The Weighted Majority Algorithm

Fabrice Zapfack   Hugo Marival

May 18, 2016

Plan

Plan

Halving Algorithm

- At each step predict the majority vote
- Keep only the functions that are consistent

Theorem

Let m be the number of mistakes when we apply Halving on $\mathcal{S}$ with the pool $\mathcal{A}$.

$$m \leq log_2(|A|)$$

ATTENTION - Works only in the realisable case.

Plan

Framework for WMA

WMA - Given a pool of algorithms $\mathcal{A}$ where one of them performs well, we will create a compound algorithm based on a weighted voting of the algorithms of $\mathcal{A}$.

Different frameworks

- **WM** - Binary WM predictions, binary pool predictions, deterministic WM and pool algorithms
- **WMG** - Binary WM predictions, continuous (in [0,1]) pool predictions, deterministic WM and pool algorithms
- **WMC** - Continuous (in [0,1]) WM and pool predictions, deterministic WM and pool algorithms
- **WMR** - Continuous (in [0,1]) WM and pool predictions, probabilistic WM and pool algorithms

Plan

Bound on the number of mistakes

Notations

- $\mathcal{S}$: sequence of instances and binary labels
- w_{init} : initial total weight of the algorithms in the pool
- w_{fin} : final total weight of the algorithms in the pool
- q_0 : total weight of the algorithms that predict 0
- q_1 : total weight of the algorithms that predict 1
- β : weight multiplication factor in case of a mistake $(0 \leq \beta < 1)$

Theorem

Let m be the number of mistakes when we apply WMA on $\mathcal{S}$ with the pool $\mathcal{A}$.

$$m \leq \frac{\log(w_{init}/w_{fin})}{\log(2/(1+\beta))}$$

PROOF SKETCH

- Step 1 - Prove that if in a trial WM makes a mistake, the sum of weights before the trial is greater than $u = \frac{1+\beta}{2}$ times the sum of weights after
- Step 2 - Recursively, we get that $w_{init} u^m \geq w_{fin}$
- Step 3 - Take the log of the above inequality to conclude

PROOF OF STEP 1 - Let us suppose wlog that the learner predicted 0 which was the wrong binary label in this trial.

- Total weight before : $q_0 + q_1$. Since 0 was predicted, $q_0 \geq q_1$
- Total weight after :
  $\beta q_0 + q_1 \leq \beta q_0 + q_1 + \frac{1-\beta}{2}(q_0 + q_1) \leq \frac{1+\beta}{2}(q_0 + q_1)$

Plan

WML - Adaptive version

What motivates this modified version (WML)?

- First version : selects a group of "good" algorithms and makes other weights shrink
- Not very good for adaptation : if throughout the trials another group becomes better it will not be seen because of its small weight
- Idea : Do not allow an algorithm's weight to be lower that $\frac{\gamma}{|\mathcal{A}|}$ times the total weight of $\mathcal{A}$

WML - Bounds

Lemma

Let m_0 be the minimum number of mistakes made on the sequence $\mathcal{S}$ by any algorithm of the pool $\mathcal{A}$ of n algorithms. If the initial weight of each algorithm is at least $\frac{\beta\gamma}{n}$ times the total initial weight, then WML applied to $\mathcal{A}$ makes at most

$$\frac{\log(n/\beta\gamma) + m_0 \log(1/\beta)}{\log(1/u)}$$

mistakes, with $u = \frac{1+\beta}{2} + (1-\beta)\gamma$. The final weight of each algorithm is at least $\frac{\beta\gamma}{n}$ times the total final weight

Plan

Assumptions for WMG and WMC

- Predictions of algorithms in the pool in [0,1] for both
- WMG predictions are binary
- WMC predictions are in [0,1]
- Labels are binary for WMG
- Labels are in [0,1] for WMC

Update step

- Updates at every trial for WMC
- Updates either at every trial or only when a mistake occurs

Notations

- *Update-trial j* : j-th trial in which an update occurs, $j = 1..t$, $|\mathcal{A}| = n$
- $x_i^{(j)}$: prediction of the i-th algorithm in update-trial j
- $\lambda^{(j)}$: prediction of the master (compound) algorithm in update-trial j
- $\rho^{(j)}$: label of update-trial j
- $w_1^{(j)}, \ldots, w_n^{(j)}$: weights at the beginning of update-trial j
- $s^{(j)} = \sum_{i=1}^{n} w_i^{(j)}, \gamma^{(j)} = \frac{\sum_{i=1}^{n} w_i^{(j)} x_i^{(j)}}{s^{(j)}}$

Predictions

- WMC : $\lambda^{(j)} = \gamma^{(j)}$
- WMG : $\lambda^{(j)} = 1$ if $\gamma^{(j)} \geq \frac{1}{2}$ and $\lambda^{(j)} = 0$ if $\gamma^{(j)} < \frac{1}{2}$

Bound for WMG

Definition (Update step for WMG and WMC)

$w_i^{(j+1)} = Fw_i^{(j)}$ *where* $F = F(\beta, x_i^{(j)}, \rho^{(j)})$ *satisfies :*

$$\beta^{|x_i^{(j)} - \rho^{(j)}|} \leq F \leq 1 - (1-\beta)|x_i^{(j)} - \rho^{(j)}|$$

Theorem (Bound for WMG)

Let m be the number of mistakes when running WMG on $\mathcal{S}$ with $0 \leq \beta < 1$.

$$m \leq \frac{\log(w_{init}/w_{fin})}{\log(2/(1+\beta))}$$

Bound for WMG - Proof

Let us start by proving that we can always find an update factor F :

Lemma

For $\beta \geq 0$ and $0 \leq r \leq 1$: $\beta^r \leq 1 + r(\beta - 1)$

PROOF - Convexity inequality on $r \mapsto \beta^r$
In the following lemma, we will assume that $w_i^{(1)} > 0$, $\rho^{(j)} \leq 1$ and $0 \leq x_i^{(j)} \leq 1$ for $i = 1..n, j = 1..t$.

Lemma (5.2)

Let us also assume that for $i = 1..n, j = 1..t, w_i^{(j+1)} \leq w_i^{(j)}(1 - (1-\beta)|x_i^{(j)} - \rho^{(j)}|)$*. If* $\beta = 0$ *and* $|\gamma^{(j)} - \rho^{(j)}| = 1$ *for some* j*, then* $w_{fin} = 0$*. Otherwise,*

$$\log\left(\frac{w_{fin}}{w_{init}}\right) \leq \sum_{j=1}^{t} \log(1 - (1-\beta)|\gamma^{(j)} - \rho^{(j)}|)$$

PROOF.

First case - If $\beta = 0$ and $|\gamma^{(j)} - \rho^{(j)}| = 1$ for some j. Then

$$\left|\frac{\sum_i w_i^{(j)} x_i^{(j)}}{\sum_i w_i^{(j)}} - \frac{\sum_i w_i^{(j)} \rho^{(j)}}{\sum_i w_i^{(j)}}\right| = 1$$

$$\Rightarrow \left|\frac{\sum_i w_i^{(j)}(x_i^{(j)} - \rho^{(j)})}{\sum_i w_i^{(j)}}\right| = 1$$

Proof of Lemma 5.2 (..page 2..)

$$\Rightarrow \frac{\sum_i w_i^{(j)}|x_i^{(j)} - \rho^{(j)}|}{\sum_i w_i^{(j)}} \geq 1$$

Since $x_i^{(j)}, \rho^{(j)} \in [0,1]$, $|x_i^{(j)} - \rho^{(j)}| \leq 1$, so $\frac{\sum_i w_i^{(j)}|x_i^{(j)} - \rho^{(j)}|}{\sum_i w_i^{(j)}}$ can only be greater than 1 if $|x_i^{(j)} - \rho^{(j)}| = 1$ for $i = 1..n$, so we have to use the update factor $(\beta = 0)$ and $w_i^{(j+1)} = 0$ for $i = 1..n$ so $w_{fin} = 0$.

Second case - Using the convexity inequality from the proof of a previous lemma :

$$s^{(j+1)} \leq \sum_{i=1}^{n} w_i^{(j)}(1-(1-\beta)|x_i^{(j)}-\rho^{(j)}|) = s^{(j)}-(1-\beta)\sum_{i=1}^{n} w_i^{(j)}|x_i^{(j)}-\rho^{(j)}|$$

Proof of Lemma 5.2 (..end)

Using the triangular inequality:

$$\begin{aligned} s^{(j)} - (1-\beta)\sum_{i=1}^{n} w_i^{(j)}|x_i^{(j)} - \rho^{(j)}| &\leq s^{(j)} - (1-\beta)|\sum_{i=1}^{n} w_i^{(j)}(x_i^{(j)} - \rho^{(j)})| \\ &= s^{(j)} - (1-\beta)|\sum_{i=1}^{n} \gamma^{(j)} s^{(j)} - \rho^{(j)} s^{(j)}| \\ &= s^{(j)}(1 - (1-\beta)|\gamma^{(j)} - \rho^{(j)}|) \end{aligned}$$

Recursively, we get :

$$s^{(t+1)} \leq s^{(1)} \prod_{j=1}^{t}(1 - (1-\beta)|\gamma^{(j)} - \rho^{(j)}|)$$

Proof of the WMG bound

PROOF OF THE WMG BOUND. First case - $\beta = 0 \Rightarrow w_{fin} = 0$ so the bound becomes ... Second case - Let $m^{(j)} = 1$ if WMG makes a mistake in update-trial j, 0 otherwise. $m = \sum_{j=1}^{t} m^{(j)}$.
Since $\log(1 - (1-\beta)|\gamma^{(j)} - \rho^{(j)}|) \leq 0$,

$$\sum_{j=1}^{t} \log(1-(1-\beta)|\gamma^{(j)}-\rho^{(j)}|) \leq \sum_{j \text{ s.t. } m^{(j)}=1} \log(1-(1-\beta)|\gamma^{(j)}-\rho^{(j)}|)$$

- If $\gamma^{(j)} < \frac{1}{2}$, $m^{(j)} = 1 \Rightarrow \rho^{(j)} = 1 \Rightarrow |\gamma^{(j)} - \rho^{(j)}| \geq \frac{1}{2}$
- If $\gamma^{(j)} \geq \frac{1}{2}$, $m^{(j)} = 1 \Rightarrow \rho^{(j)} = 0 \Rightarrow |\gamma^{(j)} - \rho^{(j)}| \geq \frac{1}{2}$

Proof of the WMG bound (..end)

So,

$$\sum_{j \text{ s.t. } m^{(j)}=1} \log(1-(1-\beta)|\gamma^{(j)}-\rho^{(j)}|) \leq m\log(1-\frac{1}{2}(1-\beta)) = m\log(\frac{1}{2}+\frac{1}{2}\beta)$$

We can use Lemma 5.2 and get :

$$\log(\frac{w_{fin}}{w_{init}}) \leq m\log(\frac{1+\beta}{2})$$

Bound for WMC

Definition (Loss m for WMC)

For continuous predictions, the loss is defined by :

$$m = \sum_{j=1}^{t} |\lambda^{(j)} - \rho^{(j)}|$$

Theorem

Let $\mathcal{S}$ be any sequence of instances and labels, with labels in [0,1]. Let m be the total loss for the WMC.

$$m \leq \frac{\log(w_{init}/w_{fin})}{1-\beta}$$

Lemma (5.3)

If the conditions of Lemma 5.2 are satisfied, then

$$\sum_{j=1}^{t} \left|\gamma^{(j)} - \rho^{(j)}\right| \leq \frac{\log(w_{init}/w_{fin})}{1-\beta}$$

PROOF

$$\log\left(1 - (1-\beta)\left|\gamma^{(j)} - \rho^{(j)}\right|\right) \leq -(1-\beta)\left|\gamma^{(j)} - \rho^{(j)}\right|$$

We then use Lemma 5.2 to get the bound.

Plan

Assumptions

- Predictions of pool members are in [0,1]
- Prediction of WMR is binary but probabilistic
- Labels associated with instances are binary

Prediction of WMR

- WMR predicts 1 with probability $\gamma^{(j)}$
- If pool members' predictions are binary, $\gamma^{(j)} = \frac{q_1}{q_0+q_1}$.

Update criterion

- Update in every trial

Update step

- Same update step as WMG and WMC.

Weak independence

In WMR, $x_i^{(j)}$ and $\rho^{(j)}$ are random variables. In order to count the mistakes made by WMR, we will use the following assumption :
WEAK INDEPENDENCE CONDITION

$$\mathbb{E}\left[\lambda^{(j)}|(x^{(1)},\rho^{(1)}),..,(x^{(j)},\rho^{(j)})\right]=\gamma^{(j)} \text{ for } j=1..t$$

REMARK If $x_i^{(j)}$ and $\rho^{(j)}$ are chosen deterministically then all of the weights and $\gamma^{(j)}$ are also deterministic and the construction of the algorithm gives us that

$$\mathbb{E}\left[\lambda^{(j)}|(x^{(1)},\rho^{(1)}),..,(x^{(j)},\rho^{(j)})\right]=\mathbb{E}\left[\lambda^{(j)}\right]=\gamma^{(j)}$$

Strong independence

To give a bound on the concentration of the total number of mistakes around it's mean, we will use the following assumption:
STRONG INDEPENDENCE CONDITION

$$\mathbb{E}\left[\lambda^{(j)} | (x^{(1)}, \rho^{(1)}), .., (x^{(t)}, \rho^{(t)}), \lambda^{(1)}, .., \lambda^{(j-1)}\right] = \gamma^{(j)}$$

REMARK
Strong independence $\Rightarrow$ weak independence.

Bound on the expected number of mistakes

Theorem

Let $\mathcal{S}$ be a sequence of instances with binary labels. Let m be the number of mistakes made by WMR on $\mathcal{S}$ when applied to a pool of probabilistic prediction algorithms. If the weak independence condition holds,

$$\mathbb{E}[m] \leq \frac{\mathbb{E}[\log(w_{init}/w_{fin})]}{1-\beta}$$

PROOF

Weak indep. condition :

$$\mathbb{E}\left[\left|\lambda^{(j)} - \rho^{(j)}\right| |(x^{(1)}, \rho^{(1)}), .., (x^{(j)}, \rho^{(j)})\right] = \left|\gamma^{(j)} - \rho^{(j)}\right|.$$

So $\mathbb{E}\left[\left|\lambda^{(j)} - \rho^{(j)}\right|\right] = \mathbb{E}\left[\left|\gamma^{(j)} - \rho^{(j)}\right|\right].$

$$\mathbb{E}[m] = \mathbb{E}\left[\sum_{j=1}^{t} \left|\lambda^{(j)} - \rho^{(j)}\right|\right]$$
$$= \mathbb{E}\left[\sum_{j=1}^{t} \left|\gamma^{(j)} - \rho^{(j)}\right|\right]$$

Then we use Lemma 5.3

Recap of the algorithms and their bounds

Recap (WMA & WMG)

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

Recap (WMC & WMR)

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

Plan

Plan

EWA

- fix $p_1 = \pi$ an arbitrary probability distribution on $\mathbb{R}^M$
- $\hat{y}_t = \int f_\theta(x_t) p_t(\mathrm{d})\theta$ and once y_t is revealed,

$$p_{t+1}(\mathrm{d}\theta) = \frac{\exp\left(-\eta \ell(y_t, f_\theta(x_t))\right) p_t(\mathrm{d}\theta)}{\int_{\mathbb{R}^M} \exp\left(-\eta \ell(y_t, f_\alpha(x_t))\right) p_t(\mathrm{d}\alpha)}.$$

Theorem

Taking $\eta = 2\sqrt{\frac{2\log(M)}{TC^2}}$ *leads to a regret in*

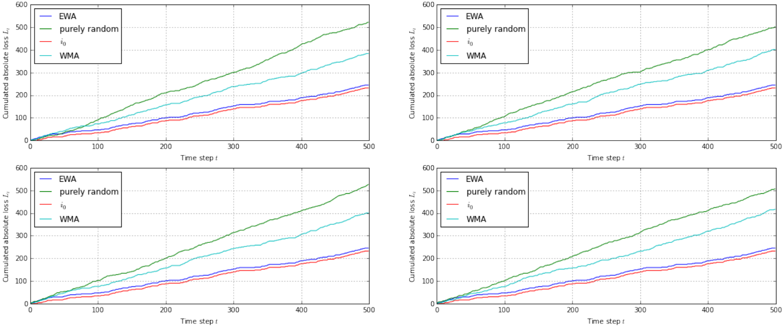
$$\mathcal{R}_T(\{f_1, \ldots, f_M\}) \leq C\sqrt{\frac{T\log(M)}{2}}.$$

ATTENTION - Applicable to L-type, C-type, MS-type aggregation

Plan

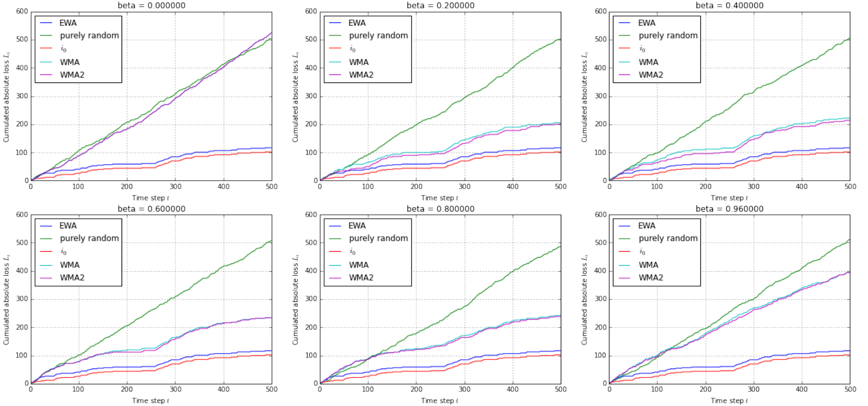
Plan

Randomly generated samples



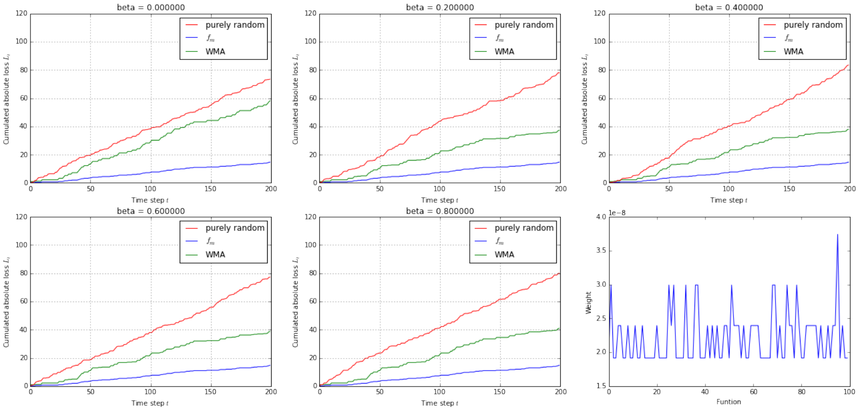
Randomly generated samples

$\mathrm{WMA1}$ - Initials weights are normally distributed



Plan

Exercise 6 (TD): Observations are randomly given



Exercise 6 (TD): Observations are given in the order of the paper