What is online learning?

Sample data are arranged in a sequence.

Each time we get a new input, the algorithm tries to predict the corresponding output.

As the number of seen samples increases, hopefully the predictions improve.

Assets

1. does not require storing all data samples

2. more plausible model for sequential problems, especially those that involve decision-making

3. typically fast algorithms

4. it is possible to give formal guarantees **not** assuming probabilistic hypotheses (*mistake bounds*)

Problems

- Performance can be worse than best **batch** algorithms
- **Generalization bounds** always require some assumption on the generation of sample data

Online setting

Sequence of sample data $z_1, z_2, \ldots, z_n$.

Each sample is an input-output couple $z_i = (x_i, y_i)$.

$x_i \in X \subset R^d$, $y_i \in Y \subset R$.

In the *classification* case $Y = \{+1, -1\}$

Estimators $f_i : X \to Y$ constructed using the first i data samples.

Online setting (cont.)

- initialization f_0
- for $i = 1, 2, \ldots, n$
  - receive x_i
  - predict $f_{i-1}(x_i)$
  - receive y_i
  - update $(f_{i-1}, z_i) \rightarrow f_i$

Note: storing efficiently f_{i-1} may require much less memory than storing all previous samples $z_1, z_2, \ldots, z_{i-1}$.

Goals

Batch learning: reducing *expected loss*

$$I[f_n] = E_z V(f_n(x), y)$$

Online learning: reducing *cumulative loss*

$$\sum_{i=1}^{n} V(f_{i-1}(x_i), y_i)$$

The Experts Framework

We will focus on the classification case.

Suppose we have a pool of prediction strategies, called experts. Denote by $E = \{E_1, \ldots, E_k\}$.

Each expert predicts y_i based on x_i.

We want to combine these experts to produce a single *master algorithm* for classification and prove bounds on how much worse it is than the *best* expert.

The Halving Algorithm*

Suppose all the experts are functions (their predictions for a point in the space do not change over time) and at least one of them is *consistent* with the data.

At each step, predict what the majority of experts that have not made a mistake so far would predict.

Note that all inconsistent experts get thrown away!

Maximum of $\log_2(|E|)$ errors.

But what if there is no consistent function in the pool? (Noise in the data, limited pool, etc.)

*Barzdin and Freivald, *On the prediction of general recursive functions*, 1972, Littlestone and Warmuth, *The Weighted Majority Algorithm*, 1994

The Weighted Majority Algorithm*

Associate a weight w_i with every expert. Initialize all weights to 1.

At example t:

$$q_{-1} = \sum_{i=1}^{|E|} w_i I[E_i \text{ predicted } y_t = -1]$$

$$q_1 = \sum_{i=1}^{|E|} w_i I[E_i \text{ predicted } y_t = 1]$$

Predict $y_t = 1$ if $q_1 > q_{-1}$, else predict $y_t = -1$

If the prediction is wrong, multiply the weights of each expert that made a wrong prediction by $0 \leq \beta < 1$.

Note that for $\beta = 0$ we get the halving algorithm.

*Littlestone and Warmuth, 1994

Mistake Bound for WM

For some example t let $W_t = \sum_{i=1}^{|E|} w_i = q_{-1} + q_1$

Then when a mistake occurs $W_{t+1} \leq uW_t$ where $u < 1$

Therefore $W_0 u^m \geq W_n$

Or $m \leq \frac{\log(W_0/W_n)}{\log(1/u)}$

Then $m \leq \frac{\log(W_0/W_n)}{\log(2/(1+\beta))}$ (setting $u = \frac{1+\beta}{2}$)

Mistake Bound for WM (contd.)

Why? Because when a mistake is made, the ratio of total weight after the trial to total weight before the trial is at most $(1+\beta)/2$.

W.L.o.G. assume WM predicted -1 and the true outcome was $+1$. Then new weight after trial is:

$\beta q_{-1}+q_1 \leq \beta q_{-1}+q_1+\frac{1-\beta}{2}(q_{-1}-q_1) = \frac{1+\beta}{2}(q_{-1}+q_1)$.

The main theorem (Littlestone & Warmuth): *Assume m_i is the number of mistakes made by the ith expert on a sequence of n instances and that $|E| = k$. Then the WM algorithm makes at most the following number of mistakes:*

$$\frac{\log(k) + m_i \log(1/\beta)}{\log(2/(1+\beta))}$$

Big fact: Ignoring leading constants, the number of errors of the pooled predictor is bounded by the sum of the number of errors of the best expert in the pool and the log of the number of experts!

Finishing the Proof

$W_0 = k$ and $W_n \geq \beta^{m_i}$

$\log(W_0/W_n) = \log(W_0) - \log(W_n)$

$\log(W_n) > m_i \log \beta$, so $-\log(W_n) < m_i \log(1/\beta)$

Therefore $\log(W_0) - \log(W_n) < \log k + m_i \log(1/\beta)$