Notes 6: Weighted Majority

1. Online regret bound model

e.g. stock market prediction: guessing whether it will go up or down for each day
A sequence of rounds/trials, each being:

(1) A new unlabeled example x arrives
(2) n experts reveal their opinions about the label for x (label is either 0 or 1)
(3) Algorithm predicts 0 or 1 according to experts' opinions
(4) Algorithm is told correct label for x

Goal: minimize number of mistakes, compared with the best expert
If every "expert" makes many mistakes, algorithm may, too

---

2. Weighted Majority

Weighted Majority

Fix parameter $0 \leqslant \beta < 1$
Initialize: $w_1 = \cdots = w_n = 1$
On input x, poll opinions from experts
  Compute total weight q_0 of experts predicting 0 and total weight q_1 predicting 1
  Predict according to weighted majority (predict 0 if $q_0 > q_1$; predict 1 otherwise)
On revealing correct label, penalize incorrect experts
  Multiply every incorrect expert i's weight w_i by β

If $\beta = 0$, Weighted Majority algorithm becomes Halving algorithm

expert	$\longleftrightarrow$	concept
expert i's opinion in jth trial	$\longleftrightarrow$	concept c's classification for jth sample

No longer assume any expert/concept correctly classifies all samples
Robust to classification noise

Theorem 1. *For any trial sequence, if the best expert (out of n experts) makes m mistakes, then number of mistakes of Weighted Majority is at most*

$$\frac{\log n + m\log(1/\beta)}{\log(\frac{2}{1+\beta})}$$

e.g. $\beta = 1/2$: $2.41(m + \log n)$
e.g. $\beta = 3/4$: $2.2m + 5.2\log n$
e.g. $\beta = 1-\varepsilon$: $\approx (2 + \frac{3}{2}\varepsilon)m + \frac{2}{\varepsilon}\log n$

Proof. let $W = q_0 + q_1 =$ total weight of all experts (initially n)
After each mistake, at least half of W shrinks by factor β
  Total weight reduces to $\leqslant \frac{W}{2} + \beta\frac{W}{2} = \frac{1+\beta}{2}W$
when Weighted Majority makes M mistakes: $W \leqslant (\frac{1+\beta}{2})^M n$
when best expert makes m mistakes: $w_i = \beta^m$
$w_i \leqslant W \implies \beta^m \leqslant (\frac{1+\beta}{2})^M n \iff m\log\beta \leqslant M\log(\frac{1+\beta}{2}) + \log n$
$\iff M\log(\frac{2}{1+\beta}) \leqslant \log n + m\log(1/\beta)$ □

Note: The bound can be interpreted as

$$\frac{\log(W_{\text{init}}/W_{\text{final}})}{\log(1/u)} \quad \text{where } u = \frac{1+\beta}{2} = \text{shrink in } W \text{ per mistake}$$

3. Randomized Weighted Majority

Randomized Weighted Majority

Fix parameter $0 \leqslant \beta < 1$
Initialize: $w_1 = \cdots = w_n = 1$
On input x, poll opinions from experts
    Predict according to a random expert i chosen with probability proportional to w_i
    i.e. probability w_i/W, where $W =$ total weight $= \sum_{1 \leqslant i \leqslant n} w_i$
On revealing correct label, penalize incorrect experts
    Multiply every incorrect expert i's weight w_i by β

Denote $\varepsilon = 1 - \beta$

Theorem 2. *Given any trial sequence with fixed correct labels, if the best expert (out of n experts) makes m mistakes, then*

$$\mathbb{E}[\#\textit{mistakes of RWM}] \leqslant \frac{\ln n - m \ln(1-\varepsilon)}{\varepsilon}$$

e.g. $\beta = 1/2$: $1.39m + 2\ln n$
e.g. $\beta = 3/4$: $1.16m + 4\ln n$
e.g. $\beta = 1-\varepsilon$: $\approx (1 + \frac{\varepsilon}{2})m + \frac{1}{\varepsilon}\ln n$
Key benefit: $\approx m$ mistakes (ignoring additive $\log n$), down from $\approx 2m$

Proof. Fix any sequence of T trials together with their correct labels
Let $F_t =$ fraction of total weight on wrong prediction at trial t
Want to bound $\mathbb{E}[\#\text{mistakes of RWM}] = \sum_{1 \leqslant t \leqslant T} F_t$
At trial t, probability of mistake is F_t, and εF_t fraction of weight is removed

$$W_{\text{final}} = W_{\text{init}}(1-\varepsilon F_1)\ldots(1-\varepsilon F_T) \qquad (W_{\text{init}} = n)$$

$$\ln W_{\text{final}} = \ln n + \ln(1-\varepsilon F_1) + \cdots + \ln(1-\varepsilon F_T)$$

Best expert makes m mistakes: $w_i = \beta^m = (1-\varepsilon)^m$

$$W_{\text{final}} \geqslant w_i \iff \ln W_{\text{final}} \geqslant \ln w_i \iff \ln n + \sum_{1 \leqslant t \leqslant T} \ln(1-\varepsilon F_t) \geqslant m \ln(1-\varepsilon)$$

Claim: $\ln(1-x) \leqslant -x$ for all $x < 1$
Take $x = \varepsilon F_t$ in Claim, we get $\ln(1-\varepsilon F_t) \leqslant -\varepsilon F_t$, and

$$\varepsilon \sum_{1 \leqslant t \leqslant T} F_t \leqslant \sum_{1 \leqslant t \leqslant T} -\ln(1-\varepsilon F_t) \leqslant \ln n - m \ln(1-\varepsilon)$$ □

Above Claim is true because for all real x $\quad 1 - x \leqslant e^{-x}$