Math 547: Mathematical Foundations of Statistical Learning Theory

Lectures 1-2 — Week 4

Instructor: S. Minsker *Scribe: S. Minsker*

1.1 Adaboost continued

Last week, we started talking about Adaboost algorithm, due to R. Schapire and Y. Freund. It was originally motivated by the following question: given a class G that satisfies a weak learnability condition, can one find $\hat{g}$ such that

$$P_n \mathbb{I}\left(y \neq \hat{g}(x)\right) \leq \varepsilon$$

for any $\varepsilon > 0$? For instance, such a $\hat{g}$ can be found by "combining" the elements of G.

We will derive Adaboost as a steepest descent method for a specific problem by asking:

Question. How can we replace minimization of the binary loss by a numerically feasible problem?

Note that since Y is a binary label we have that

$$\mathbb{P}\left(Y \neq g(X)\right) = \mathbb{P}\left(\underbrace{Yg(X)}_{\text{"the margin"}} \leq 0\right)$$

since when $Y \neq g(X)$, Y and $g(X)$ have different signs. The product $Yg(X)$ is called **the margin**.

Recall that

$$\mathbb{P}\left(Yg(X) \leq 0\right) = \mathbb{E}\left[\mathbb{I}\left(\underbrace{Yg(X)}_{\ell(t)} \leq 0\right)\right]$$

and now we want to bound this from above by some convex function $\ell(t)$, as shown in figure 1.1, namely

$$\mathbb{E}\left[\mathbb{I}\left(Yg(X) \leq 0\right)\right] \leq \mathbb{E}\left[\ell(Yg(X))\right].$$

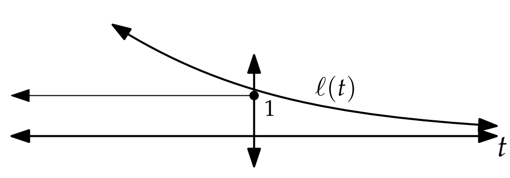

Figure 1.1. The function $Yg(X)$ being bound above by $\ell(t)$.

So let's choose a "nice" function, say $\ell(t) = e^{-t}$. Now, the key question here is: what are the properties of $\mathbb{E}\left[\exp\left(-Yg(X)\right)\right]$?

Lemma 1. Let

$$\bar{g} = \operatorname*{argmin}_{g:\mathbb{S}\to\{\pm 1\}} \mathbb{E}\left[\exp\left(-Yg(X)\right)\right].$$

Then, $\operatorname{sign}\bar{g} = \operatorname{sign}\eta$.

Proof. By the law of total expectation we have

$$\mathbb{E}\left[\exp\left(-Yg(X)\right)\right] = \mathbb{E}\left[\mathbb{E}\left[\exp\left(-Yg(X)\right)|X\right]\right].$$

Recall from the previous lecture that $\mathbb{P}\left(Y = 1|X = x\right) = \dfrac{1+\eta(x)}{2}$ and $\mathbb{P}\left(Y = -1|X = x\right) = \dfrac{1-\eta(x)}{2}$. Thus,

$$\begin{aligned}
\mathbb{E}\left[\exp\left(-Yg(X)\right)\right] &= \int \left[\exp\left(-1\cdot g(x)\right)\cdot \mathbb{P}\left(Y=1|X=x\right) + \exp\left(1\cdot g(x)\right)\cdot \mathbb{P}\left(Y=-1|X=x\right)\right] d\Pi(x) \\
&= \int \underbrace{\left[\exp\left(-g(x)\right)\left(\frac{1+\eta(x)}{2}\right) + \exp\left(g(x)\right)\left(\frac{1-\eta(x)}{2}\right)\right]}_{\text{Nonnegative expression, so it can be minimized pointwise}} d\Pi(x)
\end{aligned}$$

Let $g(X) = t$. We now want to minimize the following function $h(t)$ with respect to $t \in \mathbb{R}$:

$$h(t) = e^{-t}\left(\frac{1+\eta(x)}{2}\right) + e^{t}\left(\frac{1-\eta(x)}{2}\right).$$

Taking the derivative and setting it equal to zero gives us

$$\begin{aligned}
e^{t}\left(\frac{1-\eta(x)}{2}\right) - e^{-t}\left(\frac{1+\eta(x)}{2}\right) &= 0 \\
\Rightarrow e^{2t} &= \frac{1+\eta(x)}{1-\eta(x)} \\
\Rightarrow t &= \frac{1}{2}\log\left(\frac{1+\eta(x)}{1-\eta(x)}\right)
\end{aligned}$$

Therefore,

$$\bar{g}(x) = \frac{1}{2}\log\left(\frac{1+\eta(x)}{1-\eta(x)}\right).$$

Thus, we find that $\operatorname{sign}\bar{g} = \operatorname{sign}\eta$ since $\operatorname{sign}\bar{g}(x) = 1 \Rightarrow 1+\eta(x) > 1-\eta(x) \Rightarrow \eta(x) > 0$, and $\operatorname{sign}\bar{g}(x) = -1 \Rightarrow 1+\eta(x) < 1-\eta(x) \Rightarrow \eta(x) < 0$. □

Conclusion. We have shown that $\operatorname{sign}\bar{g}$ is the **Bayes classifier**!

We now set our sights forward on to our next goal: consider the empirical risk minimization problem

$$\frac{1}{n}\sum_{j=1}^{n} \exp\left(-Y_j g(X_j)\right) \to \min_{g\in\mathbb{G}} \tag{1.1}$$

Note that this problem is convex with respect to g as long as the class $\mathbb{G}$ is convex.

1.2 AdaBoost

Define

$$\hat{g}_n = \operatorname*{argmin}_{g\in\mathbb{G}} \frac{1}{n}\sum_{j=1}^{n} \exp\left(-Y_j g(X_j)\right), \tag{1.2}$$

where $\mathbb{G}$ is a class of functions $S \mapsto \mathbb{R}$. If $\mathbb{G}$ is convex, then $\hat{g}_n$ is the solution of the convex minimization problem. Let $\mathcal{F}$ be the "base class" (the collection of "weak learners"), and set

$$\mathbb{G} := \text{closed linear span of } \mathcal{F} = \overline{\left\{\sum_{j=1}^{k} \alpha_j f_j : k \geq 1, \alpha_0, \ldots, \alpha_k \in \mathbb{R}, f_0, \ldots, f_k \in \mathcal{F}\right\}}.$$

Then $\mathbb{G}$ is indeed convex and closed. Let's examine one step of the (version of) the steepest descent algorithm for (1.2). Assume that $g \in \mathbb{G}$ is our current guess. We will look for $\alpha \in \mathbb{R}$ and $f \in \mathcal{F}$ that minimize (at least approximately)

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

Intuitively, such an f can be seen as an "approximate gradient". Define $w_j = \frac{1}{n}e^{-Y_j f(X_j)}$, $j = 1, \ldots, n$, to be the weights. Note that $w_j \geq 0$. Let $\tilde{w}_j = \frac{w_j}{\sum_{j=1}^{n} w_j}$, so that $\sum_{j=1}^{n} \tilde{w}_j = 1$. Our problem is then to minimize $\sum_{j=1}^{n} \tilde{w}_j e^{-\alpha f(X_j) Y_j}$ over $f \in \mathcal{F},\ \alpha \in \mathbb{R}$. Since f takes only two values ± 1, we have that

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

where $e_{n,\tilde{w}}(f) = \sum_{j=1}^{n} \tilde{w}_j \mathbb{I}\left(Y_j \neq f(X_j)\right)$ is the "weighted" training error. To minimize the resulting expression, we proceed in two steps:

1. Minimize $\sum_{j=1}^{n} \tilde{w}_j \mathbb{I}\left(Y_j \neq f(X_j)\right)$ with respect to f
2. Minimize $e^{-\alpha} + (e^{\alpha} - e^{-\alpha}) e_{n,\tilde{w}}(f)$ with respect to α.

To complete step 1, we need the following "weak learnability" assumption: for any nonnegative weights $\tilde{w}_1, \ldots, \tilde{w}_n$ with $\sum_{j=1}^{n} \tilde{w}_j = 1$, $\exists f \in \mathcal{F}$ such that $e_{n,\tilde{w}}(f) \leq \frac{1}{2}$. Weak learnability is implied by symmetry, meaning that $\mathcal{F} = -\mathcal{F}$; indeed, if $e_{n,\tilde{w}}(f) > \frac{1}{2}$ then $e_{n,\tilde{w}}(-f) < \frac{1}{2}$. For instance, the class of decision stumps is symmetric. We will assume access to a "black box" weak learning algorithm that takes $\tilde{w}_1, \ldots, \tilde{w}_n$ and $(X_1, Y_1), \ldots, (X_n, Y_n)$ as inputs and outputs some $f \in \mathcal{F}$ such that $e_{n,\tilde{w}}(f) \leq \frac{1}{2}$; an example of such an algorithm for the class of decision stumps was discussed before.

Assuming that $e_{n,\tilde{w}}(f) \leq \frac{1}{2}$, the minimum of $e^{-\alpha} + (e^{\alpha} - e^{-\alpha})e_{n,\tilde{w}}(f)$ occurs for

$$\hat{\alpha} = \frac{1}{2}\log\frac{1 - e_{n,\tilde{w}}(f)}{e_{n,\tilde{w}}(f)} \geq 0.$$

Adaboost is an algorithm that repeats the steps outlined above. We present it now.

Adaboost algorithm:

Initialize $w_j^{(0)} = \frac{1}{n}$, $j = 1, \dots, n$. For $t = 0, \dots, T$ do

- Call the weak learner (WL);
- Output f_t such that $e_{n,w^{(0)}}(f_t) \leq \frac{1}{2}$;
- Set $\alpha_t = \frac{1}{2}\log\frac{1-e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)}$;
- Update the weights $w_j^{(t+1)} = \frac{w_j^{(t)}\exp(-Y_j\alpha_t f_t(X_j))}{Z_t}$, $j = 1 \dots n$, where $Z_t = \sum_{j=1}^n w_j^{(t)} \exp\left(-Y_j\alpha_t f_t(\cdot)\right)$ is the "normalizing factor."
- Output: $\widehat{g}_T(\cdot) = \text{sign}\left(\sum_{j=1}^T \alpha_t f_t(\cdot)\right)$.

Exercise 1. If f_t classifies X_j correctly, then $w_j^{(t+1)} \leq w_j^{(t)}$. If f_t classifies X_j incorrectly, then $w_j^{(t+1)} \geq w_j^{(t)}$.

Theorem 1. Assume that at each step, WL outputs f_t such that

$$e_{n,w^{(t)}}(f_t) = \sum_{j=1}^n w_j^{(t)} I\{Y_j \neq f_t(X_j)\} \leq \frac{1}{2} - \gamma,$$

for some $\gamma > 0$. Then the *training error* corresponding to the classifier $\hat{g}_T$ satisfies

$$\frac{1}{n}\sum_{j=1}^n I\{Y_j \neq \widehat{g}_T(X_j)\} \leq \exp\left(-2T\gamma^2\right).$$

Proof.

a) Note that $w_j^{(T+1)} = \frac{1}{n}\frac{e^{-Y_j\sum_{t=1}^T \alpha_t f_t(X_j)}}{\prod_{t=1}^T Z_t}$; this is easy to show by induction.

b) We have that

$$\begin{aligned}
\frac{1}{n}\sum_{j=1}^n I\{Y_j \neq \widehat{g}_T(X_j)\} &= \frac{1}{n}\sum_{j=1}^n I\{Y_j \sum_{t=1}^T \alpha_t f_t(X_j) \leq 0\} \\
&\leq \frac{1}{n}\sum_{j=1}^n e^{-Y_j\sum_{t=1}\alpha_t f_t(X_j)} \\
&= \frac{1}{n}\sum_{j=1}^n w_j^{(T+1)} n \prod_{t=1}^T Z_t \\
&= \prod_{t=1}^T Z_t.
\end{aligned}$$

c) For Z_t at each step

$$
\begin{aligned}
Z_t &= \sum_{j=1}^{n} w_j(t) \exp\left(-Y_j \alpha_t f_t(X_j)\right) \\
&= \sum_{j=1}^{n} w_j^{(t)} I\{Y_j = f_t(X_j)\} e^{-\alpha_t} + \sum_{j=1}^{n} w_j^{(t)} I\{Y_j \neq f_t(X_j)\} e^{\alpha_t} \pm \sum_{j=1}^{n} w_j I\{Y_j \neq f_t(X_j)\} e^{-\alpha_t} \\
&= e^{-\alpha_t} + (e^{\alpha_t} - e^{-\alpha_t}) \sum_{j=1}^{n} w_j^{(t)} I\{Y_j \neq f_t(X_j)\},
\end{aligned}
$$

where the last multiplicand is $e_{n,w^{(t)}}(f_t)$. Recall that $\alpha_t = \frac{1}{2}\log\left(\frac{1-e_{n,w^{(t)}}(f_t)}{e_{n,w^{(t)}}(f_t)}\right)$, we thus have that

$$
Z_t = 2\sqrt{e_{n,w^{(t)}}(f_t)(1 - e_{n,w^{(t)}}(f_t))}.
$$

d) The function $f(x) = x(1-x); x \in [0, \frac{1}{2} - \gamma]$ is maximized for $x = \frac{1}{2} - \gamma$, thus

$$
\mathbb{Z}_t \leq 2\sqrt{(1/2-\gamma)(1/2+\gamma)} \leq \sqrt{1-4\gamma^2} \leq \sqrt{e^{-4\gamma^2}} = e^{-2\gamma^2},
$$

since $1 - x \leq e^{-x}$ for $x \in [0,1]$. Therefore

$$
\frac{1}{n}\sum_{j=1}^{n} I\{Y_j \neq \widehat{g}_T(X_j)\} = \prod_{t=1}^{T} Z_t \leq \exp\left(-2T\gamma^2\right).
$$

□

In conclusion, the training error goes to 0 exponentially fast. However, the main object of interest is the *generalization error*

$$
P(Y\widehat{g}_T(X)) \leq 0.
$$

Estimating the generalization error turns out to be a much harder problem that we will consider later in this course.