CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 11**

AdaBoost and large margin classifiers

Lecturer: Peter Bartlett *Scribe: Matt Johnson*

1 Review

Algorithm 1 AdaBoost

```
1: D_1(i) ⇐ 1/n, ∀i ∈ {1, ..., n}
2: F_0(x) ⇐ 0
3: for t = 1, ..., T do
4:    choose f_t ∈ G to approximately minimize Σ_{i=1}^n D_t(i) 1[f_t(x_i) ≠ y_i]
5:    α_t ⇐ (1/2) ln((1 − ε_t)/ε_t)
6:    F_t ⇐ F_{t−1} + α_t f_t
7:    D_{t+1}(i) ⇐ (D_t(i)/Z_t) · { e^{α_t}   if f_t(x_i) ≠ y_i
                                   { e^{−α_t}  otherwise
8: end for
```

Note that the Z_t term on line 1 can be thought of as simply a normalizer to ensure that $D_t(i)$ remains a distribution. We will see later in this lecture that $Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$.

2 AdaBoost Analysis

2.1 Performance Bound

The following theorem shows that, if the ϵ_ts are significantly below 1/2, then we can get the proportion of training data misclassified arbitrarily small. The proof actually shows that we can view AdaBoost as an algorithm that greedily minimizes $\hat{\mathbb{E}}e^{-Yf(X)}$.

Theorem 2.1.

$$\hat{P}\left(YF_T(x) \le 0\right) = \frac{1}{n}\left|\{i : y_i F_T(x_i) \le 0\}\right| \quad (1)$$

$$\le \prod_{t=1}^{T} 2\sqrt{\epsilon_t(1-\epsilon_t)} \quad (2)$$

Furthermore, if we know that ϵ_t is slightly less than $\frac{1}{2}$, say $\epsilon_t \le \frac{1}{2} - \gamma\, \forall t$, the product above is no more than $(1-4\gamma^2)^{\frac{T}{2}}$.

Proof. Instead of the event $YF_T(X) \le 0$, look at the equivalent event $\exp(-YF_T(X)) \ge 1$. So, plugging in for F_T, we have

$$\hat{P}\left(YF_T(X) \leq 0\right) \leq \hat{\mathbb{E}}\left[\exp(-YF_T(X))\right] \tag{3}$$
$$= \frac{1}{n}\sum_{i=1}^{n}\exp(-y_i\sum_{t=1}^{T}\alpha_t f_t(x_i)) \tag{4}$$
$$= \frac{1}{n}\sum_{i}\prod_{t}\exp(-y_i\alpha_t f_t(x_i)) \tag{5}$$

We also know that, since $y_i, f(x_i) \in \{\pm 1\}$, their product is also in $\{\pm 1\}$. Note that the exponentiation in the above expression is in the D_{t+1} expression of the algorithm, so we have

$$= \frac{1}{n}\sum_{i}\prod_{t}\frac{D_{t+1}(i)}{D_t(i)}Z_t \tag{6}$$
$$= \frac{1}{n}\sum_{i}\left(\prod_{t}Z_t\right)\frac{D_{T+1}}{D_1(i)} \tag{7}$$
$$= \prod_{t}Z_t \tag{8}$$

Where we have the final equality because $D_1(i) = 1/n$ and D_{T+1} is a distribution, so it sums over i to one.

If we choose α_t to minimize

$$Z_t = \sum_{i:y_i=f_t(x_i)} D_t(i)e^{-\alpha_t} + \sum_{i:y_i\neq f_t(x_i)} e^{\alpha_t} \tag{9}$$
$$= (1-\epsilon_t)e^{-\alpha_t} + \epsilon_t e^{\alpha_t} \tag{10}$$

We can differentiate w.r.t. α_t and set to zero to solve the optimization to get

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

Which gives

$$Z_t = (1-\epsilon_t)\sqrt{\frac{\epsilon_t}{1-\epsilon_t}} + \epsilon_t\sqrt{\frac{1-\epsilon_t}{\epsilon_t}} \tag{11}$$
$$= 2\sqrt{\epsilon_t(1-\epsilon_t)} \tag{12}$$

We can plug in to (8) to get the desired result. □

We can extend the above theorem to include a margin as well.

Theorem 2.2. If we define $\overline{F} = \frac{F_T}{\sum_{t=1}^{T}\alpha_t} = \frac{\sum_t \alpha_t f_t}{\sum_t \alpha_t} \in co(\mathcal{G})$ (like ℓ_1 normalization) then

$$\hat{P}\left(\overline{F}(X) \leq \gamma\right) \leq \prod_t 2\sqrt{\epsilon_t^{1-\gamma}(1-\epsilon_t)^{1+\gamma}}$$

and if $\epsilon_t \leq \frac{1}{2} - 2\gamma\,\forall t$, then this decreases exponentially fast.

We can think of the first theorem (in the previous subsection) as saying: for all D_t, there exists $f_t \in \mathcal{G}$ with weighted empirical risk less than $1/2 - \gamma$, then $\exists \overline{F} \in co(\mathcal{G})$ with $\hat{P}\left(Y\overline{F}(X) \leq 0\right)$. The second theorem replaces the zero in the empirical probability with $\gamma/2$.

The converse result has a similar flavor: if $\exists \overline{F} \in co(\mathcal{G})$ margin better than γ, then we have $\epsilon_t \leq 1/2 - \gamma$.

Below we examine the converse:

Theorem 2.3. If, for $(x_1, y_1), \ldots, (x_n, y_n)$, $\exists F \in co(\mathcal{G})$ with $y_i F(x_i) > \gamma\,\forall i$, then for all probability distributions D on $\{1, \ldots, n\}, \exists f \in \mathcal{G}$ such that

$$\sum D(i) 1\left[y_i \neq f(x_i)\right] \leq \frac{1-\gamma}{2}$$

Proof. We proceed with the probabilistic method:

Suppose $F = \sum_t \alpha_t f_t$ with α_t as convex coefficients. Choose f randomly according to distribution given by $P(f = f_t) = \alpha_t$. Then

$$0 \leq \mathbb{E}\left[\sum_i D(i) 1[y_i = f(x_i)]\right] \tag{13}$$

$$= \sum_t \alpha_t \sum_i D(i) 1[y_i \neq f_t(x_i)] \tag{14}$$

$$= \sum_i D(i) \sum_t \alpha_t \frac{1 - y_i f_t(x_i)}{2} \tag{15}$$

$$= \frac{1}{2}\left(1 - \sum_i D(i) \sum_t \alpha_t f_t(x_i)\right) \qquad \leq \frac{1}{2}(1-\gamma) \tag{16}$$

□

2.2 Another interpretation: gradient descent

From last time, we know $\hat{\mathbb{E}}\exp(-YF_T(X)) = \frac{1}{n}\sum_i \frac{D_{T+1}(i)}{D_1(i)} \prod_t Z_t$. Recall also that $\frac{1}{n}\exp(-y_i F_{T-1}(x_i)) = D_T(i) \prod_{t=1}^{T-1} Z_t$.

<u>Observation:</u> Choosing f_t to minimize $\epsilon_t = \sum_{i=1}^n D_t(i) 1[y_i \neq f_t(x_i)]$ and setting $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$ is equivalent to choosing α_t, f_t to minimize

$$\hat{\mathbb{E}}\exp(-YF_t(X)) = \frac{1}{n}\sum_{i=1}^{n}\exp(-y_i(F_{t-1}(x_i)+\alpha_t f_t(x_i))) \tag{17}$$

$$\hat{\mathbb{E}}\exp(-YF_t(X)) = \frac{1}{n}\sum_{i=1}^{n}\left[(e^{\alpha_t}-e^{-\alpha_t})1[y_i \neq f_t(x_i)+e^{\alpha_t}]\right]e^{-y_iF_{t-1}(x_i)} \tag{18}$$

$$= (e^{\alpha_t}-e^{-\alpha_t})\prod_{s=1}^{t-1}Z_s\sum_{i=1}^{n}D_t(i)1[y_i \neq f_t(x_i)] + \frac{e^{-\alpha_t}}{n}\sum_{i=1}^{n}e^{-y_iF_{t-1}(x_i)} \tag{19}$$

Where the last equality holds from noting that $\frac{1}{n}e^{-y_iF_{t-1}(x_i)}$ is the weighting term recalled above. We also see that $\forall\alpha_t$, the best choice of f_t minimizes the first summation term above.

Given f_t, we can take a partial derivative with respect to α_t and set it equal to zero to find

$$\sum_{i:y_i\neq f_t(x_i)}\left(\frac{1}{n}e^{-y_iF_{t-1}(x_i)}\right)e^{\alpha_t} - \sum_{i:y_i=f_t(x_i)}\left(\frac{1}{n}e^{-y_iF_{t-1}(x_i)}\right)e^{-\alpha_t} = 0 \tag{20}$$

$$\left(\epsilon_t e^{\alpha_t}-(1-\epsilon_t)e^{-\alpha_t}\right)\prod_{s=1}^{t-1}Z_s = 0 \tag{21}$$

Which implies $\alpha_t = \frac{1}{2}\ln\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$

So this is like a coordinate descent along the α_t with the objective of $\min \frac{1}{n}\sum e^{-y_i\sum_t \alpha_t f_t(x_i)}$.

3 An alternative formulation

We can create a more general interpretation with other cost functions than the exponential:

$$\min_F J(F) = \hat{\mathbb{E}}\phi(YF(X)) = \hat{\mathbb{E}}\left[\phi(Y(F_{t-1}(X)+\alpha_t f_t(X)))\right]$$

Gradient descent would be to choose a direction $v = (\alpha_t f_t(x_i), \ldots, \alpha_t f_t(x_n))$ to minimize $v'\nabla_z J_n(F_{t-1}+z)$, i.e. choose a direction from restricted options.

$$v \text{ minimizes } \sum v_i y_i \phi'(y_iF_{t-1}(x_i)) \tag{22}$$

$$\Leftrightarrow \min\sum(-v_iy_i)(-\phi'(y_iF_{t-1}(x_i))) \tag{23}$$

$$\Leftrightarrow \min 1[v_i \neq y_i]D_t(i) \tag{24}$$

With $D_t(i) = \frac{-\phi'(y_iF_{t-1}(x_i))}{Z_t}$ and Z_t is a normalization term.