# Projet d'Advanced Machine Learning : AdaBoost

Victor Dheur

27 décembre 2020

## 1 Principe du boosting

## 2 Méthode AdaBoost

## 3 Bornes sur l'erreur de généralisation

On montre maintenant comment obtenir des bornes sur l'erreur de généralisation d'AdaBoost. Nous commençons par obtenir la dimension VC d'AdaBoost, puis nous appliquons l'inégalité VC. Cette approche est tirée du livre [SS14], et des détails supplémentaires ont été ajoutés.

Nous avons vu que la sortie de l'algorithme AdaBoost est une hypothèse composée d'une combinaison linéaire d'hypothèses faibles. AdaBoost se base sur un weak-learner dont l'espace d'hypothèses est dénoté $B$. $T$ hypothèses $h_1, ..., h_T$ sont créées par ce weak-learner. La sortie d'AdaBoost fait partie de cet ensemble d'hypothèses :

$$L(B,T) = \left\{ x \mapsto \text{sign}\left(\sum_{t=1}^{T} w_t h_t(x)\right) \mid \forall t, w_t \in \mathbb{R} \wedge h_t \in B \right\}.$$

Pour un espace d'hypothèses $\mathcal{H}$, nous dénotons $d_{VC}(\mathcal{H})$ sa dimension VC et $m_{\mathcal{H}}$ sa growth function.

**Théorème 1.** *(Dimension VC d'AdaBoost)*
*Nous allons montrer que, lorsque $T \geq 3$ et $d_{VC}(B) \geq 3$ :*

$$d_{VC}(L(B,T)) \leq T(d_{VC}(B)+1)(3\ln(T(d_{VC}(B)+1))+2).$$

*Démonstration.* Dénotons $d = d_{VC}(B)$ et supposons que $T \geq 3$ et $d_{VC}(B) \geq 3$. Soit $C = (x_1, ..., x_m)$ une séquence de points qui est shattered par $L(B,T)$. La création d'un labeling de $C$ par une hypothèse $h \in L(B,T)$ se fait en 2 étapes. D'abord, $T$ hypothèses $h_1, ..., h_T \in B$ sont sélectionnées par le weak-learner. Ensuite, un vecteur $w \in \mathbb{R}^T$ permet de créer la combinaison linéaire $\sum_{t=1}^{T} w_t h_t(x)$ pour un point $x$. On obtient ainsi un labeling $(h(x_1), ..., h(x_m))$ de $C$.

Nous allons utiliser le lemme de Sauer, qui permet de borner supérieurement la growth function $m_{\mathcal{H}}$ d'un espace d'hypothèses $\mathcal{H}$ en utilisant la VC dimension $d_{VC}(\mathcal{H})$ :

$$m_{\mathcal{H}}(m) \leq \left(\frac{em}{d_{VC}(\mathcal{H})}\right)^{d_{VC}(\mathcal{H})}.$$

Par le lemme de Sauer, au plus $\left(\frac{em}{d}\right)^d$ labelings différents de $C$ peuvent être créés à partir de l'espace d'hypothèses $B$. De plus, $T$ hypothèses qui créent ces labelings doivent être choisies, ce qui donne au plus $\left(\frac{em}{d}\right)^{dT}$ labelings différents. En utilisant encore le lemme de Sauer, puisque la dimension VC d'un perceptron (sans biais) dans $\mathbb{R}^T$ est de $T$, la combinaison linéaire entraîne $\left(\frac{em}{T}\right)^T$ labelings différents. Nous avons donc :

$$m_{L(B,T)}(m) \leq \left(\frac{em}{d}\right)^{dT} \left(\frac{em}{T}\right)^T.$$

En utilisant les hypothèses que $T \geq 3$ et $d_{VC}(B) \geq 3$, nous avons :

$$\left(\frac{em}{d}\right)^{dT} \left(\frac{em}{T}\right)^T \leq m^{(d+1)T}.$$

Puisque $C$ est shattered par $L(B,T)$, $m_{L(B,T)}(m) = 2^m$.
Nous avons donc :

$$2^m = m_{L(B,T)}(m) \leq m^{(d+1)T}$$

En passant au log :

$$m \leq \ln(m)\frac{(d+1)T}{\ln(2)}$$

Il est possible de montrer (voir [SS14] p.419 lemme A.1) que

$$\forall a > 0, x \leq a\ln(x) \implies x \leq 2a\ln(a).$$

On déduit une borne sur $m$ qu'on borne encore par une expression plus simple :

$$m \leq \frac{2(d+1)T}{\ln(2)} \ln \frac{(d+1)T}{\ln(2)} \leq (d+1)T(3\ln((d+1)T) + 2).$$

En d'autres termes, le nombre de points $m$ qui peuvent être shattered par $L(B,T)$ est borné supérieurement par une expression qui dépend de $d$ et $T$. Puisque la dimension VC correspond au nombre maximum de points qui peuvent être shattered, l'expression reste vraie lorsque $m = d_{VC}(L(B,T))$ :

$$d_{VC}(L(B,T)) \leq m \leq (d+1)T(3\ln((d+1)T) + 2).$$

□

Il ne reste plus qu'à utiliser l'inégalité VC en bornant la growth function par le lemme de Sauer pour avoir une borne sur l'erreur de généralisation. Nous utilisons l'inégalité VC présentée dans [BBLR03] à la page 192.

En supposant que la loss produit des valeurs bornées dans $[0,1]$, pour toute précision $\epsilon > 0$, on obtient :

$$\begin{aligned}\mathbb{P}\left[\sup_{h \in L(B,T)} (R(h) - R_m(h)) \geq \epsilon\right] &\leq 4m_{L(B,T)}(2m)e^{-m\epsilon^2/8} \\ &\leq 4\left(\frac{2me}{d_{VC}(L(B,T))}\right)^{d_{VC}(L(B,T))} e^{-m\epsilon^2/8}.\end{aligned}$$

Un point important de ce développement est que la dimension VC de l'ensemble des hypothèse produites par AdaBoost augmente linéairement avec la dimension VC de $B$ et avec $T$, en ignorant les facteurs constants et logarithmiques.

# Références

[BBLR03] Olivier Bousquet, Stéphane Boucheron, Gábor Lugosi, and Gunnar Rätsch. Introduction to statistical learning theory. January 2003.

[SS14] Shai Shalev-Shwartz. *Understanding Machine Learning (From Theory to Algorithms)*. Cambridge University Press, 1 edition, May 2014.