Boosting the Margin

Oleg Ivanov
Min Chi

Ensemble methods

• Bagging reduces the expected error by decreasing the variance

• Boosting reduces the expected error by decreasing the variance (can be increased (?)) and the bias

• Train errors can be driven to 0

• But the test (generalization) errors do not show overfitting (paradox ?)
Margin

- Ensemble classifier prediction = weighted vote over a set of base classifiers

- Margin – difference between the weight assigned to the correct label and the maximal weight assigned to any incorrect label

- Margin distribution graphs – margin for a set of examples

- Boosting and bagging increase the margins

- Boosting is especially useful for examples with small initial margins

Margin & test error

- Prove that achieving a large margin on the training set results in an improved bound on the generalization (test) error

- The bound does not depend on the number of classifiers that are combined in the vote

- The bound is not asymptotic
Definitions

- Output: {-1, 1}
- \( H \) the space of base classifiers
- \( h \in H \) mapping from an instance space \( X \) to {-1, +1}
- The examples are generated independently at random according to some fixed but unknown distribution \( D \) over \( X \times \{-1, +1\} \)
- The training data is a list of \( m \) pairs:
  \[
  S = \langle (x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m) \rangle \text{ chosen according to } D
  \]

Definitions

- \( P_{(x, y) \sim D}[A] \): probability of event \( A \) when the example \((x, y)\) is chosen according to \( D \). (Abbreviated by \( P_D[A] \)).
- \( P_{(x, y) \sim S}[A] \): probability with respect to choosing an example uniformly at random from the training set. (Abbreviated by \( P_S[A] \)).
Definitions

- \( C = \left\{ f : x \mapsto \sum_{h \in H} a_h h(x) \mid a_h \geq 0; \sum_{h \in H} a_h = 1 \right\} \) convex hull \( C \) of \( H \) as the set of mapping that can be generated by taking a weighted average of classifier from \( H \)
- A classifier \( f \) in \( C \) predicts label \( y \) for input \( x \) if \( f(x,y) > \max_{y' \neq y} f(x,y') \)
- Margin \( (f,x,y) = f(x,y) - \max_{y' \neq y} f(x,y') \)
- \( f \) gives the wrong prediction on \( (x,y) \) only if margin \( (f,x,y) \leq 0 \)

Definitions

- The majority vote rule that is associated with \( f \) gives the wrong prediction on the example \( (x,y) \) only if \( yf(x) \leq 0 \)
- The margin of an example \( (x,y) \) in this case is simply \( yf(x) \)
- The following two theorems state that with high probability, the generalization error of any majority vote classifier can be bounded in terms of the number of training examples with margin below a threshold \( \theta \), plus an additional term which depends on the number of training examples, some complexity measure of \( H \), and the threshold \( \theta \)
**Theorem 1 (finite classifier space)**

- Let $D$ be a distribution over $X \times \{-1, +1\}$, and let $S$ be a sample of $m$ examples chosen independently at random according to $D$. Assume that the base-classifier space $H$ is finite, and let $\delta > 0$. Then with probability at least $1 - \delta$ over the random choice of the training set $S$, every weighted average function $f \in C$ satisfies the following bound for all $\theta > 0$:

$$P_D[yf(x) \leq 0] \leq P_S[yf(x) \leq \theta] + O\left(\frac{1}{\sqrt{m}} \left(\frac{\log m \log |H|}{\theta^2} + \log(1/\delta)\right)^{1/2}\right)$$

**Theorem 1 - proof**

- Define $C_N$ to be the set of unweighted averages over $N$ elements from $H$:

$$C_N = \left\{ f : x \mapsto \frac{1}{N} \sum_{i=1}^{N} h_i(x) \mid h_i \in H \right\}$$

(approximating set)

- Any majority vote classifier $f \in C$ can be associated with a distribution over $H$ as defined by the coefficients $a_h$.

- By choosing $N$ elements of $H$ independently at random according to this distribution we can generate an element of $C_N$. 


Theorem 1 - proof

- Define a function of $g \in C_N$ distributed according to $Q$ and selected by choosing $h_1, \ldots, h_N$ independently at random according to the coefficients $a_h$

- $g(x) = (1/N) \sum_{i=1}^N h_i(x)$

\[
\begin{align*}
    P_D[yf(x) \leq 0] &= P_D[yg(x) \leq \theta/2, yf(x) \leq 0] + P_D[yg(x) > \theta/2, yf(x) \leq 0] \\
    &\leq P_D[yg(x) \leq \theta/2] + P_D[yg(x) > \theta/2, yf(x) \leq 0]
\end{align*}
\] (1)
Theorem 1 - proof

• Equation (1) holds for any $g \in C$, we can take the expected value of the right hand side wrt $Q$ and get:

$$P_D[yf(x) \leq 0]$$

$\leq P_{D,g-Q}[yg(x) \leq \theta / 2] + P_{D,g-Q}[yg(x) > \theta / 2, yf(x) \leq 0]$  

$= E_{g-Q}[P_D[yg(x) \leq \theta / 2]] + E_{D}[P_{g-Q}[yg(x) > \theta / 2, yf(x) \leq 0]]$  

$\leq E_{g-Q}[P_D[yg(x) \leq \theta / 2]] + \frac{A}{P(A \cap B) = P(A|B) \cdot P(B) \leq P(A|B) \cdot B}$

(2)

Chernoff bound

• Let $X_1, \ldots, X_n$ be discrete, independent random variables such that $E[X_i] = 0$ and $|X_i| \leq 1$ for all $i$.

Let $X = \sum_{i=1}^{N} X_i$ and $\sigma^2$ be the variance of $X$. Then

$$\Pr[|X| \geq \lambda \sigma] \leq 2e^{-\lambda^2/4}$$

for any $0 \leq \lambda \leq 2\sigma$
Theorem 1 - proof

- Both terms are bounded separately, starting with $B$

- The Chernoff bound for $B$:

$$P_{g-Q}[yg(x) > \theta/2 | yf(x) \leq 0] \leq e^{-N\theta^2/8} \quad (3)$$

Theorem 1 - proof

- The probability over the choice of $S$ that there exists any $g \in C_N$ and $\theta > 0$ for which (union bound):

$$P_{g} [yg(x) \leq \theta/2] > P_{g} [yg(x) \leq \theta/2] + \varepsilon_N$$

is at most $(N + 1)|C_N|e^{-2m\xi_N}$

- The $e^{-2m\xi_N}$ comes from the Chernoff bound which holds for any single choice of $g$ and $\theta$

- The term $(N + 1)|C_N|$ is an upper bound on the number of such choices where we have used the fact that, because of the form of functions in $C_N$, we need only consider values of $\theta$ of the form $2i/N$ for $i = 0, ..., N$.  


Theorem 1 - proof

Thus, if we set $e_N = \sqrt{(1/2m) \ln ((N+1)H)^N / \delta}$, and take expectation with respect to $Q$, we get that, with probability at least 1 - $\delta_n$

$$P_{D,S-Q}[yg(x) \leq \theta/2] \leq P_{S,S-Q}[yg(x) \leq \theta/2, yg(x) > \theta] + e_N$$

$$P_{S,S-Q}[yg(x) \leq \theta/2] \leq P_{S,S-Q}[yf(x) \leq \theta] + P_{S,S-Q}[yg(x) \leq \theta/2, yf(x) > \theta]$$

$$= P_S[yf(x) \leq \theta] + E_S[P_{S,S-Q}[yg(x) \leq \theta/2, yf(x) > \theta]]$$

$$\leq P_S[yf(x) \leq \theta] + E_S[P_{S,S-Q}[yg(x) \leq \theta/2, yf(x) > \theta]]$$

$$\leq \sum_{N=1} \delta_n = \sum_{N=1} \delta / N(N+1)$$

$$= \delta / 1 - \delta / 2 + \delta / 2 - \delta / 3 + \delta / 3 - \delta / 4 + ...$$

$$= \delta$$

To bound $E$, use Chernoff bound:

$$E_S[P_{S,S-Q}[yg(x) \leq \theta/2, yf(x) > \theta]] \leq e^{-N\theta^2/8}$$

Let $\delta_N = \delta / N(N+1)$, so the probability of failure for any $N$ will be at most:

$$\sum_{N=1} \delta_n = \sum_{N=1} \delta / N(N+1)$$

$$= \delta$$
Then combining Equations (2), (3), (4), (5) and (6), we get that, with probability at least $1-\delta$, for every $\theta>0$ and every $N \geq 1$:

$$P_{D}[yf(x) \leq 0] \leq P_{S}[yf(x) \leq \theta] + 2e^{-N\theta^2/8} + \frac{1}{2m} \ln \left( \frac{N(N + 1)^2|H|^N}{\delta} \right)$$  \hspace{1cm} (7)

And $N = \left\lceil (4l/\theta^2) \ln(m / \ln(|H|)) \right\rceil$
Theorem 1 - proof

\[ \text{If } \delta > 0 \text{ and } \theta > 0 \text{ are held fixed as } m \to \infty, \text{ then the bound converges to} \]

\[ P_s[\lfloor yf(x) \rfloor \leq \theta] \leq P_s[yf(x) \leq \theta] + \frac{2}{\sqrt{m / \ln(|H|)}} + \frac{1}{\sqrt{2m}} \left(3 \ln (\ln (m / \ln(|H|))) + (4/\theta^2) \ln (m / \ln(|H|)) \right)^{1/2} \]

\[ = P_s[yf(x) \leq \theta] + \frac{2}{\sqrt{m / \ln(|H|)}} + \frac{1}{\sqrt{2m}} \left(3 \ln (\ln (m / \ln(|H|))) + (4/\theta^2) \ln (m / \ln(|H|)) \right)^{1/2} \]

\[ \text{In fact, if } \theta \leq 1/2, \text{ then } F \text{ is a pretty good approximation of the second and third terms on the right-hand side of Equation (7)} \]

Theorem 1 - discussion

- If \( \delta > 0 \) and \( \theta > 0 \) are held fixed as \( m \to \infty \), then the bound converges to

\[ P_s[\lfloor yf(x) \rfloor \leq 0] \leq P_s[yf(x) \leq \theta] + \frac{2 \log m}{m \theta^2} + o\left(\frac{\ln m}{m}\right) \]

and \( m \geq 1000 \) then \( F \) is a pretty good approximation of the second and third terms on the right-hand side of Equation (7)

- Even though not asymptotic, the bounds are loose

i.e. start to be meaningful only with the large training set (10,000s)
Theorem 1 - discussion

Figure: A few plots of the second and third terms in the bound given in Equation (8) (solid lines) and their approximation by the second term in Equation (9) (dotted lines). The horizontal axis denotes the number of training examples (with a logarithmic scale) and the vertical axis denotes the value of the bound. All plots are for $\delta = 0.01$ and $|\mathcal{H}| = 10^5$. Each pair of close lines corresponds to a different value of $\theta$, counting the pairs from the upper right to the lower left, the values of $\theta$ are $1/20, 1/5, 1/4$ and $1/2$.

Theorem 2 (infinite classifier space)

Theorem 2
Let $D$ be a distribution over $X \times \{-1,+1\}$, and let $S$ be a sample of $m$ examples chosen independently at random according to $D$. Suppose the base-classifier space $H$ has VC-dimension $d$, and let $\delta > 0$. Assume that $m \geq d \geq 1$. Then with probability at least $1 - \delta$ over the random choice of the training set $S$, every weighted average function $f \in C$ satisfies the following bound for all $\theta > 0$:

$$
P_{H}[\{f(x) \leq 0\}] \leq P_{S}[\{f(x) \leq \theta\}] + O\left(\frac{d \log^2(m/d)}{\theta^2} + \log(1/\delta)\right)^{1/2}
$$
AdaBoost and Margin Distribution

- AdaBoost - rerunning a base learning algorithm, each time using a different distribution over training examples.
- The goal - to find a classifier $h_t$ with small error $e_t = P_{x \sim D_t} [y \neq h_t(x)]$.
- On each round $t = 1,...,T$ a distribution $D_t$ is computed over the training examples:
  $$D_{t+1}(i) = \frac{D_t(i) \exp(-y_i \alpha_t h_t(x_i))}{Z_t},$$

  where $\alpha_t = \frac{1}{2} \ln((1-e_t)/e_t)$

  $Z_t$ is a normalization factor $D_{t+1}(i)$'s should add up to 1.

---

AdaBoost and Margin Distribution

- Combined classifier - a weighted majority vote of the base classifiers:
  $$f(x) = \frac{\sum_{t=1}^{T} \alpha_t h_t(x)}{\sum_{t=1}^{T} \alpha_t}$$

- On round $t$, AdaBoost places the most weight on examples $(x, y)$ for which
  $$y \sum_{t=1}^{T} \alpha_t h_t(x)$$

is the smallest (margin of the combined classifier).
**AdaBoost and Margin Distribution**

Theorem 5

Suppose the base learning algorithm, when called by AdaBoost, generates classifiers with weighted training errors $\varepsilon_1, \ldots, \varepsilon_T$. Then for any $\theta$

$$P_{(x,y) \sim \mathcal{S}}[yf(x) \leq \theta] \leq 2^T \prod_{t=1}^{T} \sqrt{\varepsilon_t^{1-\theta}(1 - \varepsilon_t)^{1+\theta}}$$

$\sqrt{\varepsilon_t^{1-\theta}(1 - \varepsilon_t)^{1+\theta}} < 1.0$, therefore

$P_{(x,y) \sim \mathcal{S}}[yf(x) \leq \theta]$ decreases exponentially fast with $T$. 