MAP569 Machine Learning II

PC6 : Ada Boost and random forests

Ada Boost

Let $(x_i, y_i)_{1\leqslant i\leqslant n} \in (\mathsf{X}\times\{-1,1\})^n$ be n observations and $\mathsf{H} = \{h_1, \ldots, h_M\}$ be a set of M classifiers, i.e. for all $1 \leqslant i \leqslant M$, : $h_i : \mathsf{X} \to \{-1, 1\}$. It is assumed that for each $h \in \mathsf{H}$, $-h \in \mathsf{H}$ and that there exist $1 \leqslant i \neq j \leqslant n$ such that $y_i = h(x_i)$ and $y_j \neq h(x_j)$. Let F be the set of all linear combinations of elements of H:

$$\mathsf{F} = \left\{ \sum_{j=1}^{M} \theta_j h_j \, ; \, \theta \in \mathbb{R}^M \right\} .$$

Consider the following algorithm. Set $\hat{f}_0 = 0$ and for all $1 \leqslant m \leqslant M$,

$$\hat{f}_m = \hat{f}_{m-1} + \beta_m h_{j_m} \quad \text{where} \quad (\beta_m, h_{j_m}) = \underset{h\in\mathsf{H}\,,\,\beta\in\mathbb{R}}{\operatorname{argmin}}\; n^{-1} \sum_{i=1}^{n} \exp\left\{-y_i \left(\hat{f}_{m-1}(x_i) + \beta h(x_i)\right)\right\} .$$

1. Choosing $\omega_i^m = n^{-1}\exp\{-y_i \hat{f}_{m-1}(x_i)\}$, show that

$$n^{-1} \sum_{i=1}^{n} \exp\left\{-y_i \left(\hat{f}_{m-1}(x_i) + \beta h(x_i)\right)\right\} = \left(\mathrm{e}^{\beta} - \mathrm{e}^{-\beta}\right) \sum_{i=1}^{n} \omega_i^m \mathbb{1}_{h(x_i)\neq y_i} + \mathrm{e}^{-\beta} \sum_{i=1}^{n} \omega_i^m .$$

Solution.

We have

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} \exp\left\{-y_i \left(\hat{f}_{m-1}(x_i) + \beta h(x_i)\right)\right\} &= \mathrm{e}^{-\beta} \sum_{i=1}^{n} \omega_i^m \mathbb{1}_{h(x_i)= y_i} + \mathrm{e}^{\beta} \sum_{i=1}^{n} \omega_i^m \mathbb{1}_{h(x_i)\neq y_i} ,\\ &= \mathrm{e}^{-\beta} \sum_{i=1}^{n} \omega_i^m \left(1 - \mathbb{1}_{h(x_i)\neq y_i}\right) + \mathrm{e}^{\beta} \sum_{i=1}^{n} \omega_i^m \mathbb{1}_{h(x_i)\neq y_i} ,\\ &= \left(\mathrm{e}^{\beta} - \mathrm{e}^{-\beta}\right) \sum_{i=1}^{n} \omega_i^m \mathbb{1}_{h(x_i)\neq y_i} + \mathrm{e}^{-\beta} \sum_{i=1}^{n} \omega_i^m . \end{aligned}$$

□

2. For all $1 \leqslant m \leqslant M$ and $h \in \mathsf{H}$, define

$$\mathrm{err}_m(h) = \frac{\sum_{i=1}^n \omega_i^m \mathbb{1}_{h(x_i)\neq y_i}}{\sum_{i=1}^n \omega_i^m} .$$

Prove that

$$h_{j_m} = \underset{h\in\mathsf{H}}{\operatorname{argmin}}\; \mathrm{err}_m(h) \quad \text{and} \quad \beta_m = \frac{1}{2}\log\left(\frac{1-\mathrm{err}_m(h_{j_m})}{\mathrm{err}_m(h_{j_m})}\right) .$$

Solution.

According to the previous question,

$$h_{j_m} = \operatorname*{argmin}_{h \in \mathsf{H}} \sum_{i=1}^{n} \omega_i^m \mathbb{1}_{h(x_i) \neq y_i} = \operatorname*{argmin}_{h \in \mathsf{H}} \mathrm{err}_m(h) \ .$$

On the other hand, β_m is solution to

$$\left(\mathrm{e}^{\beta_m} + \mathrm{e}^{-\beta_m}\right) \sum_{i=1}^{n} \omega_i^m \mathbb{1}_{h(x_i) \neq y_i} - \mathrm{e}^{-\beta_m} \sum_{i=1}^{n} \omega_i^m = 0 \ ,$$

which yields

$$\mathrm{e}^{2\beta_m} \sum_{i=1}^{n} \omega_i^m \mathbb{1}_{h(x_i) \neq y_i} = \sum_{i=1}^{n} \omega_i^m - \sum_{i=1}^{n} \omega_i^m \mathbb{1}_{h(x_i) \neq y_i}$$

and concludes the proof. □

3. Propose an algorithm to compute $\hat{f}_M$.
   Solution.

Note that for all $1 \leqslant i \leqslant n$ and all $h \in \mathsf{H}$, $-y_i h(x_i) = 2\mathbb{1}_{y_i \neq h(x_i)} - 1$, then for all $1 \leqslant m \leqslant M$,

$$\omega_i^{m+1} = \omega_i^m \mathrm{e}^{-\beta_m y_i h_{j_m}(x_i)} = \omega_i^m \mathrm{e}^{2\beta_m \mathbb{1}_{y_i \neq h_{j_m}(x_i)}} \mathrm{e}^{-\beta_m} \ .$$

As the value of $\mathrm{err}_m(h)$ does not depend on the normalizing constant of the ω_i^m, $1 \leqslant i \leqslant n$, consider the following algorithm. For all $1 \leqslant i \leqslant n$, set $\omega_i^1 = 1/n$. Then, for $1 \leqslant m \leqslant M$,

(a) $h_{j_m} = \operatorname*{argmin}_{h \in \mathsf{H}} \mathrm{err}_m(h)$.

(b) $\beta_m = \left[\log\left(1 - \mathrm{err}_m(h_{j_m})\right) - \log\left(\mathrm{err}_m(h_{j_m})\right)\right]/2$.

(c) $\omega_i^{m+1} = \omega_i^m \mathrm{e}^{2\beta_m \mathbb{1}_{y_i \neq h_{j_m}(x_i)}}$.

The classifier obtained at the end of the algorithm is given by:

$$\hat{f}_M = \sum_{m=1}^{M} \beta_m h_{j_m} \ .$$

□

Consistency of a simple random forest

Consider a data set $\mathcal{D}_n = \{(X_i, Y_i) \in [0,1]^d \times \mathbb{R}, i = 1, \ldots, n\}$. It is assumed that the (X_i, Y_i) are i.i.d. with the same distribution as (X, Y) where

$$Y = r(X) + \varepsilon,$$

with ε a centered Gaussian noise, independent of X and r a uniformly continuous function. Define the following centered random forest estimator:

1. Grow M trees as follows:
   (a) Consider the cell $[0,1]^d$.
   (b) Select uniformly one variable $j^\star$ in $\{1, \ldots, d\}$.
   (c) Cut the cell at the middle of the $j^\star$-th side, where $j^\star$ is the coordinate chosen above.
   (d) For each of the two resulting cells, repeat $(b) - (c)$ if the cell has been cut strictly less than k_n times.
   (e) For a query point x, the m-th tree outputs the average $\hat{r}_n(x, \Theta_m)$ of the Y_i falling into the same cell as x, where Θ_m is the random variable encoding all selected splitting variables in each cell of the m-th tree.
2. For a query point x, the centered forest outputs the average $\hat{r}_{M,n}(x, \Theta_1, \ldots, \Theta_M)$ of the predictions given by the M trees.

Define the infinite random forest estimate $\hat{r}_{\infty,n}$ by considering the random forest estimate defined above and letting $M \to \infty$, that is

$$\hat{r}_{\infty,n}(x) = \mathbb{E}_\Theta[\hat{r}_n(x, \Theta)],$$

where $\mathbb{E}_\Theta$ is the expectation with respect to Θ only. For a tree built with the randomness Θ, we let $A_n(x, \Theta)$ be the cell containing x and $N_n(x, \Theta)$ be the number of observations falling into $A_n(x, \Theta)$. We want to prove the following theorem:

Theorem 1. Assume that $k_n \to \infty$ is such that $2^{k_n}/n \to 0$, as $n \to \infty$. Then the random forest fulfills $\mathbb{E}[(\hat{r}_{\infty,n}(X) - r(X))^2] \to 0$, where X is independent of $(X_i, Y_i)_{i=1,\dots,n}$ with the same distribution as the X_i on $[0, 1]^d$.

1. Prove that there exists weights $W_{ni}(x, \Theta)$ and $W_{ni}^\infty(x)$, $1 \leqslant i \leqslant n$, such that

$$\hat{r}_n(x, \Theta) = \sum_{i=1}^n W_{ni}(x, \Theta) Y_i, \quad \text{and} \quad \hat{r}_{\infty,n}(x) = \sum_{i=1}^n W_{ni}^\infty(x) Y_i.$$

Solution.

The estimation $\hat{r}_n(x, \Theta)$ outputs by a regression tree is the average of Y_i falling into the cell containing x. Then

$$\hat{r}_n(x, \Theta) = \sum_{i=1}^n \frac{\mathbb{1}_{X_i \in A_n(x,\Theta)}}{N_n(x, \Theta)} Y_i,$$

which gives the first assertion by setting

$$W_{ni}(x, \Theta) = \frac{\mathbb{1}_{X_i \in A_n(x,\Theta)}}{N_n(x, \Theta)}.$$

Regarding the random forest estimate, write

$$\hat{r}_{\infty,n}(x) = \mathbb{E}_\Theta\left[r_n(x, \Theta)\right] = \mathbb{E}_\Theta\left[\sum_{i=1}^n \frac{\mathbb{1}_{X_i \in A_n(x,\Theta)}}{N_n(x, \Theta)} Y_i\right] = \sum_{i=1}^n Y_i \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{X_i \in A_n(x,\Theta)}}{N_n(x, \Theta)}\right].$$

This leads to

$$\hat{r}_{\infty,n}(x) = \sum_{i=1}^n W_{ni}^\infty(x) Y_i,$$

where

$$W_{ni}^\infty(x) = \mathbb{E}_\Theta\left[\frac{\mathbb{1}_{X_i \in A_n(x,\Theta)}}{N_n(x, \Theta)}\right].$$

□

In this context, Stone's Theorem states that the random tree estimate $\hat{r}_n(x, \Theta)$ fulfills

$$\lim_{n\to\infty} \mathbb{E}\left[(\hat{r}_n(X, \Theta) - r(X))^2\right] = 0,$$

as soon as the two following conditions are satisfied

(i) $\mathbb{E}[\mathrm{diam}(A_n(X, \Theta))] \to 0$, as $n \to \infty$, where the diameter of any cell A is defined as

$$\mathrm{diam}(A) = \sup_{x,z \in A} \|x - z\|_2.$$

(ii) $N_n(X, \Theta) \to \infty$ in probability, as $n \to \infty$.

2. Let $x \in [0, 1]^d$. What is the distribution of the number of cuts along the coordinate $j \in \{1, \dots, d\}$ in the cell $A_n(x, \Theta)$?
Solution.

Let $K_{nj}(x,\theta)$ be the number of cuts along the j-th coordinate in the cell $A_{nj}(x,\Theta)$. Then, $K_{nj}(x,\theta)$ has a binomial $\mathcal{B}(k_n, 1/d)$ distribution. $\square$

3. Check that, for all $x \in [0,1]^d$,

$$\mathbb{E}\left[\sup_{z\in A_n(x,\Theta)} z_j - \inf_{z\in A_n(x,\Theta)} z_j\right] = \left(1-\frac{1}{2d}\right)^{k_n}.$$

Solution.

Let $V_{nj}(x,\Theta)$ be the size of the j-th dimension of the rectangle containing x. Each time there is a cut at the j-th dimension of the rectangle, the size along this dimension is divided by two. Therefore,

$$V_{nj}(x,\Theta) = 2^{-K_{nj}(x,\Theta)}.$$

Since $K_{nj}(x,\Theta)$ follows the binomial $\mathcal{B}(k_n,1/d)$ distribution, we have

$$\mathbb{E}[V_{nj}(x,\Theta)] = \mathbb{E}[2^{-K_{nj}(x,\Theta)}] = \left(1\times\left(1-\frac{1}{d}\right)+\frac{1}{2}\times\frac{1}{d}\right)^{k_n} = \left(1-\frac{1}{2d}\right)^{k_n},$$

which concludes the proof. $\square$

4. Prove that (i) holds for a random centered tree.
Solution.

Note that

$$\mathbb{E}[\mathrm{diam}(A_n(X,\Theta))]^2 \leqslant \mathbb{E}\left[(\mathrm{diam}(A_n(X,\Theta)))^2\right] \leqslant \mathbb{E}\left[\sum_{j=1}^{d} V_{nj}(X,\Theta)^2\right] \leqslant \sum_{j=1}^{d}\mathbb{E}\left[V_{nj}(X,\Theta)\right],$$

which tends to zero, according to the previous question. $\square$

5. We denote by $A_1,\dots,A_{2^{k_n}}$ the 2^{k_n} cells and by N_ℓ the number of points among $X, X_1,\dots,X_n$ which falls into A_ℓ. Then, show that for $\ell \in \{1,\dots,2^{k_n}\}$,

$$\mathbb{P}\left(X\in A_\ell | N_\ell\right) = \frac{N_\ell}{n+1}.$$

Conclude that for every integer $t \geqslant 1$,

$$\mathbb{P}\left(N_n(X,\Theta)\leqslant t\right) \leqslant t2^{k_n}/(n+1).$$

Solution.

For all $1\leqslant k\leqslant n+1$,

$$\mathbb{P}(X\in A_\ell|N_\ell = k) = \frac{\mathbb{P}(X\in A_\ell; N_\ell = k)}{\mathbb{P}(N_\ell = k)},$$

where, by writing $X_{n+1} = X$,

$$\mathbb{P}(X\in A_\ell; N_\ell = k) = \sum_{1\leqslant i_1<\ldots i_{k-1}\leqslant n} \mathbb{P}(X_{n+1}\in A_\ell; X_{i_1}\in A_\ell,\dots,X_{i_{k-1}}\in A_\ell; X_{j\notin\{n+1,i_1,\dots,i_{k-1}\}}\notin A_\ell),$$

$$\mathbb{P}(N_\ell = k) = \sum_{1\leqslant i_1<\ldots i_k\leqslant n+1} \mathbb{P}(X_{i_1}\in A_\ell,\dots,X_{i_k}\in A_\ell; X_{j\notin\{i_1,\dots,i_k\}}\notin A_\ell).$$

As the $(X_i)_{1\leqslant i\leqslant n+1}$ are i.i.d. the probabilities on the r.h.s. are equal and constant which yields

$$\mathbb{P}(X\in A_\ell|N_\ell = k) = \frac{\binom{n}{k-1}}{\binom{n+1}{k}} = \frac{k}{n+1}.$$

Thus, for every fixed $t\geqslant 1$,

$$\mathbb{P}(N_n(X,\Theta)\leqslant t) = \mathbb{E}\left[\mathbb{P}(N_n(X,\Theta)\leqslant t|\Theta)\right].$$

On the other hand,

$$
\begin{aligned}
\mathbb{P}(N_n(X,\Theta) \leqslant t|\Theta) &= \mathbb{P}(\{N_n(X,\Theta) \leqslant t\} \cap \{X \in [0,1]^d\}|\Theta)\,,\\
&= \sum_{\ell=1}^{2^{k_n}} \mathbb{P}(\{N_n(X,\Theta) \leqslant t\} \cap \{X \in A_\ell\}|\Theta)\,,\\
&= \sum_{\ell=1}^{2^{k_n}} \mathbb{E}[\mathbb{E}[\mathbb{1}_{X\in A_\ell}|N_n(X,\Theta);\Theta]\mathbb{1}_{N_n(X,\Theta)\leqslant t}|\Theta]\,,\\
&\leqslant \sum_{\ell=1}^{2^{k_n}} \frac{t}{n+1} = \frac{t2^{k_n}}{n+1}\,.
\end{aligned}
$$

□

6. Prove that the infinite centered random forest fulfills $\mathbb{E}[(\hat{r}_{\infty,n}(X) - r(X))^2] \to 0$, as $n \to \infty$.
Solution.

Combining the Stone Theorem and Jensen's inequality yields

$$
\begin{aligned}
\mathbb{E}[\hat{r}_{\infty,n}(X) - r(X)]^2 = \mathbb{E}[\mathbb{E}_\Theta[\hat{r}_n(X)] - r(X)]^2 &= \mathbb{E}[\mathbb{E}_\Theta[\hat{r}_n(X) - r(X)]]^2,\\
&\leqslant \mathbb{E}[\mathbb{E}_\Theta[\hat{r}_n(X) - r(X)]^2],\\
&\leqslant \mathbb{E}[\hat{r}_n(X) - r(X)]^2
\end{aligned}
$$

and the rhs goes to 0 as $n \to \infty$ according to the previous question.

□

7. Assume that the noise ε is Gaussian. Thus,

$$\mathbb{E}\left[\max_{1\leqslant i\leqslant n} \varepsilon_i^2\right] \leqslant \sigma^2(1 + 4\log n).$$

Find a condition on the number M_n of trees such that the finite centered random forest fulfills

$$\lim_{n\to\infty} \mathbb{E}[(\hat{r}_{M_n,n}(X,\Theta) - r(X))^2] = 0.$$

Solution.

Observe that,

$$
\begin{aligned}
\Big(\hat{r}_{M,n}(X,\Theta_1,\ldots,\Theta_m) - r(X)\Big)^2 &= \Big(\hat{r}_{M,n}(X,\Theta_1,\ldots,\Theta_m) - \mathbb{E}_\Theta\left[\hat{r}_n(X,\Theta)\right]\Big)^2 + \Big(\mathbb{E}_\Theta\left[\hat{r}_n(X,\Theta)\right] - r(X)\Big)^2\\
&\quad + 2\Big(\mathbb{E}_\Theta\left[\hat{r}_n(X,\Theta)\right] - r(X)\Big)\Big(\hat{r}_{M,n}(X,\Theta_1,\ldots,\Theta_m) - \mathbb{E}_\Theta\left[\hat{r}_n(X,\Theta)\right]\Big).
\end{aligned}
$$

Note $R(U) = \mathbb{E}[(U - r(X))^2]$. Then, taking the expectation on both sides,

$$R(\hat{r}_{M,n}) = R(\hat{r}_{\infty,n}) + \mathbb{E}\Big[\hat{r}_{M,n}(X,\Theta_1,\ldots,\Theta_m) - \mathbb{E}_\Theta\left[\hat{r}_n(X,\Theta)\right]\Big]^2,$$

by noting that

$$
\begin{aligned}
&\mathbb{E}\Big[\Big(\hat{r}_{M,n}(X,\Theta_1,\ldots,\Theta_m) - \mathbb{E}_\Theta\left[\hat{r}_n(X,\Theta)\right]\Big)\Big(\mathbb{E}_\Theta\left[\hat{r}_n(X,\Theta)\right] - m(X)\Big)\Big]\\
&\quad = \mathbb{E}_{X,\mathcal{D}_n}\left[\Big(\mathbb{E}_\Theta\left[\hat{r}_n(X,\Theta)\right] - m(X)\Big)\mathbb{E}_{\Theta_1,\ldots,\Theta_M}\Big[\hat{r}_{M,n}(X,\Theta_1,\ldots,\Theta_m) - \mathbb{E}_\Theta\Big[\hat{r}_n(X,\Theta)\Big]\Big]\right] = 0.
\end{aligned}
$$

Note that random variables $\hat{r}_n(X,\Theta_1),\ldots,\hat{r}_n(X,\Theta_m)$ are independent and identically distributed conditionnaly on X and $\mathcal{D}_n$. Thus,

$$
\begin{aligned}
\mathbb{E}\left[\hat{r}_{M,n}(X,\Theta_1,\ldots,\Theta_m) - \mathbb{E}_\Theta\left[\hat{r}_n(X,\Theta)\right]\right]^2 &= \mathbb{E}_{X,\mathcal{D}_n}\mathbb{E}_{\Theta_1,\ldots,\Theta_M}\left[\frac{1}{M}\sum_{m=1}^{M}\hat{r}_n(X,\Theta_m) - \mathbb{E}_\Theta\left[\hat{r}_n(X,\Theta)\right]\right]^2,\\
&= \frac{1}{M} \times \mathbb{E}\Big[\mathbb{V}_\Theta\left[\hat{r}_n\left(X,\Theta\right)\right]\Big],
\end{aligned}
$$

Now, note that the tree estimate $\hat{r}_n(X,\Theta)$ can be written as

$$\hat{r}_n(X,\Theta) = \sum_{i=1}^{n} W_{ni}(X,\Theta)Y_i,$$

Therefore,

$$\begin{aligned}
R(\hat{r}_{M,n}) - R(\hat{r}_{\infty,n}) = \frac{1}{M} \times \mathbb{E}\Big[\mathbb{V}_\Theta\left[\hat{r}_n(X,\Theta)\right]\Big] &= \frac{1}{M} \times \mathbb{E}\left[\mathbb{V}_\Theta\left[\sum_{i=1}^{n} W_{ni}(X,\Theta)(r(X_i)+\varepsilon_i)\right]\right] \\
&\leqslant \frac{1}{M} \times \mathbb{E}\left[\mathbb{E}_\Theta\left[\max_{1\leqslant i\leqslant n}(r(X_i)+\varepsilon_i) - \min_{1\leqslant j\leqslant n}(r(X_j)+\varepsilon_j)\right]^2\right] \\
&\leqslant \frac{1}{M} \times \mathbb{E}\Bigg[2\mathbb{E}_\Theta\left[\max_{1\leqslant i\leqslant n} r(X_i) - \min_{1\leqslant j\leqslant n} r(X_j)\right]^2 \\
&\qquad + 2\mathbb{E}_\Theta\left[\max_{1\leqslant i\leqslant n}\varepsilon_i - \min_{1\leqslant j\leqslant n}\varepsilon_j\right]^2\Bigg] \\
&\leqslant \frac{1}{M} \times \left[8\|r\|_\infty^2 + 2\mathbb{E}\left[\max_{1\leqslant i\leqslant n}\varepsilon_i - \min_{1\leqslant j\leqslant n}\varepsilon_j\right]^2\right] \\
&\leqslant \frac{1}{M} \times \left[8\|r\|_\infty^2 + 8\sigma^2\mathbb{E}\left[\max_{1\leqslant i\leqslant n}\frac{\varepsilon_i}{\sigma}\right]^2\right].
\end{aligned}$$

Therefore,

$$R(\hat{r}_{M,n}) - R(\hat{r}_{\infty,n}) \leqslant \frac{8}{M} \times \Big(\|r\|_\infty^2 + \sigma^2(1+4\log n)\Big).$$

Thus the finite random forest is consistent if $M \to \infty$ such that $(\log n)/M \to 0$. $\square$