Online Learning, Games, Boosting

Arindam Banerjee
Predicting with Expert Advice

- Basic setting: $n$ experts, boolean prediction

- Prediction proceeds in trials
  - Each expert makes a boolean prediction
  - “Master algorithm” combines the predictions
  - The true label is revealed

- No explicit notion of input

- Master algorithms with “good” cumulative/relative loss
The Halving Algorithm

If one of the experts is always known to be correct:

- Always vote based on the majority
- If the master is wrong, get rid of the experts that were wrong

Proposition: The master makes at most $\log_2 n$ mistakes
Weighted Majority (Simple)

- Initialize the weights $w_i = 1$, $[i]_1^n$
- Each expert makes a boolean prediction $z_i \in \{0, 1\}$, $[i]_1^n$
- The master predicts 1 if
  $$\sum_{i: z_i = 1} w_i \geq \sum_{i: z_i = 0} w_i,$$
  and 0 otherwise
- The true label $\ell$ is received
- Modify the weights as follows: if $z_i = \ell$, no change; if $z_i \neq \ell$, $w_i \leftarrow w_i / 2$, $[i]_1^n$

Theorem: The master makes at most $2.41(m + \log_2 n)$ mistakes, where $m$ is the number of mistakes by the best expert so far
Weighted Majority (Randomized)

- Initialize the weights $w_i = 1$, $[i]_n^1$
- Each expert makes a boolean prediction $z_i \in \{0, 1\}$, $[i]_1^n$
- The master predicts $z_i$ with probability $w_i/W$, where $W = \sum_i w_i$
- The true label $\ell$ is received
- Modify the weights as follows: if $z_i = \ell$, no change; if $z_i \neq \ell$, $w_i \leftarrow \beta w_i$, $[i]_1^n$ for $\beta \in [0, 1]$

**Theorem:** The expected number of mistakes the master makes is at most $m \ln(1/\beta) + \ln n \over 1-\beta$, where $m$ is the number of mistakes by the best expert so far

**Corollary:** If $\beta$ is adjusted dynamically, or, if the total number of trials is known upfront, the expected number of mistakes is at most $m + \ln n + O(\sqrt{m \ln n})$
Extensions

- The range of prediction is more general, i.e., not necessarily boolean
- An explicit notion of input exists
- The adversary is not arbitrary, e.g., choosing labels from a function, a fixed distribution, etc.
- Labels from a function with random noise

Learning (boolean) functions
- Realizable learning
- Agnostic learning
- Random Noise
The loss matrix $M$ for the row player

The actual game has payoffs $2(M - \frac{1}{2})$

Pure strategies: Rock, Paper, Scissors

Randomized play
- Row player chooses distribution $P$
- Column player chooses distribution $Q$
- The expected loss for the row player

$$M(P, Q) = P^T M Q = \sum_{ij} P(i)Q(j)M(i, j)$$
Sequential Play

- Row player chooses a (randomized) strategy $P$
- Given $P$, column player chooses a strategy $Q$ such that the loss of row player is maximized, i.e., $\max_Q M(P, Q)$
- Row player chooses $P$ upfront to minimize the maximum loss

$$\min_P \max_Q M(P, Q)$$

- If column player chooses first, then the loss of the row player

$$\max_Q \min_P M(P, Q)$$

- “Obviously” $\max_Q \min_P M(P, Q) \leq \min_P \max_Q M(P, Q)$
- Von Neumann says the following:

$$\max_Q \min_P M(P, Q) = \min_P \max_Q M(P, Q)$$
Repeated Play

- Row is the learner, Column is the adversary
- The game is being played repeatedly \( t = 1, \ldots, T \)
  - On round \( t \), learner chooses \( P_t \)
  - Adversary chooses \( Q_t \)
  - Learner gets losses \( M(i, Q_t) \) for individual strategies
  - Learner suffers loss \( M(P_t, Q_t) \)

- Learner’s algorithm (Hedging)
  - Initialize the weights \( w_1(i) = 1, [i]^n_1 \)
  - At round \( t \), choose
    \[
    P_t(i) = \frac{w_t(i)}{\sum_i w_t(i)}
    \]
  - Upon receiving loss \( M(i, Q_t) \), update
    \[
    w_{t+1}(i) = w_t(i) \beta^{M(i, Q_t)}
    \]
Main Results

Theorem: The sequence of strategies $P_1, \ldots, P_T$, chosen by the algorithm satisfies

$$\sum_{t=1}^{T} M(P_t, Q_t) \leq a_\beta \min_P \sum_{t=1}^{T} M(P, Q_t) + c_\beta \ln n$$

where $a_\beta = \frac{\ln(1/\beta)}{1-\beta}$ and $c_\beta = \frac{1}{1-\beta}$

Corollary: With $\beta = \frac{1}{1+\sqrt{2 \ln n} T}$, the average per-trial loss satisfies

$$\frac{1}{T} \sum_{t=1}^{T} M(P_t, Q_t) \leq \min_P \frac{1}{T} \sum_{t=1}^{T} M(P, Q_t) + \Delta_T$$

where $\Delta_T = O\left(\sqrt{\frac{\ln n}{T}}\right)$
Corollary: If $v$ is the value of the game

$$
\frac{1}{T} \sum_{t=1}^{T} M(P_t, Q_t) \leq v + \Delta_T
$$

A (new) proof of the minimax theorem

$$
\min_P \max_Q P^T M Q \leq \max_Q \frac{1}{T} \sum_t P_t^T M Q \\
\leq \frac{1}{T} \sum_t \max_Q P_t^T M Q = \frac{1}{T} \sum_t P_t^T M Q_t \\
\leq \min_P \frac{1}{T} \sum_t P^T M Q_t + \Delta_t \\
\leq \max_Q \min_P P^T M Q + \Delta_t
$$
Weak Learning

- A weak learner predicts slightly better than random

The PAC setting (basics)
- Let $\mathcal{X}$ be a instance space, $c: \mathcal{X} \mapsto \{0, 1\}$ be a target concept, $\mathcal{H}$ be a hypothesis space ($h: \mathcal{X} \mapsto \{0, 1\}$)
- Let $Q$ be a fixed but unknown distribution on $\mathcal{X}$
- An algorithm, after training on $(x_i, c(x_i)), [i]_1^m$, selects $h \in \mathcal{H}$ such that
  \[
P_{x \sim Q}[h(x) \neq c(x)] \leq \frac{1}{2} - \gamma
\]
- The algorithm is called a $\gamma$-weak learner
- We assume the existence of such a learner
The Boosting Model

- Boosting converts a weak learner to a strong learner
- Boosting proceeds in rounds
  - Booster constructs $D_t$ on $X$, the train set
  - Weak learner produces a hypothesis $h_t \in \mathcal{H}$ so that
    \[ P_{x \sim D_t}[h_t(x) \neq c(x)] \leq \frac{1}{2} - \gamma \]
  - After $T$ rounds, the weak hypotheses $h_t, [t]^T$ are combined into a final hypothesis $h_{\text{fin}}$
- We need procedures
  - for obtaining $D_t$ at each step
  - for combining the weak hypotheses
Consider the mistake matrix $M(h, x) = 1$ if $h(x) \neq c(x)$ and 0 otherwise.

Therefore, $v \cdot \frac{1}{2} < \frac{1}{2}$.

Majority voting according to $P$ is equivalent to $c$. – p.14
Minimax and Majority Votes

- Consider the mistake matrix $M(h, x) = 1$ if $h(x) \neq c(x)$ and 0 otherwise.
- The row player can choose $P$ over $\mathcal{H}$. 

 Majority voting according to $P$ is equivalent to $c$. - p.14
Consider the mistake matrix $M(h, x) = 1$ if $h(x) \neq c(x)$ and $0$ otherwise.

The row player can choose $P$ over $\mathcal{H}$.

The column player can choose $Q$ over $\mathcal{X}$.
Minimax and Majority Votes

- Consider the mistake matrix $M(h, x) = 1$ if $h(x) \neq c(x)$ and 0 otherwise.
- The row player can choose $P$ over $\mathcal{H}$.
- The column player can choose $Q$ over $\mathcal{X}$.
- Now $\min_P \max_x M(P, x) = v = \max_Q \min_h M(h, Q)$. 

Consider the mistake matrix $M(h, x) = 1$ if $h(x) \neq c(x)$ and 0 otherwise.

The row player can choose $P$ over $\mathcal{H}$.

The column player can choose $Q$ over $\mathcal{X}$.

Now $\min_P \max_x M(P, x) = v = \max_Q \min_h M(h, Q)$.

However $M(h, Q) = P_{x \sim Q}[h(x) \neq c(x)]$. 

Consider the mistake matrix \( M(h, x) = 1 \) if \( h(x) \neq c(x) \) and 0 otherwise.

The row player can choose \( P \) over \( \mathcal{H} \).

The column player can choose \( Q \) over \( \mathcal{X} \).

Now \( \min_P \max_x M(P, x) = v = \max_Q \min_h M(h, Q) \).

However \( M(h, Q) = P_{x \sim Q}[h(x) \neq c(x)] \).

Now \( \exists Q^*, M(h, Q^*) = P_{x \sim Q^*}[h(x) \neq c(x)] \geq v, \forall h \).
Consider the mistake matrix \( M(h, x) = 1 \) if \( h(x) \neq c(x) \) and 0 otherwise.

The row player can choose \( P \) over \( \mathcal{H} \).

The column player can choose \( Q \) over \( \mathcal{X} \).

Now \( \min_P \max_x M(P, x) = v = \max_Q \min_h M(h, Q) \).

However \( M(h, Q) = P_{x \sim Q}[h(x) \neq c(x)] \).

Now \( \exists Q^*, M(h, Q^*) = P_{x \sim Q^*}[h(x) \neq c(x)] \geq v, \forall h \).

From weak learning, \( \exists h, P_{x \sim Q^*}[h(x) \neq c(x)] \leq \frac{1}{2} - \gamma \).
Consider the mistake matrix $M(h, x) = 1$ if $h(x) \neq c(x)$ and 0 otherwise

The row player can choose $P$ over $\mathcal{H}$

The column player can choose $Q$ over $\mathcal{X}$

Now $\min_P \max_x M(P, x) = v = \max_Q \min_h M(h, Q)$

However $M(h, Q) = P \sim Q[h(x) \neq c(x)]$

Now $\exists Q^*, M(h, Q^*) = P \sim Q^*[h(x) \neq c(x)] \geq v, \forall h$

From weak learning, $\exists h, P \sim Q^*[h(x) \neq c(x)] \leq \frac{1}{2} - \gamma$

Hence $v \leq \frac{1}{2} - \gamma$
Consider the mistake matrix $M(h, x) = 1$ if $h(x) \not= c(x)$ and 0 otherwise.

The row player can choose $P$ over $\mathcal{H}$

The column player can choose $Q$ over $\mathcal{X}$

Now $\min_P \max_x M(P, x) = v = \max_Q \min_h M(h, Q)$

However $M(h, Q) = \mathbb{P}_{x \sim Q}[h(x) \not= c(x)]$

Now $\exists Q^*, M(h, Q^*) = \mathbb{P}_{x \sim Q^*}[h(x) \not= c(x)] \geq v$, $\forall h$

From weak learning, $\exists h$, $\mathbb{P}_{x \sim Q^*}[h(x) \not= c(x)] \leq \frac{1}{2} - \gamma$

Hence $v \leq \frac{1}{2} - \gamma$

Also $\exists P^*$ such that $\forall x$, 

$$M(P^*, x) = \mathbb{P}_{h \sim P^*}[h(x) \not= c(x)] \leq v \leq \frac{1}{2} - \gamma < \frac{1}{2}$$
Minimax and Majority Votes

Consider the mistake matrix \( M(h, x) = 1 \) if \( h(x) \neq c(x) \) and 0 otherwise.

The row player can choose \( P \) over \( \mathcal{H} \).

The column player can choose \( Q \) over \( \mathcal{X} \).

Now \( \min_P \max_x M(P, x) = v = \max_Q \min_h M(h, Q) \).

However \( M(h, Q) = P_{x \sim Q}[h(x) \neq c(x)] \).

Now \( \exists Q^*, M(h, Q^*) = P_{x \sim Q^*}[h(x) \neq c(x)] \geq v, \forall h \).

From weak learning, \( \exists h, P_{x \sim Q^*}[h(x) \neq c(x)] \leq \frac{1}{2} - \gamma \).

Hence \( v \leq \frac{1}{2} - \gamma \).

Also \( \exists P^* \) such that \( \forall x, \)

\[
M(P^*, x) = P_{h \sim P^*}[h(x) \neq c(x)] \leq v \leq \frac{1}{2} - \gamma < \frac{1}{2}
\]

Majority voting according to \( P^* \) is equivalent to \( c \).
Putting it together

- Boosting maintains distribution over data
Putting it together

- Boosting maintains distribution over data
- Minimax argument requires distribution over hypothesis
Putting it together

- Boosting maintains distribution over data
- Minimax argument requires distribution over hypothesis

Change the mistake matrix to $M' = 1 - M^T$ so that:

$$M'(x, h) = 1 - M(h, x) = 1 \text{ if } h(x) = c(x) \text{ and 0 otherwise}$$
Putting it together

- Boosting maintains distribution over data
- Minimax argument requires distribution over hypothesis
- Change the mistake matrix to $M' = 1 - M^T$ so that $M'(x, h) = 1 - M(h, x) = 1$ if $h(x) = c(x)$ and 0 otherwise
- On round $t$ of boosting
  - Let $D_t = P_t$ from hedging over rows (data)
  - Weak learner returns $h_t$ with $P_{x \sim D_t}[h_t(x) = c(x)] \geq \frac{1}{2} + \gamma$
  - Update weights using $Q_t = h_t$ by adversary
Putting it together

- Boosting maintains distribution over data
- Minimax argument requires distribution over hypothesis
- Change the mistake matrix to $M' = 1 - M^T$ so that $M'(x, h) = 1 - M(h, x) = 1$ if $h(x) = c(x)$ and 0 otherwise
- On round $t$ of boosting
  - Let $D_t = P_t$ from hedging over rows (data)
  - Weak learner returns $h_t$ with $P_{x \sim D_t}[h_t(x) = c(x)] \geq \frac{1}{2} + \gamma$
  - Update weights using $Q_t = h_t$ by adversary
- Let $h_{\text{fin}}$ be the majority vote of $h_t, [t]_1^T$
We have $M'(P_t, h_t) = P_{x \sim P_t}[h_t(x) = c(x)] \geq \frac{1}{2} + \gamma$
Analysis of Boosting

- We have $M'(P_t, h_t) = P_{x \sim P_t} [h_t(x) = c(x)] \geq \frac{1}{2} + \gamma$
- Hence, from Corollary,

$$\frac{1}{2} + \gamma \leq \frac{1}{T} \sum_t M'(P_t, h_t) \leq \min_x \frac{1}{T} \sum_t M'(x, h_t) + \Delta_T$$
Analysis of Boosting

- We have $M'(P_t, h_t) = \mathbb{P}_{x \sim P_t}[h_t(x) = c(x)] \geq \frac{1}{2} + \gamma$
- Hence, from Corollary,

$$\frac{1}{2} + \gamma \leq \frac{1}{T} \sum_t M'(P_t, h_t) \leq \min_x \frac{1}{T} \sum_t M'(x, h_t) + \Delta_T$$

- Therefore, $\forall x$ (and sufficiently large $T$)

$$\frac{1}{T} \sum_t M'(x, h_t) \geq \frac{1}{2} + \gamma - \Delta_T > \frac{1}{2}$$
Analysis of Boosting

- We have $M'(P_t, h_t) = P_{x \sim P_t}[h_t(x) = c(x)] \geq \frac{1}{2} + \gamma$

- Hence, from Corollary,

$$\frac{1}{2} + \gamma \leq \frac{1}{T} \sum_t M'(P_t, h_t) \leq \min_x \frac{1}{T} \sum_t M'(x, h_t) + \Delta_T$$

- Therefore, $\forall x$ (and sufficiently large $T$)

$$\frac{1}{T} \sum_t M'(x, h_t) \geq \frac{1}{2} + \gamma - \Delta_T > \frac{1}{2}$$

- Hence, the majority vote $h_{\text{fin}}(x) = c(x)$, $\forall x$
Analysis of Boosting

- We have $M'(P_t, h_t) = P_{x \sim P_t}[h_t(x) = c(x)] \geq \frac{1}{2} + \gamma$
- Hence, from Corollary,
  \[
  \frac{1}{2} + \gamma \leq \frac{1}{T} \sum_t M'(P_t, h_t) \leq \min_x \frac{1}{T} \sum_t M'(x, h_t) + \Delta_T
  \]
- Therefore, $\forall x$ (and sufficiently large $T$)
  \[
  \frac{1}{T} \sum_t M'(x, h_t) \geq \frac{1}{2} + \gamma - \Delta_T > \frac{1}{2}
  \]
- Hence, the majority vote $h_{\text{fin}}(x) = c(x), \forall x$
- But what happens on unseen data?