CS281B/Stat241B. Statistical Learning Theory. Lecture 25.

Peter Bartlett

Overview

- AdaBoost
  - Weak learning implies strong learning.
  - Weak learning is equivalent to large margin combination.
  - AdaBoost minimizes exponential surrogate loss.
  - Coordinate descent with other losses.

AdaBoost

- Pattern classification method: makes a decision based on the predictions of a committee of classifiers.
- One motivation: often easy to come up with simple rules that give some edge over random guessing. Hope that a combination of such classifiers will produce an accurate decision rule.
- A way to improve the performance of a base learning algorithm by combining predictions of decision rules.

AdaBoost: Examples of base classifiers

- Decision trees, $f_T : \mathbb{R}^d \to \{\pm 1\}$, defined by a binary tree T with nodes labeled with decision stumps:

$$f_T(x) = \begin{cases} s(x) & \text{if no descendants,} \\ f_L(x) & \text{if } s(x) = -1, \\ f_R(x) & \text{if } s(x) = 1, \end{cases}$$

  where L, R are the left and right subtrees and s is the decision stump at the root:

$$s_{a,b,i}(x) = \text{sign}(b(x_i - a) > 0).$$

AdaBoost: Examples of base classifiers

- Linear threshold functions, $f_\theta(x) = \text{sign}(\theta^T x)$.

  Then a weighted combination of these functions is

$$f(x) = \text{sign}\left(\sum_i \alpha_i \, \text{sign}(\theta_i^T x)\right),$$

  which is a two-layer neural network.

- Fixed dictionary: family of simple rules.

  For example, in parsing, we might use simple rules that distinguish between simple families of subtrees. In detecting objects in images, we might use simple image features, such as differences of rectangular sums, or similarity to image patches.

AdaBoost

AdaBoost (and other ensemble methods, such as *bagging*), work with a function class

$$F = \left\{ x \mapsto \text{sign}\left(\sum \alpha_t f_t(x) \right) : \alpha_t \in \mathbb{R},\ f_t \in \mathcal{G} \right\},$$

where $\mathcal{G} \subseteq \{\pm 1\}^{\mathcal{X}}$ is the class of base classifiers.

AdaBoost:

- Maintains weighting (probability distribution) D_t over training data $\{1, \ldots, n\}$.
- Chooses f_t sequentially, to minimize weighted empirical risk.
- Adjusts weighting to emphasize mistakes.

AdaBoost

$D_1(i) = \frac{1}{n}, i = 1, \ldots, n.$

$F_0(x) = 0.$

for $t = 1, \ldots, T$ **do**

Choose $f_t \in \mathcal{G}$ to (approximately) minimize

$$\epsilon_t = \sum_{i=1}^{n} D_t(i) 1[f_t(x_i) \neq y_i].$$

$F_t = F_{t-1} + \alpha_t f_t.$

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{\alpha_t} & \text{if } f_t(x_i) \neq y_i, \\ e^{-\alpha_t} & \text{otherwise.} \end{cases}$$

end for

Here, $\alpha_t = \frac{1}{2} \log \frac{1-\epsilon_t}{\epsilon_t}$, and the normalization constant is $Z_t = \sum_i D_t(i) \exp\left(-y_i f_t(x_i) \alpha_t\right) = 2\sqrt{\epsilon_t(1-\epsilon_t)}$.

AdaBoost: Weak learning implies strong learning

Theorem: Define

$$\overline{F} = \frac{F_T}{\sum_{t=1}^{T} \alpha_t} = \frac{\sum_t \alpha_t f_t}{\sum_t \alpha_t} \in \mathrm{co}(\mathcal{G}).$$

Then

$$P_n\left(Y\overline{F}(X) \leq 0\right) = \frac{1}{n}\left|\left\{i : y_i \overline{F}(x_i) \leq 0\right\}\right| \leq \prod_{t=1}^{T} 2\sqrt{\epsilon_t(1-\epsilon_t)}.$$

AdaBoost: Weak learning implies strong learning

Furthermore, if $\epsilon_t \leq 1/2 - \gamma$ for all t, we have

$$\prod_{t=1}^{T} 2\sqrt{\epsilon_t(1-\epsilon_t)} = 2^T \left(\frac{1}{4} - \gamma^2\right)^{T/2} = \left(1 - 4\gamma^2\right)^{T/2},$$

which is no more than ϵ for

$$T \geq \frac{\ln 1/\epsilon}{2\gamma^2}.$$

Weak learning implies strong learning: Proof

Use the Chernoff idea:

$$y\overline{F}(x) \leq 0 \Leftrightarrow y\sum_t \alpha_t f_t(x) \leq 0$$

$$\Leftrightarrow \exp\left(-y\sum_t \alpha_t f_t(x)\right) \geq 1.$$

Let Z_t denote the normalization constant at round t (we'll calculate its value later):

$$D_{t+1}(i) = \frac{D_t(i)\exp\left(-y_i f_t(x_i)\alpha_t\right)}{Z_t}.$$

So $$\frac{D_{t+1}(i)}{D_t(i)}Z_t = \exp\left(-y_i f_t(x_i)\alpha_t\right).$$

Weak learning implies strong learning: Proof

$$
\begin{aligned}
P_n\left(Y\overline{F}(x) \le 0\right) &\le \mathbb{E}_n \exp\left(-Y\overline{F}(X)\right) \\
&= \mathbb{E}_n \exp\left(-Y\sum_{t=1}^{T}\alpha_t f_t(X)\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\prod_{t=1}^{T}\exp\left(-y_i f_t(x_i)\alpha_t\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\prod_{t=1}^{T}\frac{D_{t+1}(i)}{D_t(i)}Z_t \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{D_{T+1}(i)}{D_1(i)}\prod_t Z_t = \prod_t Z_t.
\end{aligned}
$$

Weak learning implies strong learning: Proof

Hence, it makes sense to choose α_t to minimize the normalization factor:

$$\begin{aligned} Z_t &= \sum_{i:y_i=f_t(x_i)} D_t(i)e^{-\alpha_t} + \sum_{i:y_i\neq f_t(x_i)} D_t(i)e^{\alpha_t} \\ &= (1-\epsilon_t)e^{-\alpha_t} + \epsilon_t e^{\alpha_t}. \end{aligned}$$

Differentiating wrt α_t and solving gives

$$\begin{aligned} \alpha_t &= \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right), \\ Z_t &= 2\sqrt{(1-\epsilon_t)\epsilon_t}. \end{aligned}$$

Weak learning implies large margins

Can extend to show large margins on training data:

Theorem:

$$P_n\left(Y\overline{F}(X) \leq \gamma\right) \leq \prod_{t=1}^{T} 2\sqrt{\epsilon_t^{1-\gamma}(1-\epsilon_t)^{1+\gamma}},$$

and if $\epsilon_t \leq 1/2 - 2\gamma$ for all t, this decreases to zero exponentially quickly.

Weak learning implies large margins

The earlier theorem implies that if, for any distribution D_t on $\{1, \ldots, n\}$, there is an $f_t \in \mathcal{G}$ with weighted empirical risk no more than $1/2 - \gamma$, then there is a $\overline{F} \in \operatorname{co} \mathcal{G}$ with zero empirical risk.

This theorem implies that if, for any distribution D_t on $\{1, \ldots, n\}$, there is an $f_t \in \mathcal{G}$ with weighted empirical risk no more than $1/2 - \gamma$, then there is a $\overline{F} \in \operatorname{co} \mathcal{G}$ with large margins: $P_n(Y\overline{F}(X) \leq \gamma/2) = 0$.

The following converse result shows that the assumption of the existence of a 'weak learner' (that produces an f_t with risk no more than $1/2 - \gamma$) is equivalent to the existence of a large margin convex combination.

Weak learning is equivalent to large margins

Theorem: If, for $(x_1, y_1), \ldots, (x_n, y_n)$, there is a $F \in \operatorname{co} \mathcal{G}$ with

$$y_i F(x_i) \geq \gamma \qquad \text{for } i = 1, \ldots, n,$$

then for all probability distributions D on $\{1, \ldots, n\}$, there is an $f \in \mathcal{G}$ with

$$\sum_{i=1}^{n} D(i) 1[y_i \neq f(x_i)] \leq \frac{1-\gamma}{2}.$$

Weak learning is equivalent to large margins: Proof

The proof uses the probabilistic method. Suppose $F = \sum_t \alpha_t f_t$ with $\sum \alpha_t = 1$, $\alpha_t \geq 0$. Choose f randomly, with $\Pr(f = f_t) = \alpha_t$. Then for any D,

$$\begin{aligned}
0 \leq \mathbb{E} \sum_{i=1}^{n} D(i) 1[f(x_i) \neq y_i] &= \sum_t \alpha_t \sum_i D(i) 1[f_t(x_i) \neq y_i] \\
&= \sum_i D(i) \sum_t \alpha_t \frac{1 - y_i f_t(x_i)}{2} \\
&\leq \sum_i D(i) \frac{1 - \gamma}{2} \\
&= \frac{1 - \gamma}{2}.
\end{aligned}$$

Thus, there exists an f with this property.

AdaBoost minimizes exponential surrogate loss

In the proof of weak $\Rightarrow$ strong, we used this observation:

$$\frac{D_{s+1}(i)}{D_s(i)} Z_s = \exp(-y_i f_s(x_i)\alpha_s).$$

Since the product telescopes, this implies that

$$D_t(i) \prod_{s=1}^{t-1} Z_s = D_1(i) \prod_{s=1}^{t-1} \exp(-y_i f_s(x_i)\alpha_s) = \frac{1}{n} \exp(-y_i F_{t-1}(x_i)).$$

That is, the weighting $D_t(i)$ is proportional to $\exp(-y_i F_{t-1}(x_i))$.

AdaBoost minimizes exponential surrogate loss

Theorem: If $\mathcal{G} = -\mathcal{G}$ (that is, $g \in \mathcal{G} \Rightarrow -g \in \mathcal{G}$), then choosing $f_t \in \mathcal{G}$ to minimize

$$\sum_{i=1}^{n} D_t(i) 1[y_i \neq f_t(x_i)]$$

and choosing

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

is equivalent to choosing f_t, α_t to minimize

$$\mathbb{E}_n \exp\left(-Y F_t(X)\right) = \mathbb{E}_n \exp\left(-Y \left(\alpha_t f_t(X) + F_{t-1}(X)\right)\right).$$

AdaBoost minimizes exponential surrogate loss: Proof

Because $\frac{1}{n}\exp(-y_i F_{t-1}(x_i)) = D_t(i)\prod_{s=1}^{t-1} Z_s$, we have

$$\mathbb{E}_n \exp\left(-Y F_t(X)\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(\left(e^{\alpha_t} - e^{-\alpha_t}\right) 1[y_i \neq f_t(x_i)] + e^{-\alpha_t}\right)\exp\left(-y_i F_{t-1}(x_i)\right)$$

$$= \left(e^{\alpha_t} - e^{-\alpha_t}\right)\prod_{s=1}^{t-1} Z_s \underbrace{\sum_{i=1}^{n} D_t(i) 1[y_i \neq f_t(x_i)]}_{(*)}$$

$$+ \frac{e^{-\alpha_t}}{n}\sum_{i=1}^{n}\exp\left(-y_i F_{t-1}(x_i)\right).$$

AdaBoost minimizes exponential surrogate loss: Proof

Clearly, for any $\alpha_t > 0$, the f_t that minimizes $\mathbb{E}_n \exp(-Y F_t(X))$ minimizes $(*)$. And given f_t, we have

$$\frac{\partial}{\partial \alpha_t} \mathbb{E}_n \exp\left(-Y\left(F_{t-1}(X) + \alpha_t f_t(X)\right)\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i F_{t-1}(x_i)\right)\left(-y_i f_t(x_i)\right) \exp\left(-y_i \alpha_t f_t(x_i)\right)$$

$$= \sum_{i: y_i \neq f_t(x_i)} \underbrace{\left(\frac{1}{n} e^{-y_i F_{t-1}(x_i)}\right)}_{= D_t(i) \prod_{s=1}^{t-1} Z_s} e^{\alpha_t} - \sum_{i: y_i = f_t(x_i)} \underbrace{\left(\frac{1}{n} e^{-y_i F_{t-1}(x_i)}\right)}_{= D_t(i) \prod_{s=1}^{t-1} Z_s} e^{-\alpha_t}$$

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

AdaBoost minimizes exponential surrogate loss: Proof

Setting the derivative to zero and solving gives

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

So we can think of AdaBoost as choosing $F \in \operatorname{span}(\mathcal{G})$ to minimize

$$\mathbb{E}_n \exp\left(-Y F(X)\right),$$

but it does this in a stepwise way: with $F_{t-1} = \sum_{s=1}^{t-1} \alpha_s f_s$ fixed, choose $\alpha_t \in \mathbb{R}$ and $f_t \in \mathcal{G}$ to minimize

$$\mathbb{E}_n \exp\left(-Y \left(F_{t-1}(X) + \alpha_t f_t(X)\right)\right).$$