CS281B/Stat241B. Statistical Learning Theory. Lecture 26.

Peter Bartlett

Overview

- AdaBoost
  - Coordinate descent with other losses.
  - Dual problem: maximum entropy/I-projection.
  - AdaBoost is iterative projection method.
  - Weakly learnable $\Leftrightarrow$ infeasible.
  - Unnormalized KL projection.
  - Convergence of AdaBoost.

Boosting—coordinate descent—with other losses

We have seen that we can think of AdaBoost as choosing $F \in \mathrm{span}(\mathcal{G})$ to minimize $\mathbb{E}_n \exp\left(-YF(X)\right)$, in a greedy, 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

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

We can use similar ideas for loss functions other than $\phi(yF(x)) = \exp(-yF(x))$.

For example, logistic loss (LogitBoost) and quadratic loss (c.f. Tukey's "twicing"):

$$\phi_{logistic}(yf(x)) = \log(1 + \exp(-yf(x))),$$
$$\phi_2(yf(x)) = (1 - yf(x))^2.$$

Boosting—coordinate descent—with other losses

Consider the minimization of

$$J(F) = \mathbb{E}_n \phi(Y F_t(X)) = \mathbb{E}_n \phi(Y(F_{t-1}(X) + \alpha_t f_t(X)).$$

Fix F_{t-1} and consider gradient descent: choose a direction $v \in \mathbb{R}^n$ to minimize $v^T \nabla_v J(F_{t-1}(x_1^n) + v)$. We have

$$\frac{\partial}{\partial v_i} J(F_{t-1}(x_1^n) + v) = \frac{1}{n} \phi'(y_i F_{t-1}(x_i)) y_i,$$

so v should minimize

$$\sum_{i=1}^{n} v_i y_i \phi'\left(y_i F_{t-1}(x_i)\right) = \sum_{i=1}^{n} (-v_i y_i)\left(-\phi'\left(y_i F_{t-1}(x_i)\right)\right).$$

Boosting—coordinate descent—with other losses

If $v_i, y_i \in \{\pm 1\}$, this is equivalent to minimizing

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

where $D_t(i)$ is $-\phi'(y_i F_{t-1}(x_i))$, appropriately normalized. More generally (for instance, if the $f \in \mathcal{G}$ are real-valued), v should be chosen to maximize the inner product

$$\sum_{i=1}^{n} D_t(i) v_i y_i.$$

Boosting—coordinate descent—with other losses

$D_1(i) = \frac{1}{n}$, $i = 1, \ldots, n$.
$F_0(x) = 0$.
for $t = 1, \ldots, T$ **do**

Choose $f_t \in \mathcal{G}$ to minimize

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

Choose $\alpha_t \in \mathbb{R}$ to minimize

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

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

$$D_{t+1}(i) = \frac{-\phi'(y_i F_t(x_i))}{Z_t}.$$

end for

Dual problem: maximum entropy/I-projection

Consider the following minimization problem:

KL Minimization:

$$\begin{aligned}
\min_{p} \quad & D_{KL}(p, u) \\
\text{s.t.} \quad & \sum_{i=1}^{n} p_i y_i f(x_i) = 0 \qquad \text{for } f \in \mathcal{G}. \\
& p \geq 0 \\
& p^T 1 = 1,
\end{aligned}$$

where u is the uniform distribution, $u_i = 1/n$, and D_{KL} is the KL-divergence, $D_{KL}(p, u) = \sum_i p_i \ln(p_i / u_i)$.

Dual problem: maximum entropy/I-projection

Ignoring the positivity constraint (we'll see we get it for free), the Lagrangian is

$$L = \sum_{i=1}^{n} p_i \ln(np_i) + \sum_{f \in \mathcal{G}} \alpha_f \left(\sum_{i=1}^{n} p_i y_i f(x_i) \right) + \beta \left(p^T 1 - 1 \right).$$

$$\frac{\partial L}{\partial p_i} = \ln(np_i) + 1 + \sum_{f \in \mathcal{G}} \alpha_f y_i f(x_i) + \beta.$$

$$p_i = \exp\left(-(\ln n + 1 + \beta)\right) \exp\left(- \sum_{f \in \mathcal{G}} \alpha_f y_i f(x_i) \right).$$

$$= \frac{1}{Z} \exp\left(- \sum_{f \in \mathcal{G}} \alpha_f y_i f(x_i) \right). \qquad (\ln Z = \ln n + 1 + \beta.)$$

Dual problem: maximum entropy/I-projection

If we set $0 = \sum_i p_i \frac{\partial L}{\partial p_i} = L + p^T 1 + \beta$, we see that $L = -(\beta + 1) = \ln n - \ln Z$, so the dual problem is

Exponential Minimization:

$$\min_{\alpha} \qquad n \exp(-g(\alpha)) = Z = \sum_{i=1}^{n} \exp\left(-y_i \sum_{h \in \mathcal{G}} \alpha_h h(x_i)\right).$$

And this is the criterion that AdaBoost minimizes.

Iterative projection algorithm

Some notation: for $f \in \mathcal{G}$, define the constraint

$$C(f) = \left\{ p \in \Delta^n : \sum_{i=1}^{n} p_i y_i f(x_i) = 0 \right\}.$$

Recall the definition of the KL-projection,
$\Pi_S(p_t) := \arg\min_{p \in S} D_{KL}(p, p_t)$.

Iterative projection algorithm

$p_1 = u$.
for $t = 1, 2, \ldots, T$ **do**
Choose $f_t \in \mathcal{G}$ to maximize

$$D_{KL}\left(\Pi_{C(f_t)}(p_t), p_t\right).$$

Set $p_{t+1} = \Pi_{C(f_t)}(p_t)$.
end for

At each step, projects p_t onto a constraint $C(f_t)$.

AdaBoost is iterative projection algorithm

Theorem: At iteration t, this iterative projection algorithm chooses f_t so that it and α_t minimize

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

and the algorithm sets

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

i.e., it is AdaBoost.

AdaBoost is iterative projection algorithm: proof

For a fixed f_t, the Lagrangian of the KL-projection on to $C(f_t)$ is

$$L(p, \alpha, \mu) = \sum_i p_i \ln \frac{p_i}{p_{t,i}} + \alpha \sum_i p_i y_i f_t(x_i) + \mu \left(\sum_i p_i - 1 \right),$$

and setting $\partial L / \partial p_i = 0$ gives

$$\begin{aligned} p_{t+1,i} &= p_{t,i} \exp\left(-\alpha y_i f_t(x_i)\right) \exp(-1-\mu) \\ &= \frac{p_{t,i}}{Z_t} \exp\left(-\alpha y_i f_t(x_i)\right). \end{aligned}$$

Substituting into L shows that the dual problem is maximization of $g(\alpha, \mu) = -\ln Z_t$. And so the dual variable α_t is chosen to minimize Z_t.

AdaBoost is iterative projection algorithm: proof

Also,

$$
\begin{aligned}
D_{KL}(p_{t+1}, p_t) &= \sum_i p_{t+1,i} \ln \frac{p_{t+1,i}}{p_{t,i}} \\
&= \sum_i p_{t+1,i} \left(-\alpha_t y_i f_t(x_i) - \ln Z_t\right) \\
&= -\ln Z_t.
\end{aligned}
$$

So f_t is also chosen to minimize Z_t.

Iterative projection does not converge if weakly learnable

Theorem: If $\mathcal{G} = -\mathcal{G}$, the feasible set $\bigcap_{f \in \mathcal{G}} C(f)$ is empty iff there is a weak learner, that is, for some $\gamma > 0$, for all distributions p, there is an $f \in \mathcal{G}$ such that

$$\sum_i p_i 1[y_i \neq f(x_i)] \leq \frac{1}{2} - \gamma.$$

And notice that, if there is a γ-weak learner, then $Z_t \leq \sqrt{1 - 4\gamma^2}$, so

$$D_{KL}(p_{t+1}, p_t) = -\ln Z_t \geq \frac{1}{2} \ln \frac{1}{1 - 4\gamma^2},$$

so the iterative projection algorithm does not converge.

Iterative projection with unnormalized KL divergence

We can avoid this difficulty if we replace D_{KL} with D_{uKL}:

Unnormalized KL Minimization:

$$\begin{aligned} \min_{p} \quad & D_{uKL}(p, 1) \\ \text{s.t.} \quad & \sum_{i=1}^{n} p_i y_i f(x_i) = 0 \qquad \text{for } f \in \mathcal{G}, \\ & p \geq 0, \end{aligned}$$

where 1 is the all 1s vector, and D_{uKL} is the unnormalized KL-divergence,

$$D_{uKL}(p, q) = \sum_{i=1}^{n} \left(p_i \ln \left(\frac{pi}{q_i} \right) + q_i - p_i \right).$$

Iterative projection with unnormalized KL divergence

Again we can ignore the positivity constraint, and compute the Lagrangian:

$$L = \sum_{i=1}^{n} (p_i \ln(p_i) + 1 - p_i) + \sum_{f \in \mathcal{G}} \alpha_f \left(\sum_{i=1}^{n} p_i y_i f(x_i) \right).$$

$$\frac{\partial L}{\partial p_i} = \ln(p_i) + \sum_{f \in \mathcal{G}} \alpha_f y_i f(x_i).$$

$$p_i = \exp\left(-y_i \sum_{f \in \mathcal{G}} \alpha_f y_i f(x_i) \right).$$

Iterative projection with unnormalized KL divergence

We see that the dual problem is:

Exponential Minimization:

$$\min_{\alpha} \qquad n - g(\alpha) = \sum_{i=1}^{n} \exp\left(-y_i \sum_{f \in \mathcal{G}} \alpha_f f(x_i) \right).$$

Again, this is the criterion that AdaBoost minimizes.

Unnormalized iterative projection algorithm

$p_1 = 1$.

for $t = 1, 2, \ldots, T$ **do**

Choose $f_t \in \mathcal{G}$ to maximize

$$D_{uKL}\left(\Pi_{C(f_t)}(p_t), p_t\right).$$

Set $p_{t+1} = \Pi_{C(f_t)}(p_t)$.

end for

At each step, projects p_t onto a constraint $C(f_t)$.

The projection $\Pi_{C(f_t)}(p_t)$ is wrt D_{uKL}.

AdaBoost is unnormalized iterative projection algorithm

Theorem: At iteration t, the unnormalized iterative projection algorithm chooses f_t so that it and α_t minimize

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

and the algorithm sets

$$p_{t+1,i} = p_{t,i} \exp\left(-\alpha_t y_i f_t(x_i)\right).$$

i.e., it is (unnormalized) AdaBoost.

AdaBoost is iterative projection: Proof

Note that $p = 0$ shows that this problem is always feasible.

The proof of the theorem is similar to the normalized case: For a fixed f_t, the Lagrangian of the unnormalized KL-projection on to $C(f_t)$ is

$$L(p, \alpha) = \sum_i \left(p_i \ln \frac{p_i}{p_{t,i}} + p_{t,i} - p_i \right) + \alpha \sum_i p_i y_i f_t(x_i).$$

Setting $\partial L / \partial p(i) = 0$ gives

$$p_{t+1,i} = p_{t,i} \exp\left(-\alpha y_i f_t(x_i)\right).$$

AdaBoost is iterative projection: Proof

Substituting into L shows that the dual problem is maximization of

$$g(\alpha) = \sum_{i=1}^{n} (p_{t,i} - p_i) = (1 - Z) \sum_{i=1}^{n} p_{t,i},$$

where, as in the AdaBoost notation,

$$Z = \frac{\sum_{i=1}^{n} p_{t,i} \exp\left(-\alpha y_i f_t(x_i)\right)}{\sum_{i=1}^{n} p_{t,i}}.$$

Once again, the dual variable α_t is chosen to minimize Z_t.

AdaBoost is iterative projection: Proof

And since

$$
\begin{aligned}
D_{uKL}(p_{t+1}, p_t) &= -\alpha_t \underbrace{\sum_{i=1}^{n} p_{t+1,i} y_i f_t(x_i)}_{=0} + \sum_{i=1}^{n} (p_{t,i} - p_{t+1,i}) \\
&= (1 - Z) \sum_{i=1}^{n} p_{t,i},
\end{aligned}
$$

maximizing this quantity over f_t is equivalent to minimizing Z_t.

Convergence of AdaBoost

To understand the convergence of p_t, consider the two sets:

$$\mathcal{P} = \bigcap_{f \in \mathcal{G}} \mathcal{C}(f) = \bigcap_{f \in \mathcal{G}} \left\{ p \in \mathbb{R}^n : \sum_{i=1}^{n} p_i y_i f(x_i) = 0 \right\},$$

$$\mathcal{Q} = \left\{ p \in \mathbb{R}^n : p_i = \exp\left(-y_i \sum_{f \in \mathcal{G}} \lambda(f) f(x_i) \right), \lambda \in \mathbb{R}^{\mathcal{G}} \right\}$$

If the data is γ-weakly learnable, then the only feasible point is $p = 0$. And in that case, we've seen that the p_t converge to 0, but the direction of the p_t does not converge. What about other cases?