption **H**, with explicit constants  $\kappa$ ,  $c_0$ , and  $\varepsilon_0$ .
10. Prove that, under assumption (1), one has, for all  $\delta \in (0, t^*]$ ,

$$d(G, G^*) \leq 2C_{\eta} \delta^{1+\alpha} + \mathbb{E} \left( |2\eta(X) - 1| \mathbb{1}_{[g(X) \neq g^*(X)]} \mathbb{1}_{||\eta(X) - 1/2| > \delta} \right).$$

Let us now be given a sample  $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$  of independent random variables, all distributed as (and independent of) the pair  $(X, Y)$ , and let  $\eta_n$  be an estimate of the regression function  $\eta$ .

11. How can we naturally define a classifier  $g_n$  and a (random) associated set  $G_n$ ?
12. Show that, under assumption (1), one has, for all  $\delta \in (0, t^*]$ ,
$$\mathbb{E}d(G_n, G^*) \leq 2C_\eta \delta^{1+\alpha} + 2\mathbb{E} \left( |\eta_n(X) - \eta(X)| \mathbb{1}_{[|\eta_n(X) - \eta(X)| > \delta]} \right).$$
13. Interpret this result.