Chapter 11: Boosting and PAC

Elvis Cui

March 17, 2020

Contents

1 Generic Adaptive Boosting

1.1 The PAC Framework

In this subsection, we define some important concepts in statistics and learning theory. For definitions of hypothesis and concept class, see Mohri, 2018, pp9.

Definition 1 (Risk) *(A.K.A. generalization error.) Given a hypothesis (or an estimator of c) $h \in \mathcal{H}$, a target concept $c \in \mathcal{C}$, the risk or generalization error of h is define by*

$$R(h) = \mathbb{P}_{x\sim\mathcal{D}}[h(x) \neq c(x)] = \mathbb{E}_{x\sim\mathcal{D}}[1_{h(x)\neq c(x)}]$$

Given a sample $S = (x_1, \cdots, x_m)$, the empirical error or empirical risk of h is defined by

$$\widehat{R}_S(h) = \frac{1}{m}\sum_{i=1}^{m} 1_{h(x_i)\neq c(x_i)}$$

where 1_ω is the indicator function.

Definition 2 (PAC-Learning (Mohri, 2018, pp11)) *A concept class $\mathcal{C}$ is said to be PAC-learnable if there exists an algorithm $\mathcal{A}$ and a polynomial function $p(\cdot,\cdot,\cdot,\cdot)$ s.t. for any $\epsilon > 0$ and $\delta > 0$, for all distributions $\mathcal{D}$ on $\mathcal{X}$ and for any target concept $c \in \mathcal{C}$, the following holds for any sample size $m > p(q/\epsilon, 1\delta, n, size(c))$:*

$$\mathbb{P}_{S\sim\mathcal{D}^m}(R(h_S) \leq \epsilon) \geq 1 - \delta$$

where h_S is the hypothesis returned by $\mathcal{A}$ and S.

1.2 Weak Learnability

Definition 3 (Weak learning) *A concept class $\mathcal{C}$ is said to be weakly (PAC-)learnable is there exists an algorithm $\mathcal{A}$, $\gamma > 0$, and a polynomial function $p(\cdot,\cdot,\cdot)$ s.t. for any $\delta > 0$, for all distributions $\mathcal{D}$ on $\mathcal{X}$ and for any target $c \in \mathcal{C}$, the following holds for any sample size $m \geq p(\frac{1}{\delta}, n, size(c))$:*

$$\mathbb{P}_{S\sim\mathcal{D}^m}(R(h_S) \leq 0.5 - \gamma) \geq 1 - \delta,$$

where h_S is the hypothesis returned by algorithm $\mathcal{A}$ when trained on sample S. When such $\mathcal{A}$ exists, it is called a weak learner for $\mathcal{C}$. The hypotheses returned by a weak learning algorithm are called base classifiers.

1.3 Adaptive Boosting

Given a class of base classifiers $\mathcal{H}$, a training sample $S = ((x_1, y_1), \cdots, (x_m, y_m))$ of size m, let's do

Algorithm 1: Generic Adaptive Boosting

Result: $f(x) = sgn\{\sum_{i=1}^{T} h_i(x)\}$
Initializing coefficients: $\mathcal{D}_0(i) = \frac{1}{m}$;
while *t=1, t++, $t \leq T$* **do**
  $h_t \leftarrow$ base classifier in $\mathcal{H}$ that approximately minimize

$$\epsilon_t := \frac{\sum_{i=1}^{m} \mathcal{D}_{t-1}(i) 1(h_t(x_i) \neq y_i)}{\sum_{i=1}^{m} \mathcal{D}_{t-1}(i)} = \mathbb{P}_n\left[1(h_t(x) \neq y)\right]$$

  Update $\alpha_t \leftarrow \frac{1}{2}\log\frac{1-\epsilon_t}{\epsilon_t}$
  For $i = 1 \cdots, m$, update $\mathcal{D}_t(i) \leftarrow \mathcal{D}_{t-1}(i)e^{-\alpha_t y_i h_t(x_i)}$
  Normalize $\mathcal{D}_t(i)$ as $\frac{\mathcal{D}_t(i)}{Z_t}$, $Z_t = \sum_{i=1}^{m} \mathcal{D}_{t-1}(i)e^{-\alpha_t y_i h_t(x_i)}$.
end
Output $g(x) \leftarrow \sum_{t=1}^{T} \alpha_t h_t(x)$ and $f(x) = sgn(g(x))$.

2 AdaBoost as Exponential Surrogate Loss

An alternative point of view of Adaboost after Freud and Schapire's original paper (1996) was given by Friedman at Stanford on Annals of Statistics (1999). Later, along with Hastie and Tibshirani, Friedman connected boosting method and the classical logistic regression under the framework of generalized additive models (2000).

2.1 Exponential Surrogate Loss (Friedman et al, 2000)

The exponential surrogate loss for binary classification is defined as

$$L(y, f(x)) = e^{-yf(x)}, y, f(x) \in \{-1, +1\}$$

The risk is

$$R(f) = \mathbb{E}_{X,y} L(y, f(X))$$

By the statistical decision theory, to derive a Bayes estimator, it is enough to consider the conditional case:

$$\begin{aligned} R(f|X = x) &= \mathbb{E}_y(L(y, f(X))|X = x) \\ &= e^{-f(x)} p(y = 1|X = x) + e^{f(x)} p(y = -1|X = x) \\ &= e^{-f(x)} p_1 + e^{f(x)} p_{-1} \end{aligned}$$

where we used $p_i = p(y = i|X = x)$ for short. Taking derivative w.r.t. f (see section 6 for details on derivatives of functions), we derive the optimal f is:

$$\widehat{f}(x) = \frac{1}{2} \log \frac{p_{-1}}{p_1} \tag{1}$$

If we use empirical error to replace p_{-1} and p_1, then this is nothing but α in the generic Adaboost algorithm.

2.2 Coordinate Descent

Another perspective of Adaboost relates to coordinate descent in optimization. Note that ignoring normalizing constant, at k^{th} iteration, the loss function of Adaboost is

$$\begin{aligned} L_k(y, f(x)) &= \sum_{i=1}^{m} e^{-y_i \sum_{t=1}^{k} \alpha_t h_t(x_i)} \\ &= \sum_{i=1}^{m} e^{-y_i f_{k-1}(x_i) - y_i \alpha_k h_k(x_i)} \\ &= \sum_{i=1}^{m} \mathcal{D}_{k-1}(i) e^{-y_i \alpha_k h_k(x_i)} \end{aligned}$$

Where we have defined $f_{k-1} = \sum_{t=1}^{k-1} h_t$ and $\mathcal{D}_{k-1}(i) = e^{-y_i f_{k-1}(x_i)}$. But the last term can be rewritten as

$$\sum_{\{i: y_i h_k(x_i) = 1\}} \mathcal{D}_{k-1}(i) e^{-y_i \alpha_k h_k(x_i)} + \sum_{\{i: y_i h_k(x_i) = -1\}} \mathcal{D}_{k-1}(i) e^{-y_i \alpha_k h_k(x_i)}$$

which is equivalent to (substitute $y_i h_k(x_i) = \pm 1$)

$$\sum_{i=1}^{m} \mathcal{D}_{k-1}(i) e^{-\alpha_k} + \sum_{\{i: y_i h_k(x_i) = -1\}} \mathcal{D}_{k-1}(i)(e^{\alpha_k} - e^{-\alpha_k})$$

Since e^{α_k} is independent of i^{th} data point, so the first term is $e^{-\alpha_k} \sum_{i=1}^{m} \mathcal{D}_{k-1}(i)$. Taking derivative w.r.t. α_k and setting it to 0, we get

$$\alpha_k = \frac{1}{2} \log \frac{\sum_{\{i: y_i h_k(x_i) = 1\}} \mathcal{D}_{k-1}(i)}{\sum_{\{i: y_i h_k(x_i) = -1\}} \mathcal{D}_{k-1}(i)}$$

which is exactly the empirical version of (1).

3 Prerequisites for Theoretical Analysis

To derive theoretical properties of Adaboost and other alternatives, we need to equip with more advanced tools in probability known as concentration inequalities. Since the pioneering paper written by Hoeffding (1963), concentration phenomenon in high dimensions has been a popular and deep field of both probability theory and statistics.

3.1 Concentration Inequalities

Theorem 1 (Azuma-Hoeffding inequality (Dabrowska, 2020, pp1051)) *Let $M_n = \sum_{i=1}^n X_n$ be a mean 0 martingale s.t. $a_n \leq X_n \leq b_n$ a.s., then*

$$\mathbb{P}(M_n > t) \leq e^{-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2}}$$

Theorem 2 (Bounded difference (Wainwright, 2019, pp39)) *Suppose f has bounded difference with parameters $(L_1, \cdots, L_n)$ and $X = (X_1, \cdots, X_n)$ has independent coordinates. Then*

$$\mathbb{P}(|f(X) - \mathbb{E}(f(X))| \geq t) \leq 2e^{-\frac{2t^2}{\sum_{k=1}^n L_k^2}}$$

Theorem 3 (Binomial tails (Mohri et al, 2018, pp440)) *Let $X_i, i = 1, \cdots, m \sim_{iid} Ber(p)$, denote $\bar{X}$ as $\frac{1}{m}\sum_{i=1}^m X_i$. Then the following holds for the upper bounds of $\mathbb{P}(\bar{X} - p > \epsilon)$:*

$$e^{-2m\epsilon^2} \text{ [Hoeffding]}$$
$$e^{-\frac{m\epsilon^2}{2\sigma^2 + \frac{2\epsilon}{3}}} \text{ [Bernstein]}$$
$$e^{-m\sigma^2\theta(x)} \text{ [Bennett]}$$
$$e^{-mD(p+\epsilon\|p)} \text{ [Sanov]}$$

where $\sigma^2 = p(1-p), \theta(x) = (1+x)\log(1+x) - x$.

Theorem 4 (Maximal inequality (Mohri et al, 2018, pp445)) *Let independent variables $Y_{ij} \in [-r_i, +r_i]$ a.s. with mean 0, $j = 1, \cdots, p$ and $X_i = \sum_{j=1}^p Y_{ij}$. Then we have*

$$\mathbb{E}(\max_{i=1,\cdots,m} X_i) \leq \sqrt{2(\sum_{i=1}^p r_i^2)\log m}$$

Theorem 5 (Gaussian concentration (Amini, 2019, slide47)) *Let $X \sim \mathcal{N}(0, I_m)$ be a standard Gaussian random vector and assume that $f : \mathbb{R}^n \to \mathbb{R}$ is L-lipschitz w.r.t. Euclidean norm. Then,*

$$\mathbb{P}(|f(X) - \mathbb{E}f(X)| \geq t) \leq 2\exp\{-\frac{t^2}{2L^2}\}, t \geq 0.$$

3.2 Growth Function, VC Dimension, Rademacher Complexity and Metric Entropy

Definition 4 (Growth function) *The growth function $\Pi_{\mathcal{H}} : \mathbb{N} \to \mathbb{N}$ for a hypotheses set $\mathcal{H}$ is defined by:*

$$\forall m \in \mathbb{N}, \Pi_{\mathcal{H}}(m) = \max_{x \subset \mathcal{X}} |\{(h(x_1), \cdots, h(x_m)) : h \in \mathcal{H}\}|$$

That is, $\Pi_{\mathcal{H}}(m)$ is the maximal number of distinct ways in which m points can be classified using hypotheses in $\mathcal{H}$.

Definition 5 (Shatter coefficient (Wasserman, 2007, pp18)) *Let $\mathcal{A}$ be a family of sets and given a finite set $R = \{x_1, \cdots, x_m\}$, let*

$$\mathcal{N}_{\mathcal{A}}(R) = |\{R \cap A : A \in \mathcal{A}\}|$$

The shatter coefficient is defined to be

$$s(\mathcal{A}, m) = \max_{R \in \mathcal{F}_n} N_{\mathcal{A}}(R)$$

Here $\mathcal{F}_n$ consists of any set with cardinality m.

Definition 6 (VC dimension) *The VC dimension of a hypothesis set $\mathcal{H}$ is the size of the largest set that can be shattered by $\mathcal{H}$:*

$$VCdim(\mathcal{H}) = \max\{m : \Pi_{\mathcal{H}}(m) = 2^m\} = \arg\max_m(s(\mathcal{H}, m) = 2^m)$$

Definition 7 (Rademacher complexity) *Let $\mathcal{D}$ denote the distribution according to which samples are drawn. For any $m \in \mathbb{Z}_+$, the Rademacher complexity of $\mathcal{G}$, a class of functions, is the expectation of the empirical Rademacher complexity (to be defined) over all samples of size m drawn from $\mathcal{D}$:*

$$\Re_m(\mathcal{G}) = \mathbb{E}_{S\sim\mathcal{D}^m}[\widehat{\Re}_S(\mathcal{G})]$$
$$\widehat{\Re}_S(\mathcal{G}) = \mathbb{E}_\sigma\left[\sup_{g\in\mathcal{G}}\frac{1}{m}\sum_{i=1}^m \sigma_i g(z_i)\right]$$

Where $(z_1, \cdots, z_m)$ is the sample and $\sigma_1, \cdots, \sigma_m$ are iid Rademacher variables.

3.2.1 Binary Classification

Theorem 6 (Radmacher complexity bounds for binary classification (Mohri et al, pp33)) *Let $\mathcal{H}$ be a family of functions from $\mathcal{X}$ to $\{-1, +1\}$ and let $\mathcal{D}$ be any distribution (i.e. probability measure) over $\mathcal{X}$. Then, for any $\delta > 0$, w.p. larger than $1 - \delta$ over a sample S of size m drawn according to $\mathcal{D}$, each of the following holds for any $h \in \mathcal{H}$:*

$$R(h) \le \widehat{R}_S(h) + \Re_m(\mathcal{H}) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{2}$$

$$R(h) \le \widehat{R}_S(h) + \widehat{\Re}_m(\mathcal{H}) + 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}} \tag{3}$$

Lemma 1 (Bounding Radmacher complexity via shatter coefficient (Mohri et al, 2018, pp35)) *Let $\mathcal{G}$ be a family of functions from $\mathcal{X}$ to $\{-1, +1\}$. Then we have:*

$$\Re_m(\mathcal{G}) \le \sqrt{\frac{2\log\Pi_{\mathcal{G}}(m)}{m}} = \sqrt{\frac{2\log s(\mathcal{G}, m)}{m}} \tag{4}$$

Proof 1 *By maximal inequality, for fixed $S = (X_1, \cdots, X_m)$, we have*

$$\mathbb{E}_\sigma \sup_{g\in\mathcal{G}}\sum_{i=1}^m \sigma_i g(x_i) \le \sqrt{2m\log s(\mathcal{G}, m)}$$

Since the vector $g(x) = (g(x_1), \cdots, g(x_m))^T$ can take at most $\Pi_{\mathcal{G}}(m) = s(\mathcal{G}, m)$ distinct values for any $g \in \mathcal{G}$. Then what left is trivial.

Using the above result we immediately get another upper bound of generalization error. That is, given $\mathcal{G}$, a family of functions from $\mathcal{X}$ to $\{-1, +1\}$, for any $\delta > 0$, w.p. larger than $1 - \delta$, we have for any $g \in \mathcal{G}$,

$$R(g) \le \widehat{R}_S(g) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} + \sqrt{\frac{2\log\Pi_{\mathcal{G}}(m)}{m}}$$

Equivalently, we have the following theorem originally due to Vapnik and Cervonenkis:

Theorem 7 (VC theorem (1971)) *Under the above assumptions, we have the following bound:*

$$\mathbb{P}(|R(g) - \widehat{R}_S(g)| \ge \epsilon) \le 2\Pi^2_{\mathcal{G}}(m)\exp\{-m\epsilon^2\}$$

Note that in the original theorem, the bound is (Wasserman, 2007, pp18)

$$\mathbb{P}(|R(g) - \widehat{R}_S(g)| \ge \epsilon) \le 8\Pi_{\mathcal{G}}(m)\exp\{-m\epsilon^2/32\}$$

Ignoring constants, these two bounds are the same.

3.3 A Detour on Glivenko-Cantelli Class

4 More on VC Dimension

Theorem 8 (VC dimension of finite dimensional vector spaces (Mohri et al, 2018, exercise 3.19)) *Let F be a finite-dimensional vector space of real functions on $\mathcal{R}^d$, $\dim(F) = r < \infty$. Let $\mathcal{H}$ be the family of hypotheses*

$$\mathcal{H} = \{\{x : f(x) \geq 0\} : f \in F\}$$

Then we have $VC\dim(\mathcal{H}) \leq r$.

Proof 2 *Given a function $f \in F$ and the points $(x_1, \cdots, x_{r+1})$, $x_1 \in \mathcal{R}^d$, define*

$$L(f) = (f(x_1), \cdots, f(x_{r+1}))$$

Then obviously this is a linear mapping:

$$L(af + bg) = aL(f) + bL(g), a, b \in \mathcal{R}, f, g \in F$$

Since $\dim(F) = r < r + 1$, we have

$$\{L(f) : f \in F\} \subset \mathcal{R}^r$$

Thus, for any f, there exists a linear combination s.t.

$$\sum_{i=1}^{r+1} \alpha_i f(x_i) = 0$$

and not all $\alpha_i's$ are zero. Now if all $\alpha_i \geq 0$, then we are not able to pick up the case $f(x_i) > 0, i = 1, \cdots, r + 1$. So WLOG, we assume there exists at least one $\alpha_i < 0$. Then rewrite the above equation:

$$\sum_{\{i:\alpha_i \geq 0\}} \alpha_i f(x_i) = \sum_{\{i:\alpha_i < 0\}} -\alpha_i f(x_i)$$

However, we are not able to pick up the case (thus, a contradiction)

$$f(x_i) \geq 0 \text{ if } i \in \{i : \alpha_i \geq 0\}$$

$$f(x_i) < 0 \text{ if } i \in \{i : \alpha_i < 0\}$$

In general, we have the following theorem on connections between growth function and VC dimension:

Theorem 9 (Sauer's lemma; Vapnik-Chervonenkis theorem) *(Wainwright, 2019, pp123; Mohri, 2018, pp41) Consider a hypothesis class $\mathcal{H}$ with $VC\dim(\mathcal{H}) = d < \infty$. Then for any $m \in \mathbb{N}$, the following holds*

$$\Pi_m(\mathcal{H}) \leq \sum_{i=0}^{d} \binom{m}{i} \leq (\frac{em}{d})^d \leq (m + 1)^d$$

Where the last equality holds if $d > e$.

Using the previous theorem and inequality under (4) we get

Lemma 2 (VC-dimension generalization bounds (Mohri, 2018, pp42)) *Let $\mathcal{H}$ be a family of functions taking values in $\{-1, +1\}$ with VC-dimension d. Then, for any $\delta > 0$, w.p. larger than $1 - \delta$, the following holds for any $h \in \mathcal{H}$:*

$$R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{2d \log \frac{em}{d}}{m}} + \sqrt{\frac{\log \frac{1}{\delta}}{2m}} = \widehat{R}_S(h) + O\left(\sqrt{\frac{\log(m/d)}{m/d}}\right) \tag{5}$$

In fact, according to Mohri, these VC-dim based bounds can be derived directly without using Rademacher complexity as an intermediate ingredient:

$$R(h) \leq \widehat{R}_S(h) + \sqrt{\frac{8d \log \frac{2em}{d} + 8 \log \frac{4}{\delta}}{m}}$$

5 Theoretical Results on AdaBoosting

5.1 Bounding Empirical Error via Hoeffding's Inequality

The following theorem shows why we could boost weak learners into a strong one.

Theorem 10 *Suppose f is the function (classifier) returned by Adaboost, i.e. $f(x) = sgn(\sum_{t=1}^{T} \alpha_t h_t(x))$. Then the empirical error of f satisfies the following exponential bound:*

$$\widehat{R}_S(f) \leq \exp\left[-2\sum_{t=1}^{T}(\frac{1}{2} - \epsilon_t)^2\right]$$

where $\epsilon_t = \sum_{i=1}^{m} \mathcal{D}_{t-1}(i) 1_{\{h_t(x_i) \neq y_i\}} / \sum_{t=1}^{T} \mathcal{D}_{t-1}(i)$. Furthermore, if $\epsilon_t \leq \frac{1}{2} - \gamma$ ($\gamma > 0$ by weak learnability), then

$$\widehat{R}_S(f) \leq \exp\{-2\gamma^2 T\}$$

So the empirical error decreases exponentially as iteration increases.

Proof 3 *First note that $g(x) = \sum_{t=1}^{T} \alpha_t h_t(x)$ and we have*

$$\widehat{R}_S(f) = \frac{1}{m}\sum_{i=1}^{m} 1_{y_i f(x_i) \neq 1} \tag{6}$$

$$= \frac{1}{m}\sum_{i=1}^{m} 1_{y_i g(x_i) \leq 0} \tag{7}$$

$$\leq \frac{1}{m}\sum_{i=1}^{m} e^{-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)} \tag{8}$$

where I have used the inequality $1_{x \leq 0} \leq e^{-x}$. Next, denote $Z_t = \sum_{i=1}^{m} \mathcal{D}_{t-1}(i) \exp(-y_i \alpha_t h_t(x_i))$, that is, the normalization constant at t^{th} iteration. We have the following iterative relation:

$$\mathcal{D}_T(i) = \frac{\mathcal{D}_{T-1}(i)\exp(-y_T \alpha_T h_T(x_i))}{Z_T} \tag{9}$$

$$= \frac{\mathcal{D}_{T-1}(i)\exp(-(\alpha_T y_T h_T(x_i) + \alpha_{T-1} y_{T-1} h_{T-1}(x_i)))}{Z_T Z_{T-1}} \tag{10}$$

$$= \cdots \tag{11}$$

$$= \frac{\mathcal{D}_0(i)\exp(-y_i g(x_i))}{\prod_{t=1}^{T} Z_t} \tag{12}$$

$$= \frac{\exp(-y_i g(x_i))}{m\prod_{t=1}^{T} Z_i} \tag{13}$$

The last step comes from the initialization $\mathcal{D}_0(i) = \frac{1}{m}, i = 1, \cdots, m$. Since $\sum_{i=1}^{m} \mathcal{D}_t(i) = 1$ holds for any t, we have

$$\widehat{R}_S(f) \leq \frac{1}{m}\sum_{i=1}^{m} e^{-y_i \sum_{t=1}^{T} \alpha_t h_t(x_i)} \tag{14}$$

$$= \frac{1}{m}\sum_{i=1}^{m} e^{-y_i g(x_i)} \tag{15}$$

$$= \prod_{t=1}^{T} Z_t \tag{16}$$

To bound the product of Z_t, note that

$$Z_t = \sum_{i=1}^{m} \mathcal{D}_{t-1}(i) \exp(-y_i \alpha_t h_t(x_i)) \tag{17}$$

$$= \sum_{\{i:y_i h_t = 1\}} \mathcal{D}_{t-1}(i) e^{-\alpha_t} + \sum_{\{i:y_i h_t = -1\}} \mathcal{D}_{t-1}(i) e^{\alpha_t} \tag{18}$$

$$= (1-\epsilon_t) e^{-\alpha_t} + \epsilon_t e^{-\alpha_t} \tag{19}$$

$$= 2\sqrt{\epsilon_t(1-\epsilon_t)} \tag{20}$$

The last step comes from the relation $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$. *Last but not least, we have to use the fact:*

$$2\sqrt{\epsilon_t(1-\epsilon_t)} = \sqrt{1 - 4(\frac{1}{2} - \epsilon_t)^2}$$

as well as the numerical inequality which holds for all $x \leq 1$*:*

$$\log(1-x) \leq 1 - x \leq e^{-x}$$

Therefore, we get

$$\widehat{R}_S(f) \leq \prod_{t=1}^{T} Z_t \tag{21}$$

$$= \prod_{t=1}^{T} \sqrt{1 - 4(\frac{1}{2} - \epsilon_t)^2} \tag{22}$$

$$\leq \exp(-2(\frac{1}{2} - \epsilon_t)) \tag{23}$$

$$\leq \exp\{-2\gamma^2 T\} \tag{24}$$

5.2 VC Dimension Based Analysis

Define the following function class:

$$\mathcal{F}_T = \{sgn(\sum_{t=1}^{T} \alpha_t h_t) : \alpha_t \geq 0, h_t \in \mathcal{H}, t = 1, \cdots, T\}$$

Given the VC-dimension of the class $\mathcal{H}$, i.e., $VCdim(\mathcal{H}) = d$, then the VC-dimension of $\mathcal{F}_T$ can be bounded by:

$$\text{VCdim}(\mathcal{F}_T) \leq 2(d+1)(T+1)\log_2((T+1)e)$$

The upper bound grows as $O(dT \log T)$ which implies that Adaboost may overfit as iteration grows. However, it is very rare in practice.

5.3 Rademacher Complexity Based Analysis

If we renormalize the coefficients for h_t, that is:

$$\bar{\alpha}_t = \frac{\alpha_t}{\sum_{t=1}^{T} \alpha_t} = \frac{\alpha_t}{\|\boldsymbol{\alpha}\|_1}$$

Then we find that

$$\mathcal{F}_T = \text{conv}(\mathcal{H})$$

where $\text{conv}(\cdot)$ denotes the convex hull of a given set. In terms of empirical Rademacher complexity, we have the following result which is useful in margin-based analysis for general machine learning algorithms (not just Boosting):

Lemma 3 *Let $\mathcal{H}$ be a set of functions mapping from $\mathcal{X}$ to $\mathbb{R}$. Then, for any sample $S = (X_1, \cdots, X_m)$, we have*

$$\widehat{\Re}_S(conv(\mathcal{H})) = \widehat{\Re}_S(\mathcal{H})$$

5.4 From Large Margin to Weak Learnability

In contrast to the empirical exponential loss bound, we have the following theorem:

Theorem 11 *If for any sample $S = (X_1, \cdots, X_m)$ and the corresponding labels $y_1, \cdots, y_m$, there is a*

$$g \in \mathcal{F}_T = conv(\mathcal{H}) \text{ s.t.}$$

$$y_i g(x_i) \geq \gamma, i = 1, \cdots, m$$

then for all probability distributions $\mathcal{D}$ on $\{1, \cdots, m\}$, there is a $f \in \mathcal{H}$ with

$$\sum_{i=1}^{m} \mathcal{D}(i) 1_{y_1 \neq f(x_i)} \leq \frac{1-\gamma}{2}$$

6 Gradient Boosting

6.1 A Detour on Generalized Additive Models

7 Weak Classifiers

7.1 Tree-based Models

7.2 Other Choices of Weak Learners

8 Simulation Study of Boosting

9 Real Dataset Analysis

10 Reference List