Statistical learning, 2021-2022
Gérard Biau

Problem

Throughout the problem, we let $\mathscr{B}$ be the Borel subsets of $\mathbb{R}^d$.

A. Preliminaries. Let f and g be two probability densities on $\mathbb{R}^d$, that is, nonnegative functions such that

$$\int f = \int g = 1.$$

(All integrals are evaluated with respect to the Lebesgue measure.)

1. Show that
$$\int |f-g| = 2\int_{A_{fg}} (f-g),$$
where A_{fg} is the set $\{f > g\}$, i.e.,
$$A_{fg} = \big\{x \in \mathbb{R}^d : f(x) > g(x)\big\}.$$

2. Deduce that
$$\int |f-g| = 2 \sup_{B\in\mathscr{B}} \Big| \int_B f - \int_B g \Big|.$$

This result is known as *Scheffé's theorem.*

B. A selection problem. Assume we are given a sample of independent random variables $X_1, \ldots, X_n$ with common **unknown** density f. We denote by $\mathscr{F}$ a collection of densities parameterized by θ:

$$\mathscr{F} = \{f_\theta : \theta \in \Theta\}.$$

Our goal is to select in $\mathscr{F}$ the "best" possible density, using only $X_1, \ldots, X_n$.

1. Let μ_n be the empirical measure associated with $X_1, \ldots, X_n$. Explain why the strategy that chooses θ in Θ by minimizing the quantity
$$\sup_{B\in\mathscr{B}} \Big| \int_B f_\theta - \mu_n(B) \Big|$$
is not a good idea.

2. Introduce the collection of sets

$$\mathscr{A} = \big\{\{f_\theta > f_{\theta'}\} : (\theta, \theta') \in \Theta^2\big\}.$$

In order to choose the "best" density in $\mathscr{F}$, a possible route is to minimize in θ the following criterion:

$$\Delta(\theta) = \sup_{A \in \mathscr{A}} \Big| \int_A f_\theta - \mu_n(A) \Big|.$$

We denote by $\theta^\star$ an element of Θ such that $\Delta(\theta^\star) = \inf_{\theta \in \Theta} \Delta(\theta)$.

2.a Let $\bar{\theta}$ be an element of Θ such that

$$\int |f_{\bar{\theta}} - f| = \inf_{\theta \in \Theta} \int |f_\theta - f|.$$

Prove that

$$\int |f_{\theta^\star} - f_{\bar{\theta}}| \leq 4 \sup_{A \in \mathscr{A}} \Big| \int_A f_{\bar{\theta}} - \mu_n(A) \Big|.$$

2.b Next, show that

$$\int |f_{\theta^\star} - f| \leq 3 \inf_{\theta \in \Theta} \int |f_\theta - f| + 4\Delta_n,$$

where Δ_n is some explicit random quantity.

3. 3.a Recall the definition of $\mathbf{S}_{\mathscr{A}}(n)$, the shatter coefficient of n points by the class $\mathscr{A}$.

3.b Show that

$$\mathbb{E}\Big(\int |f_{\theta^\star} - f| \Big) \leq 3 \inf_{\theta \in \Theta} \int |f_\theta - f| + \mathrm{O}\left(\sqrt{\frac{\log(\mathbf{S}_{\mathscr{A}}(n))}{n}} \right).$$

3.c Provide a statistical interpretation of this inequality.

C. Application. On the real line $\mathbb{R}$, we let $\mathscr{F}$ be the set of Gaussian densities, parameterized by their mean and variance, i.e.,

$$\mathscr{F} = \Big\{ f_\theta(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-(x-m)^2/(2\sigma^2)} : \theta = (m, \sigma^2) \in \mathbb{R} \times (0, \infty) \Big\}.$$

1. Prove that $\mathscr{A}$ is contained in a class of sets $\mathscr{B}_2$ that can be easily described.

2. Determine the Vapnik-Chervonenkis dimension V of $\mathscr{B}_2$.

3. Conclude that

$$\mathbb{E}\Big(\int |f_{\theta^\star} - f|\Big) \leq 3 \inf_{\theta \in \Theta} \int |f_\theta - f| + \mathrm{O}\left(\sqrt{\frac{V \log n}{n}}\right).$$