Part II: Multi-class AdaBoost

We want to address the K-class supervised classification problem. We observe n independent replicates $(X^{(i)}, Y^{(i)})_{1\leq i\leq n}$ of (X, Y), where $X \in \mathbb{R}^d$ and $Y \in \mathcal{Y} = \{1, \ldots, K\}$. The goal is to construct a classifier $g : \mathbb{R}^d \to \mathcal{Y}$ that predicts Y accurately via $g(X)$. In what follows,

$$\forall\, x \in \mathbb{R}^d,\ \forall\, k \in \mathcal{Y},\ \pi_k(x) = \mathbb{P}(Y = k|X = x).$$

Q1: Derive the classifier $g^\star$ that minimizes $R(g) = \mathbb{P}(g(X) \neq Y)$.

The class Y can be encoded as a vector $Z = (Z_1, \ldots, Z_K)$ of $\mathbb{R}^K$: $Y = k$ if and only if $Z_k = 1$ and $Z_j = \frac{-1}{K-1}$ for $j \neq k$. Let $\mathcal{Z}$ denote the set of possible values of Z.

For any $z \in \mathcal{Z}$ and $f \in \mathbb{R}^K$, consider the function

$$L(z, f) = \exp\left(-\frac{1}{K}\sum_{k=1}^{K} z_k f_k\right).$$

Q2: Let $x \in \mathbb{R}^d$. Show that the vector $f^\star(x) = (f_1^\star(x), \ldots, f_K^\star(x))$ minimizing the function $f \mapsto \mathbb{E}[L(Z, f)|X = x]$ under the constraint $f_1 + \cdots + f_K = 0$ is given by

$$\forall\, 1 \leq k \leq K,\ f_k^\star(x) = (K-1)\left[\ln\, \pi_k(x) - \frac{1}{K}\sum_{j=1}^{K} \ln\, \pi_j(x)\right].$$

Q3: Deduce that $\pi_k(x) \propto \exp(f_k^\star(x)/(K-1))$, and express $g^\star$ in terms of the components of $f^\star$.

From now on, let $\mathcal{H}$ denote a finite set of functions from $\mathbb{R}^d$ to $\mathcal{Z}$.

Q4: For any positive weights $(\omega^{(i)})_{1\leq i\leq n}$, show that the pair $(\beta^\star, h^\star)$ minimizing

$$(\beta, h) \mapsto \sum_{i=1}^{n} \omega^{(i)} L(Z^{(i)}, \beta h(X^{(i)}))$$

under the constraints $\beta > 0$ and $h \in \mathcal{H}$, is given by

$$\begin{cases} h^\star &= \underset{h\in\mathcal{H}}{\arg\min} \sum_{i=1}^{n} \omega^{(i)} \mathbb{I}_{\{h(X^{(i)})\neq Z^{(i)}\}}, \\ \beta^\star &= \dfrac{(K-1)^2}{K}\alpha^\star, \end{cases}$$

with

$$\alpha^\star = \ln\left(\frac{1-\epsilon^\star}{\epsilon^\star}\right) + \ln(K-1) \qquad \text{and} \qquad \epsilon^\star = \frac{\sum_{i=1}^{n} \omega^{(i)} \mathbb{I}_{\{h^\star(X^{(i)})\neq Z^{(i)}\}}}{\sum_{i=1}^{n} \omega^{(i)}}.$$

Our aim is to approximate $f^\star$ as a linear combination of functions in $\mathcal{H}$. Specifically, we consider

$$f_M = \sum_{m=1}^{M} \beta_m h_m,$$

where $\beta_m > 0$ and $h_m \in \mathcal{H}$ for all m.

Q5: Show that

$$\sum_{i=1}^{n} L(Z^{(i)}, f_M(X^{(i)})) = \sum_{i=1}^{n} \omega_M^{(i)} L(Z^{(i)}, \beta_M h_M(X^{(i)})),$$

where $(\omega_m)_{1\leq m\leq M}$ satisfies a recurrence relation to be determined.

Q6: Using the results of Q4 and Q5, propose an algorithm to estimate $f^\star$ (as defined in Q2).

Q7: Show that the classifier $G : \mathbb{R}^d \to \mathcal{Y}$ derived from Q3 and Q6 is given by

$$G(x) = \arg\max_{k\in\mathcal{Y}} \sum_{m=1}^{M} \alpha_m \mathbb{I}_{\{g_m(x)=k\}},$$

where $(\alpha_m)_{1\leq m\leq M}$ and $(g_m)_{1\leq m\leq M}$ are to be specified.

Q8: Identify where the non-trivial generalization lies compared to AdaBoost.