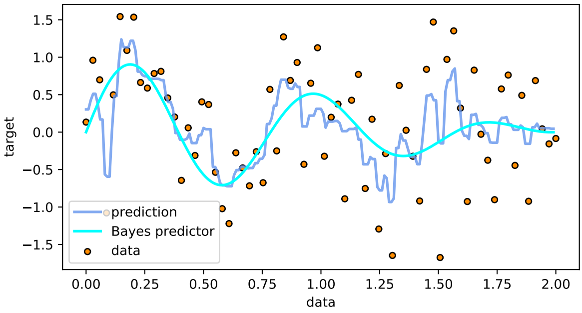
Fondamentaux théoriques du machine learning

Random forest regression
number of estimators: 20
max depth: 5
test error: 8.09E-01
Bayes risk: 7.00E-01



Step 1

We have seen that

$$R(f_n) - R(f_a) \leq 2 \sup_{h \in F} |R(h) - R_n(h)| \tag{1}$$

As a consequence, for all $t \geq 0$:

$$P\Big(R(f_n) - R(f_a) \geq t\Big) \leq P\Big(2 \sup_{h \in F} |R(h) - R_n(h)| \geq t\Big) \tag{2}$$

Step 2

The fact that

$$2 \sup_{h \in F} |R(h) - R_n(h)| \geq t \tag{3}$$

is equivalent to :

$$\cup_{h \in F} \Big(2|R(h) - R_n(h)| \geq t \Big) \tag{4}$$

Step 3

Boole's inequality shows that :

$$P\Big(\cup_{h \in F} \left(2|R(h) - R_n(h)| \geq t\right)\Big) \leq \sum_{h \in F} P\big(2|R(h) - R_n(h)| \geq t\big) \tag{5}$$

Step 4

For each $h \in F$, we need to bound

$$P\big(2|R(h) - R_n(h)| \geq t\big) \tag{6}$$

Step 4

For each $h \in F$, we need to bound

$$P\left(2|R(h) - R_n(h)| \geq t\right) \tag{7}$$

With Hoeffding's inequality we get

$$P\left(2|R(h) - R_n(h)| \geq t\right) \leq 2\exp\left(-\frac{nt^2}{2(b-a)^2}\right) \tag{8}$$

Step 5

With Hoeffding's inequality we get

$$P\big(2|R(h) - R_n(h)| \geq t\big) \leq 2\exp\left(-\frac{nt^2}{2(b-a)^2}\right) \tag{9}$$

Finally, putting everything together :

$$\begin{aligned} P\Big(R(f_n) - R(f_a) \geq t\Big) &\leq \sum_{h\in F} P\big(2|R(h) - R_n(h)| \geq t\big) \\ &\leq \sum_{h\in F} 2\exp\left(-\frac{nt^2}{2(b-a)^2}\right) \\ &= 2|F|\exp\left(-\frac{nt^2}{2(b-a)^2}\right) \end{aligned} \tag{10}$$

Conclusion

$$P\Big(R(f_n) - R(f_a) \geq t\Big) \leq 2|F| \exp\left(-\frac{nt^2}{2(b-a)^2}\right) \tag{11}$$

Conclusion

We write

$$\delta = 2|F| \exp\left(- \frac{nt^2}{2(b-a)^2} \right) \tag{12}$$

We assume that $b - a = 1$. Then, with probability $1 - \delta$, we can compute and show that

$$R(f_n) \leq R(f_a) + 2\sqrt{\frac{\log(|F|) + \log(\frac{2}{\delta})}{2n}} \tag{13}$$