CSC 588: Machine learning theory **Spring 2021**

Lecture 12: Finish SRM; Adaboost and its training error analysis

Lecturer: Chicheng Zhang *Scribe: Yichen Li*

In previous class, how we can choose the classifier in unfixed hypothesis class $\mathcal{H}$ was discussed. In this class, we finished this topic and start the introduction and proofs of the Boosting methods, especially Adaboost.

1 Continue on Model Selection

Q: How to use $\mathcal{H}_1, ..., \mathcal{H}_k$ to find a good $\hat{h}$ with low error? $\hat{h} = \operatorname{argmin}_{h \in \cup_i \mathcal{H}_i} err(h, S)$ is not a good idea since $\hat{h}_k$ may not be the best among $\{\hat{h}_1, ..., \hat{h}_k\}$.

Idea 1:Validation:

$$\hat{\mathcal{H}} = \{\hat{h}_1, ..., \hat{h}_k\}$$

$\hat{h} = \operatorname{argmin}_{h \in \hat{\mathcal{H}}} err(h, V)$, where V is a fresh validation sample set.

Idea 2: Structural risk minimization(penalized ERM)

$$\hat{i} = \operatorname*{argmin}_{i \in \{1,...,k\}} (err(\hat{h}_i, S) + \sqrt{\frac{\ln \frac{2k|\mathcal{H}_i|}{\delta}}{2m}})^1$$

Output $\hat{h} = \hat{h}_{\hat{i}}$

Performance of Idea 2:

Claim 1. *With probability 1-δ, $\forall i$, $\forall h \in \mathcal{H}_i$*

$$|err(h, S) - err(h, D)| \le \sqrt{\frac{\ln \frac{2k|\mathcal{H}_i|}{\delta}}{2m}}$$

(from standard ERM analysis + union bound over all i)

Claim 2. *If we use structural risk minimization, then with probability 1-δ:*

$$err(\hat{h}, D) \le \min_{i \in \{1,...,k\}} err(h_i^*, D) + 2\alpha(\mathcal{H}_i, m)$$

where,$h_i^* = \operatorname{argmin}_{h \in \mathcal{H}_i} err(h, D)$*,* $\alpha(\mathcal{H}_i, m) = \sqrt{\frac{\ln \frac{2k|\mathcal{H}_i|}{\delta}}{2m}}$

Proof. First we can write the full definition of the output.

$$(\hat{i}, \hat{h}) = \operatorname*{argmin}_{i \in \{1,...,k\}} \min_{h \in \mathcal{H}_i} (err(\hat{h}_i, S) + \alpha(\mathcal{H}_i, m))$$

From claim 1, with probably probability 1-δ, $\forall i$,$\forall h \in \mathcal{H}_i$

$$|err(h, S) - err(h, D)| \le \alpha(\mathcal{H}_i, m)$$

then

[1] penalty for complexity term, define it as $\alpha(\mathcal{H}_i, m)$

$$
\begin{aligned}
err(\hat{h}, D) &= err(\hat{h}_{\hat{i}}, D) \\
&\leq err(\hat{h}_{\hat{i}}, S) + \alpha(\mathcal{H}_{\hat{i}}, m) \\
&\leq err(\hat{h}_i, S) + \alpha(\mathcal{H}_i, m) \\
&\leq err(h_i^*, S) + \alpha(\mathcal{H}_i, m) \\
&\leq err(h_i^*, D) + 2\alpha(\mathcal{H}_i, m)
\end{aligned}
$$

Note that the first and last inequation come from claim 1. The second inequation comes from the optimality of $\hat{h}_{\hat{i}}$ on S with respect to the penalized ERM function. The third comes from optimality of $\hat{h}_i$ on S with respect to the penalized ERM function in the i th hypothesis class and h_i^* must belong to some hypothesis class.

Chicheng notes: the above derivation is slightly different and improves over the derivation in the class; the concentration factor in the lecture was $4\alpha(\mathcal{H}_i, m)$ as opposed to $2\alpha(\mathcal{H}_i, m)$.

Since $\forall i$ the above inequation is true, we have $err(\hat{h}, D) \leq \min_{i \in \{1,\ldots,k\}} err(h_i^*, D) + 4\alpha(\mathcal{H}_i, m)$ □

Remarks:

- **α can pessimistic in practice.**
- **It may be possible to refine the generalization err bounds.**

2 Introduction for Boosting methods

Motivation: Combine weak classification rules to obtain strong ones.
Example: When doing spam filtering, we can use "free offer" or "million dollar" to detect spams. These simple words based filters are weak by individual but when combined together they can perform better.
Weak PAC Learning Theory:
γ-weak PAC learner:
Given $\mathcal{H}$, we call $\mathcal{A}$ is a γ-weak PAC learner for $\mathcal{H}$, if exists a function $f : (0, 1) \to \mathbb{N}$ for any $\delta > 0$ and any D realizable by $\mathcal{H}$, with $m \geq f(\delta)$ iid training examples, $\mathcal{A}$ produces a classifier h $s.t.$ with probability $1 - \delta$, $\text{err}(h, D) \leq \frac{1}{2} - \gamma$.
γ-weak PAC learnable:
$\mathcal{H}$ is said to be γ-weak PAC learnable if there exist a γ-weak PAC learner for $\mathcal{H}$.
History:
$\mathcal{H}$ PAC learnable can induce $\mathcal{H}$ weak PAC learnable, what about the inverse?
1988 Kearns raised this question.
1990 Schapire proposed boosting, build a PAC learner with black-box access to a weak PAC learner based on recursion.
1990 Freund proposed boost by majority, combining the outputs of weak learners by a unweighted majority vote.
1997 Freund and Schapire proposed Adsboost(adaptive boost) which is a γ free method (only require the weak PAC learner to be slightly better than random guess).

3 AdaBoost

Basic idea: Given training examples $(x_1, y_1)...(x_m, y_m)$, the algorithm maintains a weighting on them which starts with uniform and then adjusted by training errors. (Assign higher weights on examples that were mistaken in the training process.)
Adaboost Algorithm:
Initialize $(D_{t=1}(i) = \frac{1}{m})_{i=1}^m$

```
for $t = 1, \dots, T$: do
    Train predictor $h_t$ on weighted examples $(x_i, y_i, D_t(i)_{i=1}^m)$
    Receive weighted error $\epsilon_t = P_{(x,y)\in D_t}(h_t(x) \neq y) = \sum_{i=1}^m I(h_t(x_i) \neq y_i) D_t(i) \leq \frac{1}{2} - \gamma$
    Assign classifier weight $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$
    Update weights on training samples $D_{t+1}(i) = \frac{1}{Z_t} D_t(i) \exp(-\alpha_t y_i h_t(x_i))$, where $Z_t$ is a normalization
    factor obtained by $Z_t = \sum_{i=1}^m D_t(i) \exp(-\alpha_t y_i h_t(x_i))$.
end for
return $H_T(x) = \text{sign}(\sum_{t=1}^T \alpha_t h_t(x))$
```

Analysis:

Theorem 3. *Suppose for every* $t, \epsilon_t \leq \frac{1}{2} - \gamma$, *then*

$$err(H_T, S) \leq \exp(-2T\gamma^2)$$

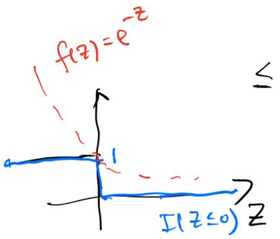
Proof steps:

1. Relax to exponential loss.
2. Adaboost optimizes exponential loss.

Proof.

$$\begin{aligned} err(H_T, S) &= \frac{1}{m}\sum_{i=1}^m (H_T(x_i) \neq y_i) \\ &= \frac{1}{m}\sum_{i=1}^m I(y_i f_T(x_i) \leq 0) \\ &\leq \frac{1}{m}\sum_{i=1}^m \exp(-y_i f_T(x_i)) \\ &= L_T \end{aligned}$$

The second equation comes from $f_t = \sum_{s=1}^t \alpha_s h_s(x)$, while the first inequation comes from $I(z \leq 0) \leq \exp(-z)$ which is shown below. Finally, L_T is defined as $\frac{1}{m}\sum_{i=1}^m \exp(-y_i f_T(x_i))$.



Then we just need to show that L_T (cumulative exponential loss) is decreasing exponentially to T. First, we can take a look of $\frac{L_t}{L_{t-1}}$

$$\frac{L_t}{L_{t-1}} = \frac{\frac{1}{m}\sum_{i=1}^m \exp(-y_i f_t(x_i))}{\frac{1}{m}\sum_{i=1}^m \exp(-y_i f_{t-1}(x_i))}$$

Observation:

$$D_t(i) \propto \exp(-\sum_{s=1}^{t-1} \alpha_s y_i h_s(x_i)) = \exp(-y_i f_{t-1}(x_i))$$

We can see from the definition that $\sum_{i=1}^m D_t(i) = 1$ and there exist N_t such that $D_t(i) = \exp(-y_i f_{t-1}(x_i))/N_t$. Also, by definition, $f_t(x_i) = f_{t-1}(x_i) + \alpha_t h_t(x_i)$ By plugging them back to $\frac{L_t}{L_{t-1}}$ we have

$$\begin{aligned}
\frac{L_t}{L_{t-1}} &= \frac{\sum_{i=1}^m N_t D_t(i) \exp(-y_i \alpha_t h_t(x_i))}{\sum_{i=1}^m N_t D_t(i)} \\
&= \frac{N_t \sum_{i=1}^m D_t(i) \exp(-y_i \alpha_t h_t(x_i))}{N_t \sum_{i=1}^m D_t(i)} \\
&= \sum_{i=1}^m D_t(i) \exp(-y_i \alpha_t h_t(x_i)) \\
&= Z_t
\end{aligned}$$

We can see $L_T = L_{T-1} Z_T = L_0 \prod_{t=1}^T Z_t$
Finally, by upper bounding Z_t we can conclude our proof.

$$\begin{aligned}
Z_t &= \sum_{i=1}^m D_t(i) \exp(-y_i \alpha_t h_t(x_i)) \\
&= \sum_{i:y_i=h_t(x_i)} D_t(i) \exp(-\alpha_t) + \sum_{i:y_i \neq h_t(x_i)} D_t(i) \exp(\alpha_t) \\
&= \exp(-\alpha_t) * (1-\epsilon_t) + \exp(\alpha_t) * \epsilon_t \\
&= \sqrt{\frac{\epsilon_t}{1-\epsilon_t}}(1-\epsilon_t) + \sqrt{\frac{1-\epsilon_t}{\epsilon_t}}\epsilon_t \\
&= 2\sqrt{\epsilon_t(1-\epsilon_t)} \\
&\leq \sqrt{1-4\gamma^2} \\
&\leq \exp(-2\gamma^2)
\end{aligned}$$

The third equation comes from the definition of $\alpha_t = \frac{1}{2}\ln(\frac{1-\epsilon_t}{\epsilon_t})$,while the first inequation comes from $\epsilon_t \leq \frac{1}{2} - \gamma$, $x^2 \geq (x-y)(x+y)$ and the second inequation comes from $1 - x \leq e^{-x}$.
With Z_t bounded and $L_0 = 1$, we conclude

$$L_T = L_0 \prod_{t=0}^T Z_t \leq \exp(-2T\gamma^2)$$

□