Ensemble - Boosting

김지수 (Jisu KIM)

통계적 기계학습(Statistical Machine Learning), 2024 1st semester

The lecture note is a minor modification of the lecture notes from Prof. Yongdai Kim's "Statistical Machine Learning", and Prof Larry Wasserman and Ryan Tibshirani's "Statistical Machine Learning". Also, see Section 10 from [1].

1 Boosting

Boosting is a method of combing weak learners (learners slightly better than random guess) to produce a strong committee. Freund and Schapire (1997) first proposed a practically usable boosting algorithm called "AdaBoost (Adaptive Boost)". Since then, many researches have been done to understand and extend boosting. A motivating example is the horse race. Suppose there are 100 gamblers who claim that they are expert in predicting the horse race results. Since they spend lots of time to study horse race, we can admit that their winning probabilities are slightly better than random guess (i.e. 50%). Now, the question is "Is it possible to combine 100 predictions of the 100 experts to make a better prediction?". Surprisingly, it is possible, which is called *weak learnability*. AdaBoot is the first algorithm to implement the idea of weak learnability. Let $Z_i = (X_i, Y_i)$ where $Y_i \in \{-1, +1\}$. We make the weak learning assumption: for some $\gamma > 0$ we have an algorithm returns $h \in \mathcal{H}$ such that, for all P,

$$P(R(h) \leq 1/2 - \gamma) \geq 1 - \delta$$

where $\gamma > 0$ is the edge.

AdaBoost algorithm is as follows:

1. Set $D_1(i) = 1/n$ for $i = 1, \ldots, n$.
2. Repeat for $t = 1, \ldots, T$:
   (a) Let $h_t = \text{argmin}_{h \in \mathcal{H}} P_{D_t}(Y_i \neq h(X_i))$.
   (b) $\epsilon_t = P_{D_t}(Y_i \neq h_t(X_i))$.
   (c) $\alpha_t = (1/2)\log((1-\epsilon_t)/\epsilon_t)$.
   (d) Let
   $$D_{t+1}(i) = \frac{D_t(i) e^{-Y_i \alpha_t h_t(X_i)}}{Z_t}$$
   where Z_t is a normalizing constant.
3. Set $g(x) = \sum_t \alpha_t h_t(x)$.
4. Return $h(x) = \text{sign} g(x)$.

AdaBoost increases the weights for misclassified observations and decreases the weights for correctly classified observations.

1.1 Empirical Results

Data Set	Single	AdaBoost	**Decrease**	Bagging
waveform	29.0	18.2	37%	19.4
breast cancer	6.0	3.2	47%	5.3
ionosphere	11.2	5.9	47%	8.6
diabetes	23.4	20.2	14%	18.8
glass	32.0	22.0	31%	24.9

AdaBoost outperforms the single best model. Moreover, AdaBoost is more accurate than Bagging in most cases (except **diabetes**).

1.2 Training Error

$h_t, t = 1, \ldots, T$, called *base learners*, are typically slightly better than random guess. In particular, Freund and Schapire (1997) showed that the training error converges to 0 exponentially fast if

$$\text{err}_t < 0.5 - \gamma$$

for some $\gamma > 0$. They conjectured that the overfitting emerges if T is too large since the model is too complex.

Lemma. *We have*

$$Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}.$$

Proof. Since $\sum_i D_t(i) = 1$ we have

$$\begin{aligned} Z_t &= \sum_i D_t(i) e^{-\alpha_t Y_i h_t(X_i)} = \sum_{Y_i h_t(X_i)=1} D_t(i) e^{-\alpha_t} + \sum_{Y_i h_t(X_i)=-1} D_t(i) e^{\alpha_t} \\ &= (1-\epsilon_t)e^{-\alpha_t} + \epsilon_t e^{\alpha_t} = 2\sqrt{\epsilon_t(1-\epsilon_t)}. \end{aligned}$$

since $\alpha_t = (1/2)\log((1-\epsilon_t)/\epsilon_t)$. □

Theorem. *Suppose that $\gamma \le (1/2) - \epsilon_t$ for all t. Then*

$$\hat{R}(h) \le e^{-2\gamma^2 T}.$$

Hence, the training error goes to 0 quickly.

Proof. Recall that $D_1(i) = 1/n$. So

$$\begin{aligned} D_{t+1}(i) &= \frac{D_t(i) e^{-\alpha_t Y_i h_t(X_i)}}{Z_t} = \frac{D_{t-1}(i) e^{-\alpha_{t-1} Y_i h_{t-1}(X_i)} e^{-\alpha_t Y_i h_t(X_i)}}{Z_t Z_{t-1}} \\ &= \cdots = \frac{e^{-Y_i \sum_t \alpha_t h_t(X_i)}}{n \prod_t Z_t} = \frac{e^{-Y_i g(X_i)}}{n \prod_t Z_t} \end{aligned}$$

which implies that

$$e^{-Y_i g(X_i)} = n D_{T+1}(i) \prod_t Z_t. \tag{1}$$

Since $I(u \le 0) \le e^{-u}$ we have

$$\begin{aligned} \hat{R}(h) &= \frac{1}{n}\sum_i I(Y_i g(X_i) \le 0) \le \frac{1}{n}\sum_i e^{-Y_i g(X_i)} = \frac{1}{n}\sum_i n(\prod_t Z_t) D_{T+1}(i) = \prod_{t=1}^{T} Z_t \\ &= \prod_t 2\sqrt{\epsilon_t(1-\epsilon_t)} = \prod_t \sqrt{1 - 4(1/2-\epsilon_t)^2} \\ &\le \prod_t e^{-2(1/2-\epsilon_t)^2} \qquad \text{since } 1 - x \le e^{-x} \\ &= e^{-2\sum_t (1/2-\epsilon_t)^2} \le e^{-2\gamma^2 T}. \end{aligned}$$

□

1.3 Generalization Error

The training error gets small very quickly. But how well do we do in terms of prediction error?

Let

$$\mathcal{F} = \left\{ \text{sign}(\sum_t \alpha_t h_t) : \ \alpha_t \in \mathbb{R}, \ h_t \in \mathcal{H} \right\}.$$

For fixed $h = (h_1, \ldots, h_T)$ this is just a set of linear classifiers which has VC dimension T. So the shattering number is

$$\left(\frac{en}{T}\right)^T.$$

If $\mathcal{H}$ is finite then the shattering number is

$$\left(\frac{en}{T}\right)^T . |\mathcal{H}|^T.$$

If $\mathcal{H}$ is infinite but has VC dimension d then the shattering number is bounded by

$$\left(\frac{en}{T}\right)^T \left(\frac{en}{d}\right)^{dT} \preceq n^{Td}.$$

By the VC theorem, with probability at least $1-\delta$,

$$R(\hat{h}) \leq \hat{R}(h) + \sqrt{\frac{Td \log n}{n}}.$$

Unfortunately this depends on T. We can fix this using margin theory.

Margins. Consider the classifier $h(x) = \text{sign}(g(x))$ where $g(x) = \sum_t \alpha_t h_t(x)$. The classifier is unchanged if we multiply g by a scalar. In particular, we can replace g with $\tilde{g} = g/||\alpha||_1$. This form of the classifier is a convex combination of the h_t's.

We define the *margin at* x of $g = \sum_t \alpha_t h_t$ by

$$\rho(x) = \frac{yg(x)}{||\alpha||_1} = y\tilde{g}(x).$$

Think of $|\rho(x)|$ as our confidence in classifying x. The margin of g is defined to be

$$\rho = \min_i \rho(X_i) = \min_i \frac{Y_i g(X_i)}{||\alpha||_1}.$$

Note that $\rho \in [-1, 1]$.

To proceed we need to review Radamacher complexity. Given a class of functions $\mathcal{F}$ with $-1 \leq f(x) \leq 1$ we define

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_\sigma \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_i \sigma_i f(Z_i) \right]$$

where $P(\sigma_i = 1) = P(\sigma_i = -1) = 1/2$. If $\mathcal{H}$ is finite then

$$\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{2 \log |\mathcal{H}|}{n}}.$$

If $\mathcal{H}$ has VC dimension d then

$$\mathcal{R}_n(\mathcal{H}) \leq \sqrt{\frac{2d \log(en/d)}{n}}.$$

We will need the following two facts. First,

$$\mathcal{R}_n(\text{conv}(\mathcal{H})) = \mathcal{R}_n(\mathcal{H})$$

where $\text{conv}(\mathcal{H})$ is the convex hull of $\mathcal{H}$. Second, if

$$|\phi(x) - \phi(y)| \leq L||x - y||$$

for all x, y then

$$\mathcal{R}_n(\phi \circ \mathcal{F}) \leq L\mathcal{R}_n(\mathcal{F}).$$

The set of margin functions is

$$\mathcal{M} = \{yf(x) : \ f \in \text{conv}(\mathcal{H})\}.$$

We then have

$$\mathcal{R}_n(\mathcal{M}) = \mathcal{R}_n(\text{conv}(\mathcal{H})) = \mathcal{R}_n(\mathcal{H}).$$

A key result is that, with probability at least $1-\delta$, for all $f \in \mathcal{F}$,

$$\mathbb{E}[f(Z)] \leq \frac{1}{n}\sum_i f(Z_i) + 2\mathcal{R}_n(\mathcal{F}) + \sqrt{\frac{2\log(1/\delta)}{n}}. \tag{2}$$

Now fix a number ρ and define the margin-sensitive loss function

$$\phi(u) = \begin{cases} 1 & u \leq 0 \\ 1 - \frac{u}{\rho} & 0 \leq \rho \\ 0 & u \geq \rho. \end{cases}$$

Note that

$$I(u \leq 0) \leq \phi(u) \leq I(u \leq \rho).$$

Assume that $\mathcal{H}$ has VC dimension d. Then

$$\mathcal{R}_n(\phi \circ \mathcal{M}) \leq L\mathcal{R}_n(\mathcal{M}) \leq L\mathcal{R}_n(\mathcal{H}) \leq \frac{1}{\rho}\sqrt{\frac{2d\log(en/d)}{n}}.$$

Now define the empirical margin sensitive loss of a classifer f by

$$\hat{R}_\rho = \frac{1}{n}\sum_i I(Y_i f(X_i) \leq \rho).$$

Theorem. *With probability at least* $1-\delta$,

$$R(g) \leq \hat{R}_\rho(g/||\alpha||_1) \leq \frac{1}{\rho}\sqrt{\frac{2d\log(en/d)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Proof. Recall that $I(u \leq 0) \leq \phi(u) \leq I(u \leq \rho)$. Also recall that g and $\tilde{g} = g/||\alpha||_1$ are equivalent classifiers. Then using (2) we have

$$\begin{aligned} R(g) = R(\tilde{g}) = P(Y\tilde{g}(X) \leq 0) &\leq \frac{1}{n}\sum_i \phi(Y_i\tilde{g}(X_i)) + 2\mathcal{R}_n(\phi\circ\mathcal{M}) + \sqrt{\frac{2\log(2/\delta)}{n}} \\ &\leq \frac{1}{n}\sum_i \phi(Y_i\tilde{g}(X_i)) + \frac{1}{\rho}\sqrt{\frac{2d\log(en/d)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}} \\ &= \hat{R}_\rho(g/||\alpha||_1) + \frac{1}{\rho}\sqrt{\frac{2d\log(en/d)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}. \end{aligned}$$

□

Next we bound $\hat{R}_\rho(g/||\alpha||_1)$.

Theorem. *We have*

$$\hat{R}_\rho(g/||\alpha||_1) \leq \prod_{t=1}^{T}\sqrt{4\epsilon_t^{1-\rho}(1-\epsilon_t)^{1+\rho}}.$$

Proof. Since $\phi(u) \leq I(u \leq \rho)$ we have

$$
\begin{aligned}
\hat{R}_\rho(g/||\alpha||_1) &\leq \frac{1}{n}\sum_i I(Y_i g(X_i) - \rho||\alpha||_1 \leq 0) \\
&\leq e^{\rho||\alpha||_1}\frac{1}{n}\sum_i e^{-Y_i g(X_i)} \\
&= e^{\rho||\alpha||_1}\frac{1}{n}\sum_i nD_{T+1}(i)\prod_t Z_t = e^{\rho||\alpha||_1}\prod_t Z_t \\
&= \prod_{t=1}^{T}\sqrt{4\epsilon_t^{1-\rho}(1-\epsilon_t)^{1+\rho}}
\end{aligned}
$$

since $Z_t = 2\sqrt{\epsilon_t(1-\epsilon_t)}$ and $\alpha_t = (1/2)\log((1-\epsilon_t)/\epsilon_t)$. □

Assuming $\gamma \leq (1/2 - \epsilon_t)$ and $\rho < \gamma$ then it can be shown that $\sqrt{4\epsilon_t^{1-\rho}(1-\epsilon_t)^{1+\rho}} \equiv b < 1$. So $\hat{R}_\rho(g/||\alpha||_1) \leq b^T$. Combining with the previous result we have, with probability at least $1-\delta$,

$$
R(g) \leq b^T + \frac{1}{\rho}\sqrt{\frac{2d\log(en/d)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}.
$$

This shows that we get small error even with T large (unlike the earlier bound based only on VC theory).

References

[1] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning.* Springer Series in Statistics. Springer, New York, second edition, 2009. Data mining, inference, and prediction.