CMSC 35900 (Spring 2008) Learning Theory **Lecture: 6**

AdaBoost

Instructors: Sham Kakade and Ambuj Tewari

1 AdaBoost

AdaBoost (*Ada*ptive *Boost*ing) is for the case where the parameter γ is not known. The algorithm adapts to the performace of the weak learner.

Algorithm 1 AdaBoost

Input parameters: T

Initialize $w_1 \leftarrow \frac{1}{m}\mathbf{1}$
for $t = 1$ to T **do**
Call γ-WeakLearner with distribution w_t, and receive hypothesis $h_t : X \to [-1, 1]$.
Calculate the error

$$\gamma_t = \frac{1}{2} - \sum_{i=1}^{m} w_{t,i} \frac{|h(x_i) - y_i|}{2}$$

Set

$$\beta_t = \frac{\frac{1}{2} - \gamma_t}{\frac{1}{2} + \gamma_t}, \qquad l_{t,i} = 1 - \frac{|h_t(x_i) - y_i|}{2}$$

and update the weights

$$w_{t+1,i} = \frac{w_{t,i}\beta_t^{l_{t,i}}}{Z_t}, \quad Z_t = \sum_i w_{t,i}\beta_t^{l_{t,i}}$$

end for
OUTPUT the hypothesis:

$$h(x) = \mathrm{sgn}\left(\sum_{t=1}^{T}\left(\log\frac{1}{\beta_t}\right)h_t(x)\right)$$

AdaBoost enjoys the following performance guarantee:

Theorem 1.1. *Let h be the output hypothesis of AdaBoost. Let M be the set of mistakes on the traning set, i.e. $M = \{i : \ h(x_i) \neq y_i\}$. We have:*

$$\frac{|M|}{m} \leq \Pi_{t=1}^{T}\sqrt{1 - 4\gamma_t^2} \leq e^{-2\sum_{t=1}^{T}\gamma_t^2}$$

Proof. We first bound the normalizing constant Z_t using $\beta^x \leq 1 - (1-\beta)x$ for any $x \in [0, 1]$,

$$Z_t = \sum_{i=1}^{m} w_{t,i}\beta^{l_{t,i}} \leq \sum_{i=1}^{m} w_{t,i}\left(1 - (1-\beta_t)l_{t,i}\right) = 1 - (1-\beta_t)\left(\frac{1}{2} + \gamma_t\right) . \tag{1}$$

Next we observe that

$$w_{T+1,i} = w_{1,i}\frac{\prod_{t=1}^{t}\beta^{l_{t,i}}}{\prod_{t=1}^{T} Z_t} . \tag{2}$$

If the output hypothesis h makes a mistake on example i, then

$$y_i \left(\sum_{t=1}^{T} \left(\log \frac{1}{\beta_t} \right) h_t(x_i) \right) \leq 0 \, .$$

Since $y_i \in \{-1, +1\}$, this implies, for all $i \in M$,

$$\prod_{t=1}^{T} \beta_t^{1-\frac{|h_t(x_i)-y_i|}{2}} \geq \left(\prod_{t=1}^{T} \beta_t \right)^{1/2} . \tag{3}$$

Combining (2) and (3), we get

$$\begin{aligned} \sum_{i=1}^{m} w_{T+1,i} \prod_{t=1}^{T} Z_t = \prod_{t=1}^{T} Z_t \\ &= \sum_{i=1}^{m} w_{1,i} \prod_{t=1}^{T} \beta^{l_{t,i}} \\ &\geq \sum_{i \in M} w_{1,i} \left(\prod_{t=1}^{T} \beta^{l_{t,i}} \right)^{1/2} = \frac{|M|}{m} \left(\prod_{t=1}^{T} \beta^{l_{t,i}} \right)^{1/2} . \end{aligned}$$

Rearranging, this gives,

$$\frac{|M|}{m} \leq \prod_{t=1}^{T} \frac{Z_t}{\sqrt{\beta_t}} \, .$$

Combining this with (1), we get

$$\frac{|M|}{m} \leq \prod_{t=1}^{T} \frac{(1 - (1 - \beta_t)(1/2 + \gamma_t))}{\sqrt{\beta_t}} \, .$$

Now substituting $\beta_t = (1/2 - \gamma_t)/(1/2 + \gamma_t)$ proves the theorem. $\square$

2 L1 Margins and Weak Learning

While it may seem that the weak learning is assumption is rather mild, we now show that it is considerably stronger than what one might initially think. In particular, the weak learning assumption is equavalent to a seperability assumption.

We say that we have a γ-weak learner if for every distribution w over the traning set, we can find a hypothesis $h : X \to [-1, 1]$ such that:

$$\sum_{i=1}^{m} w_i \frac{|h(x_i) - y_i|}{2} \leq \frac{1}{2} - \gamma$$

which is equivalent to the condition

$$\sum_{i=1}^{m} w_i y_i h(x_i) \geq 2\gamma$$

which is straightforward to show since $|h(x_i) - y_i| = 1 - y_i h(x_i)$

Let us assume that we have a set of hypothesis

$$\mathcal{H} = \{h_1(\cdot), h_2(\cdot), \ldots h_k(\cdot)\}$$

such that if h is in this set then $-h$ is in this set. Also assume that our weak learning assumpiton holds with respect to this set of hypothesis, meaning that the output of our weak learning always lies in this set $\mathcal{H}$. Note then that our final prediction will be of the form:

$$h_{\text{output}}(x) = \sum_{j=1}^{k} w_j h_j(x)$$

where w is a weight vector.

Define the matrix A such that:

$$A_{i,j} = y_i h_j(x_i) \ .$$

so A is an $m \times k$. Letting S denote the n-dimensional simplex, the weak γ-learning assumption can be stated as follows:

$$\begin{aligned}
2\gamma &\leq \min_{p \in S} \max_{j \in [k]} \sum_{i=1}^{m} p_i y_i h_j(x_i) \\
&= \min_{p \in S} \max_{j \in [k]} |\sum_{i=1}^{m} p_i y_i h_j(x_i)| \\
&= \min_{p \in S} \max_{j \in [k]} |\sum_{i=1}^{m} p_i A_{i,j}| \\
&= \min_{p \in S} \max_{j \in [k]} |[p^{\dagger} A]_j|
\end{aligned}$$

where $\gamma \geq 0$ and we have stated the assumption in matrix notation, in terms of A.

Now let B_1 deonte the L_1 ball of dimension k. We can say that our data-set A is linearly separable with L_1 margin $\alpha \geq 0$ if:

$$\begin{aligned}
\alpha &\leq \max_{w \in B_1} \min_{i \in [m]} y_i (\sum_{j=1}^{k} w_j h_j(x_i)) \\
&= \max_{w \in B_1} \min_{i \in [m]} \sum_{j=1}^{k} w_j A_{i,j} \\
&= \max_{w \in B_1} \min_{i \in [m]} [Aw]_i
\end{aligned}$$

Theorem 2.1. *A is γ weak learnable if and only if A is linearly separable with L_1 margin 2γ.*

Proof. Using the minimax theorem:

$$\begin{aligned}
\min_{p \in S} \max_{j \in [k]} |[p^{\dagger} A]_j| &= \min_{p \in S} \max_{w \in B_1} p^{\dagger} A w \\
&= \max_{w \in B_1} \min_{p \in S} p^{\dagger} A w \\
&= \max_{w \in B_1} \min_{i \in [m]} [Aw]_i
\end{aligned}$$

which completes the proof. $\square$