Statistical learning, 2023-2024
Gérard Biau

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, μ is the distribution of X and η is the regression function

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

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

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

1. Prove that
$$g^\star(x) = \mathbb{1}_{[x \in G^\star]}, \quad x \in \mathbb{R}^d,$$
where $G^\star$ is some measurable set.

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

3. Let
$$d(G, G^\star) = \mathbb{P}(g(X) \neq Y) - L^\star.$$
Conclude from the above that
$$d(G, G^\star) = \int_{G \triangle G^\star} |2\eta(x) - 1| \mu(\mathrm{d}x),$$
where $\triangle$ is the symmetric difference operator[1].

[1] For two sets A and B, $A \triangle B = (A \cap B^c) \cup (A^c \cap B)$.

From now on, we let

$$d_\triangle(G, G^\star) = \mu(G \triangle G^\star).$$

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

$$d(G, G^\star) \geq c_0 d_\triangle^\kappa(G, G^\star)$$

as soon as G satisfies $d_\triangle(G, G^\star) \leq \varepsilon_0$.

4. Prove that $d(G, G^\star) \leq d_\triangle(G, G^\star) \leq 1$.

5. Assume now that, for all $t \in (0, t^\star]$ (where $0 < t^\star \leq 1/2$), one has

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

where C_η and α are two positive constants. Give an interpretation of this assumption by a careful examination of the cases $\alpha \to 0$ and $\alpha \to \infty$.

6. **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.

7. Prove that, under assumption (1), one has, for all $t \in (0, t^\star]$,

$$d(G, G^\star) \geq 2t\,[d_\triangle(G, G^\star) - C_\eta t^\alpha]\,.$$

8. Deduce that assumption (1) implies assumption **H**, with explicit constants κ, c_0, and ε_0.

9. Prove that, under assumption (1), one has, for all $\delta \in (0, t^\star]$,

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

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

10. How can we naturally define a classifier g_n and a (random) associated set G_n?

11. Show that, under assumption (1), one has, for all $\delta \in (0, t^\star]$,

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

12. Interpret this result.