CS281B/Stat241B. Statistical Learning Theory. Lecture 27.

Peter Bartlett

Overview

- AdaBoost
  - Recall: unnormalized KL projection.
  - Convergence of AdaBoost.
- Model selection
  - Complexity penalization.
  - Oracle inequalities.
  - Universal consistency.
- Universal consistency of AdaBoost.

Recall: unnormalized KL projection

Consider the two sets:

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

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

and the dual optimization problems:

$$\begin{array}{ll} \min\limits_{p} & D_{uKL}(p, 1) \\ \text{s.t.} & p \in \mathcal{P}. \end{array}$$

$$\begin{array}{ll} \min\limits_{q} & D_{uKL}(0, q) \\ \text{s.t.} & q \in \mathcal{Q}. \end{array}$$

Recall: 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

Recall: AdaBoost is iterative projection

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.

Convergence of AdaBoost

Some notation:

Write the update step of the unnormalized iterative projection algorithm as A, so that at iteration t it sets

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

Notice that the exponential loss after iteration t is $p_{t+1}^T 1$:

$$\begin{aligned}
&\sum_{i=1}^{n} \exp\left(-y_i \sum_{s=1}^{t} \alpha_s f_s(x_i)\right) \\
&= \sum_{i=1}^{n} \prod_{s=1}^{t} \exp\left(-y_i \alpha_s f_s(x_i)\right) \\
&= \sum_{i=1}^{n} \prod_{s=1}^{t} \frac{p_{s+1,i}}{p_{s,i}} = \sum_{i=1}^{n} \frac{p_{t+1,i}}{p_{1,i}} = p_{t+1}^T 1.
\end{aligned}$$

Convergence of AdaBoost

Theorem: For the sequence $p_1, p_2, \ldots$ chosen by the algorithm,

1. $p_t \in \mathcal{Q}$.
2. The exponential loss is non-increasing: $(A(p_t) - p_t)^T 1 \leq 0$.
3. Since $p_t^T 1 \geq 0$, $(A(p_t) - p_t)^T 1 \to 0$.
4. If $(A(p_t) - p_t)^T 1 = 0$, then $p_t \in \mathcal{P}$.
5. The sequence $p_1, p_2, \ldots$ contains a limit point.
6. Since $(A(p_t) - p_t)^T 1$ is continuous, all limit points are in $\mathcal{P} \cap \overline{\mathcal{Q}}$.

Convergence of AdaBoost: Proof idea

1: $1 \in \mathcal{Q}$ and we have seen that the unnormalized KL-projections involve multiplication by exponentials, leaving $p_{t+1} \in \mathcal{Q}$.

2: We've seen that the algorithm is equivalent to AdaBoost, which greedily minimizes the exponential loss.

3: $p_t^T 1$ is bounded below and monotonically non-increasing, so it approaches a limit.

4: It's straightforward to show that

$$(A(p) - p)^T 1 = -\max_{f \in \mathcal{G}} \left(\sqrt{\sum_{i: y_i f(x_i) = 1} p_i} - \sqrt{\sum_{i: y_i f(x_i) = -1} p_i} \right)^2 .$$

And if this is zero, then for all $f \in \mathcal{G}$, the two terms are equal, which implies $p \in \mathcal{P}$.

Convergence of AdaBoost: Proof idea

5: Since the p_t come from a compact set ($0 \leq p_{t,i} \leq p_t^T 1 \leq p_1^T 1 = n$, so $p_t \in [0, n]^n$), they have a limit point.

6: It's clear that $(A(p) - p)^T 1$ is continuous, so any limit point p^* must have $(A(p^*) - p^*)^T 1 = 0$, so $p^* \in \mathcal{P}$. And $p_t \in \mathcal{Q}$, so $p^* \in \mathcal{P} \cap \overline{\mathcal{Q}}$.

Pythagorean Theorem

Lemma: If $p^* \in \mathcal{P} \cap \overline{\mathcal{Q}}$, then for all $p \in \mathcal{P}$ and $q \in \overline{\mathcal{Q}}$,

$$D_{uKL}(p, q) = D_{uKL}(p, p^*) + D_{uKL}(p^*, q).$$

Because $p \in \mathcal{P}$ and $q \in \overline{\mathcal{Q}}$, $\sum_i p_i \ln q_i = 0$. Hence,

$$D_{uKL}(p, p^*) = \sum_i (p_i \ln p_i + p_i^* - p_i),$$

$$D_{uKL}(p^*, q) = \sum_i (q_i - p_i^*),$$

$$D_{uKL}(p, q) = \sum_i (p_i \ln p_i + q_i - p_i),$$

which implies the result.

Convergence of AdaBoost

The Pythagorean theorem, applied to $1 \in \mathcal{Q}$ and $0 \in \mathcal{P}$, shows that there can be no more than one point in $\mathcal{P} \cap \overline{\mathcal{Q}}$, and it solves both optimization problems.

Theorem: Suppose there is a $p^* \in \mathcal{P} \cap \overline{\mathcal{Q}}$.

1. For any $p \in \mathcal{P}$ with $p \neq p^*$,

$$\begin{aligned} D_{uKL}(p, 1) &= D_{uKL}(p, p^*) + D_{uKL}(p^*, 1) \\ &> D_{uKL}(p^*, 1). \end{aligned}$$

2. For any $q \in \overline{\mathcal{Q}}$ with $q \neq p^*$,

$$\begin{aligned} D_{uKL}(0, q) &= D_{uKL}(0, p^*) + D_{uKL}(p^*, q) \\ &> D_{uKL}(0, p^*). \end{aligned}$$

Convergence of AdaBoost

But we've seen that the p_t sequence produced by the iterative projection method (i.e., AdaBoost) has at least one limit point, and all of its limit points are in $\mathcal{P} \cap \overline{\mathcal{Q}}$. So it must converge to the unique solution $p^* \in \mathcal{P} \cap \overline{\mathcal{Q}}$ of both optimization problems.

Model selection

Let's return to the problem of minimizing risk. We can decompose the excess risk (over the Bayes risk R^*) as

$$R(\hat{f}) - R^* = \underbrace{\left(R(\hat{f}) - \inf_{f \in F} R(f) \right)}_{\text{estimation error}} + \underbrace{\left(\inf_{f \in F} R(f) - R^* \right)}_{\text{approximation error}}$$

The first (second) term increases (decreases) with the complexity of the class F. **Model Selection** is the problem of automatically choosing the complexity to optimize this trade-off.

For instance, if we are combining classifiers (as AdaBoost does), as the size of the combination grows (measured perhaps in terms of the number of base classifiers, or perhaps in terms of the total weight of the combination), the complexity of the combination increases. How can we choose this complexity to minimize risk?

Complexity-penalized model selection

1. Define a complexity hierarchy $F_1 \subseteq F_2 \subseteq \cdots$.
2. Set $f_n^k = \arg\min_{f \in F_k} \hat{R}(f)$.
3. Choose $\hat{f} = \arg\min \hat{R}(f_n^k) + p_k(n)$, ($p_k(n)$ = *complexity penalty*).

Examples:

- Maximum *a posteriori* estimate: $\ell(\hat{p}, z) = -\log(\hat{p}(z))$, $p_k(n) = \log(1/\pi(k))/n$.
- Akaike Information Criterion: $p_k(n) = \dim(F_k)/n$.
- Structural risk minimization: $\ell = 0/1$ loss, $p_k(n) = \sqrt{d_{VC}(F_k)/n}$.
- Risannen's minimum description length: $p_k(n)$ =codelength.
- Error estimates: $p_k(n)$ = high confidence upper bound on $R(f_n^k) - \hat{R}(f_n^k)$, e.g., based on VCdim, Rademacher averages, etc.

Oracle inequalities

Theorem: Suppose that

$$P\left(R(f_n^k) > \hat{R}(f_n^k) + p_k(n) + \epsilon\right) \leq c_1 \exp(-c_2 n \epsilon^2).$$

If $\hat{f}$ is chosen to minimize $\hat{R}(f_n^k) + p_k(n) + \sqrt{\frac{\log k}{c_2 n}}$, then

$$P\left(R(\hat{f}) > \inf_k \left(\inf_{f \in F_k} R(f) + p_k(n) + 2\sqrt{\frac{\log k}{c_2 n}}\right) + \epsilon\right)$$
$$\leq (c_1 + 1) \exp(-c_2 n \epsilon^2/2).$$

Oracle inequalities

Notice that, for each k, the condition of the theorem ensures that with probability at least $1 - \delta$,

$$R(f_n^k) \leq \inf_{f \in F_i} R(f) + p_k(n) + c\sqrt{\frac{\log 1/\delta}{n}}.$$

So the theorem shows that $\hat{f}$ satisfies the best of these inequalities (with the addition of a $\sqrt{\log k/n}$ term). It is as if an oracle told us which class k would give the best performance guarantee, and we have (almost) the performance guarantee for that class.

The proof involves splitting the confidence δ across the classes ($\delta_k = \delta/k^2$), applying a union bound, and summing tail inequalities.

Universal consistency

We can rewrite the conclusion of the theorem in terms of excess risk:

$$R(\hat{f}) - R^* \leq \min_k \left(\inf_{f \in F_k} R(f) - R^* + p_k(n) + c\sqrt{\frac{\log k}{n}} \right) + c\sqrt{\frac{\log 1/\delta}{n}}.$$

As long as $\lim_{n\to\infty} p_i(n) = 0$, and

$$\lim_{i\to\infty} \inf_{f \in F_i} R(f) - R^* = 0,$$

then taking $n \to \infty$ ensures that $R(\hat{f}) \to R^*$.

This property is called *universal consistency*.

Universal consistency of AdaBoost

There are three difficulties in applying an analogous argument to AdaBoost:

1. AdaBoost minimizes exponential loss.
2. Concentration inequalities rely on boundedness, but F_T is unbounded.
3. AdaBoost does not have explicit complexity regularization. Is *early stopping* (restricting T) sufficient?

Universal consistency of AdaBoost: Definitions

1. $F_T = \sum_{t=1}^{T} \alpha_t f_t$.
2. Clipped version:

$$\pi_C(F_T(x)) = \begin{cases} C & \text{if } F_T(x) \geq C, \\ -C & \text{if } F_T(x) \leq -C, \\ F_T(x) & \text{otherwise.} \end{cases}$$

3. Optimal with norm B:

$$F_B^* = \arg \min_{\|F\| \leq B} R_\phi(F)$$

where $\|F\| = \sum_t |\alpha_t|$ for $F = \sum_t \alpha_t f_t$.

Universal consistency of AdaBoost: Key ideas

1. $R_\phi(\pi_C \circ F_T) \leq \hat{R}_\phi(\pi_C \circ F_T) + \epsilon_1(n, T, C, \delta)$.
   (Uniform law of large numbers for a T-combination.)
2. $\hat{R}_\phi(\pi_C \circ F_T) \leq \hat{R}_\phi(F_T) + \epsilon_2(C)$.
   (Approximation error of clipping for exponential loss.)
3. $\hat{R}_\phi(F_T) \leq \hat{R}_\phi(F_B^*) + \epsilon_3(B, T, n)$.
   (convergence rate)
4. $\hat{R}_\phi(F_B^*) \leq R_\phi(F_B^*) + \epsilon_4(B, n, \delta)$.
   (e.g., Hoeffding)
5. $R_\phi(F_B^*) \leq R_\phi^* + \epsilon_5(B)$.
   (For sufficiently rich $\mathcal{G}$, $\lim_{B\to\infty} \epsilon_5(B) = 0$)

Universal consistency of AdaBoost

Then:

$$\begin{aligned}\psi\left(R(F_T) - R^*\right) &\leq R_\phi(\pi_C \circ F_T) - R_\phi^* \\ &\leq \epsilon_1(n, T, C, \delta) + \epsilon_2(C) + \epsilon_3(B, T, n) + \\ &\quad + \epsilon_4(B, n, \delta) + \epsilon_5(B) + R_\phi^*.\end{aligned}$$

Allowing $T, n, C, B \to \infty$ at appropriate rates demonstrates universal consistency. The ϵ_1 term shows that we need $T = o(n)$ iterations: early stopping suffices for regularization.

Convergence rate: key ideas

For $L_t = \ln \dfrac{\hat{R}_\phi(F_t)}{\hat{R}_\phi(F_B^*)}$, $\quad S_t = |F_t| + |F_B^*|$, $\quad \gamma_t = \dfrac{1}{2} - \epsilon_t$, we have:

$$R_{t-1} \leq 2\gamma_t S_{t-1},$$

$$\frac{R_{t-1} - R_t}{S_t - S_{t-1}} \geq \gamma_t,$$

$$R_{t-1} - R_t \geq 2\gamma_t^2.$$

Hence,

$$\hat{R}_\phi(F_T) \leq \hat{R}_\phi(F_B^*) + 2\left(\frac{B^6}{T}\right)^{1/5}.$$