Lecture 6 — September 1

Instructor: Stas Minsker *Scribe: Mose Wintner*

6.1 Boosting and Adaboost (adaptive boosting)

Invented by R. Schapire and Yoav Freund.

Generalization error of g:

$$P(Y \neq g(x)) = \mathbb{E}I\{Yg(X) \leq 0\}.$$

We estimated the binary loss of the marginal with an exponential:

$$\ell_{0-1}(y, g(x)) = I\{yg(x) \leq 0\} \leq e^{-yg(x)}.$$

Then

$$\overline{g} = \operatorname*{argmin}_{g:\mathbb{S}\to\mathbb{R}} \mathbb{E}e^{-yg(x)} = \frac{1}{2}\ln\frac{1+\eta}{1-\eta}$$

where we see $\operatorname{sign}(\overline{g}(x)) = \operatorname{sign}(\eta(x))$. We seek to estimate $\overline{g}$ by

$$\widehat{g}_n = \operatorname*{argmin}_{g\in G} \frac{1}{n}\sum_{j=1}^{n} e^{-Y_j g(X_j)} \approx \mathbb{E}e^{-Yg(X)}$$

for any fixed g. The map

$$g \mapsto \frac{1}{n}\sum_{j=1}^{n} e^{-Y_j g(X_j)}$$

is convex. If G is convex, then $\widehat{g}_n$ is the solution of the convex minimization problem.

Let $\mathcal{F}$ be the "base class" (also called the "weak learners") consisting of the binary classifiers (i.e. $f \in \mathcal{F}$ takes values in $\{\pm 1\}$.) Let G be the closure of the real linear span of $\mathcal{F}$. Let's examine one step of the steepest descent algorithm for this convex minimization problem. Assume that $g \in G$ is our current guess. We want to look for for $\alpha \in \mathbb{R}$ and $f \in \mathcal{F}$ that minimizes

$$\frac{1}{n}\sum_{j=1}^{n} e^{-Y_j(g(X_j)+\alpha f(X_j))} = \frac{1}{n}\sum_{j=1}^{n} e^{-Y_j g(X_j)} e^{-\alpha f(X_j)Y_j} =: \frac{1}{n}\sum_{j=1}^{n} w_j e^{-\alpha f(X_j)Y_j}.$$

We'll normalize the weights: let $\tilde{w}_j = w_j/(\sum w_j)$. Then our problem is to minimize

$$\frac{1}{n}\sum_{j=1}^{n} \tilde{w}_j e^{-\alpha f(X_j)Y_j}$$

over $f \in \mathcal{F}$ and $\alpha \in \mathbb{R}$.

Because we're dealing with binary classifiers, we have

$$
\begin{aligned}
&\sum_{j=1}^{n} \tilde{w}_j e^{-\alpha f(X_j)Y_j} = \\
&\quad = \sum_{j=1}^{n} \tilde{w}_j e^{-\alpha} I\{Y_j = f(X_j)\} + \sum_{j=1}^{n} \tilde{w}_j e^{\alpha} I\{Y_j \neq f(X_j)\} \pm \sum_{j=1}^{n} \tilde{w}_j e^{-\alpha} I\{Y_j \neq f(X_j)\} \\
&\quad = e^{-\alpha} + (e^{\alpha} - e^{-\alpha}) \sum_{j=1}^{n} \tilde{w}_j I\{Y_j \neq f(X_j)\}
\end{aligned}
$$

where the last term is referred to as the "weighted training error" $e_{n,\tilde{\omega}}(f)$.

to minimze the resulting expression, we proceed in two steps:

1) minimize $\sum_{j=1}^{n} \tilde{w}_j I\{Y_j \neq f(X_j)\}$ with respect to f

2) minimize $e^{-\alpha} + 2\sinh(\alpha) e_{n,\tilde{\omega}}(f)$ with respect to $\alpha \in \mathbb{R}$.

By weak learnability, we mean that for any $\tilde{w}_1, \ldots \tilde{w}_n$ such that $\tilde{w}_j \geq 0$ for all $j = 1, \ldots, n$ and $\sum \tilde{w}_j = 1$, there exists $f \in \mathcal{F}$ such that $e_{n,\tilde{w}}(f) \leq 1/2$. We always have this due to symmetry: if our classifier f has error $\geq 1/2$, we can replace it with $-f$, which has error $\leq 1/2$.

We make the "weak learnability assumption", to satisfy 1), i.e. that we have access to a black box weak learning algorithm that takes $\tilde{w}_j$, $j = 1, \ldots, n$ and data $(X_1, Y_1), \ldots, (X_n, Y_n)$ as input and outputs $f \in \mathcal{F}$ such that $e_{n,\tilde{\omega}}(f) \leq 1/2$.

Step 2) is a calculus problem:

$$
\begin{aligned}
0 &= -e^{-\alpha} + 2\cosh(\alpha) e_{n,\tilde{\omega}}(f) \\
1 &= (e^{2\widehat{\alpha}} + 1) e_{n,\tilde{\omega}} \\
\widehat{\alpha} &= \frac{1}{2} \log\left(\frac{1 - e_{n,\tilde{\omega}}(f)}{e_{n,\tilde{\omega}}(f)}\right)
\end{aligned}
$$

This is the Adaboost algorithm; it is the first example of weak learnability $\Rightarrow$ strong learnability.

6.1.1 Adaboost

- $\omega_j^{(0)} ;= \frac{1}{n}$.
- For $t = 0, \ldots T$, call "weak learner" (WL) that
  - outputs binary classifier f_t such that the weighted training error $e_{n,w^{(t)}}(f_t) \leq 1/2$

  - sets $\alpha_t = \frac{1}{2}\log\left(\frac{1-e_{n,\tilde{\omega}}(f_t)}{e_{n,\tilde{\omega}}(f_t)}\right)$

- update the weights $w_j^{(t+1)} = w_j^{(t)} \exp\left(-Y_j \alpha_t f_t(X_j)\right))$ for each j.

Output:

$$\widehat{g}^T(\cdot) = \text{sign}\left(\sum_{j=1}^{T} \alpha_t f_t(\cdot)\right).$$

Can be thought of as an average vote among the weak learners.