CIS 606 Spring 2013
Machine Learning
Lecture 12

Online Learning

Wei Lee Woon and

Zeyar Aung

Online Learning: the General Setting

- For $t = 1 \dots T$
  1. I receive some input $\underline{x}^{(t)}$
  2. I make some prediction $\hat{y}^{(t)} \in \{-1, +1\}$
  3. I receive the correct label $y^{(t)} \in \{-1, +1\}$. If $\hat{y}^{(t)} \neq y^{(t)}$ I have made an error.
  4. I potentially update my hypothesis based on the new information $\underline{x}^{(t)}, y^{(t)}$.

The goal: to minimize the number of errors I make on the sequence

Online Learning

Examples:

- *Weather prediction.* For each day I receive some input $\underline{x}^{(t)}$ summarizing various measurements. My goal is to predict whether it will rain or not.

- *Stock market prediction.* For each day I receive some input $\underline{x}^{(t)}$ summarizing various measurements. My goal is to predict whether the stock market will go up or not.

Online Learning: Predicting from Expert Advice

- A useful setting is as follows. Assume each $\underline{x}^{(t)}$ is a vector in $\{-1, +1\}^d$ summarizing the advice from d *experts*.

- $x_j^{(t)}$ is the prediction of the j'th expert. I.e., $x_j^{(t)} = +1$ if the j'th expert predicts $+1$ (similarly for -1).

- For any sequence $(x^{(1)}, y^{(1)}) \ldots (x^{(T)}, y^{(T)})$, the loss of the j'th expert is

$$L_j = \sum_t [[x_j^{(t)} \neq y^{(t)}]]$$

- We'd like to design an online algorithm that for **any sequence** does **nearly as well as the best expert**

The Halving Algorithm

- First case: we assume that there is at least one $j \in \{1 \ldots d\}$ such that

$$L_j = 0$$

The Halving Algorithm

- Initialization: $\mathcal{Q}^{(1)} = \{1, 2, 3, \ldots d\}$
- For $t = 1 \ldots T$
  1. I receive some input $\underline{x}^{(t)}$
  2. Define
  $$\mathcal{Q}_{+}^{(t)} = \{j \in \mathcal{Q}^{(t)} : x_j^{(t)} = +1\}$$
  $$\mathcal{Q}_{-}^{(t)} = \{j \in \mathcal{Q}^{(t)} : x_j^{(t)} = -1\}$$
     If $|\mathcal{Q}_{+}^{(t)}| > |\mathcal{Q}_{-}^{(t)}|$ predict $\hat{y}^{(t)} = +1$, else $\hat{y}^{(t)} = -1$
  3. I receive the correct label $y^{(t)} \in \{-1, +1\}$. If $\hat{y}^{(t)} \neq y^{(t)}$ I have made an error.
  4. Update: if $y^{(t)} = +1$ then $\mathcal{Q}^{(t+1)} = \mathcal{Q}_{+}^{(t)}$, else $\mathcal{Q}^{(t+1)} = \mathcal{Q}_{-}^{(t)}$.

The Halving Algorithm: Guarantees

- For any value of T, for any sequence $(x^{(1)}, y^{(1)}) \ldots (x^{(T)}, y^{(T)})$ such that at least one expert j has $L_j = 0$ errors on the sequence, the halving algorithm makes at most $\log_2 d$ errors.

- Proof:
  - $|\mathcal{Q}^{(1)}| = d$
  - $|\mathcal{Q}^{(T+1)}| \geq 1$ (note that $\mathcal{Q}^{(T+1)}$ contains all experts that have $L_j = 0$ on the sequence)
  - If we make an error on the t'th example, we have

$$|\mathcal{Q}^{(t+1)}| \leq \frac{1}{2}|\mathcal{Q}^{(t)}|$$

  - We can halve $\mathcal{Q}^{(1)}$ at most $\log_2 d$ times before reaching a set of size zero, hence the number of mistakes is at most $\log_2 d$.

The Weighted Majority Algorithm

- Parameter: $0 < \beta < 1$
- Initialization: set $w_j = 1$ for $j = 1 \ldots d$.

- For $t = 1 \ldots T$
  1. I receive some input $\underline{x}^{(t)}$
  2. Define
  $$q_+^{(t)} = \sum_{j: x_j^{(t)} = +1} w_j; \quad q_-^{(t)} = \sum_{j: x_j^{(t)} = -1} w_j$$
  If $q_+^{(t)} > q_-^{(t)}$ predict $\hat{y}^{(t)} = +1$, else $\hat{y}^{(t)} = -1$
  3. I receive the correct label $y^{(t)} \in \{-1, +1\}$. If $\hat{y}^{(t)} \neq y^{(t)}$ I have made an error.
  4. Update: for all j such that $x_j^{(t)} \neq y^{(t)}$, set $w_j = w_j \times \beta$

Guarantees for the Weighted Majority Algorithm

- For any value of T, for any sequence $(x^{(1)}, y^{(1)}) \ldots (x^{(T)}, y^{(T)})$, the weighted majority algorithm makes at most

$$f(\beta) \times \min_j L_j + g(\beta) \times \log d$$

mistakes, where L_j is the loss of the j'th expert on the sequence, and

$$f(\beta) = \frac{\log\left(\frac{1}{\beta}\right)}{\log\left(\frac{2}{1+\beta}\right)} \qquad g(\beta) = \frac{1}{\log\left(\frac{2}{1+\beta}\right)}$$

- E.g., with $\beta = 1/e$, $f(\beta) = g(\beta) = 2.63$.

Behaviour of $f(\beta)$ and $g(\beta)$

β	$f(\beta)$	$g(\beta)$
0.1	3.85	1.67
0.2	3.15	1.95
0.3	2.79	2.32
0.4	2.56	2.80
0.5	2.40	3.47
0.6	2.28	4.48
0.7	2.19	6.15
0.8	2.11	9.49
0.9	2.05	19.49

Original source:

- http://courses.csail.mit.edu/6.867/lectures/lecture7_slides.pdf (by Prof Michael Collins)