Predictors that Specialize

Yoav Freund

February 16, 2006
Outline

The specialists setup

bounding cumulative loss using relative entropy

Applications of specialists
The specialists setup

- Up till now we assumed that each expert makes a prediction at each iteration.
- Imagine that experts are specialists, they predict only some of the time.
- Gives the designer a lot of flexibility.
- Generalizes the switching experts setup.
The specialists game

On each iteration $t = 1, 2, 3, \ldots$
- Adversary chooses a set $E^t \subseteq \{1, \ldots, N\}$ of awake specialists.
- Adversary chooses predictions for specialists in $E^t$
- Algorithm chooses its prediction.
- Adversary chooses outcome.
Desired bound

- Algorithm has to predict on each iteration
- Each specialist might sleep some of the time.
- \(\Rightarrow\) makes no sense to compare to total loss of best specialist.
- \(u\): a probability distributions, \(u_i \geq 0, \sum_i u_i = 1\).
- Average loss w.r.t. \(u\): \(\ell^t_u = \frac{\sum_{i \in E^t} u_i \ell_i^t}{\sum_{i \in E^t} u_i}\)
- Goal: \(L_A \leq \min_u \sum_{t=1}^T \ell^t_u + \text{something small}\)
Applying Vovk-style algs to specialists

- We use normalized weights:

\[ v_i^t = \frac{w_i^t}{\sum_{j=1}^{N} w_j^t}, \quad v^t = \frac{w^t}{W^t} \]

- Algorithm: treat the set \( E_t \) as the set of experts.
- Normalize the weights of specialists in \( E_t \) so that

\[ \sum_{i \in E_t} v_i^t = \sum_{i \in E_t} v_i^{t+1} \]

- In particular: total weight is always 1.
Bound for log-loss case

- Bound for log loss (Theorem 1), for any distribution $u$:
  \[ \sum_{t=1}^{T} u(E_t) \ell_A^t \leq \sum_{t=1}^{T} \sum_{i \in E_t} u_i \ell_i^t + \text{RE}(u \| v^1) \]
- $\text{RE}(u \| v) \doteq \sum_i u_i \log \frac{u_i}{v_i}$
- $u(E_t) \doteq \sum_{i \in E_t} u_i$
- If we assume that $u(E_t) = U$ is constant, we get
  \[ L_A \leq \sum_{t=1}^{T} \ell_u^t + \frac{\text{RE}(u \| v^1)}{U} \]
Cumulative loss vs. Final total weight

Total weight: \( W^t = \sum_{i=1}^{N} w_i^t \)

\[
\frac{W^{t+1}}{W^t} = \frac{\sum_{i=1}^{N} w_i^t e^{\log p_i^t(c^t)}}{\sum_{i=1}^{N} w_i^t} = \frac{\sum_{i=1}^{N} w_i^t p_i^t(c^t)}{\sum_{i=1}^{N} w_i^t} = p_A(c^t)
\]

\[-\log \frac{W^{t+1}}{W^t} = -\log p_A^t(c^t)\]

\[-\log W^{T+1} = -\log \frac{W^{T+1}}{W^1} = -\sum_{t=1}^{T} \log p_A^t(c^t) = L_A^T\]

**EQUALITY** not bound!
Relative Entropy

- $u, v$: probability distributions, $u_i \geq 0$, $\sum_i u_i = 1$.
- $\text{RE}(u \| v) = \sum_i u_i \log \frac{u_i}{v_i}$
- $\text{RE}(u \| v) \geq 0$, $\text{RE}(u \| v) = 0$ iff $u = v$
- $\exists u, v$, $\text{RE}(u \| v) \neq \text{RE}(v \| u)$
- $\exists u_1, u_2, u_3$, $\text{RE}(u_1 \| u_3) > \text{RE}(u_1 \| u_2) + \text{RE}(u_2 \| u_3)$
Normalized weights notation

- $p^t_i$: distribution (of letters) predicted by expert $i$ at time $t$
- Experts losses at time $t$: 
  \[ \ell^t = \langle \ell^t_1, \ldots, \ell^t_N \rangle = -\langle \log p^t_1(c^t), \ldots, \log p^t_N(c^t) \rangle \]
- Prediction of algorithm: 
  \[ p^t_A = \sum_{i=1}^{N} v^t_i p^t_i \]
- Loss of algorithm at time $t$: 
  \[ \ell^t_A = -\log p^t_A(c^t) \]
Bounding cumulative log loss using relative entropy

- Let \( u \) be an arbitrary distribution vector over experts.
- **Lemma:** \( \text{RE}(u \| v^t) - \text{RE}(u \| v^{t+1}) = \ell_A^t - u \cdot \ell^t \)
- Summing over \( t = 1, \ldots, T \) we get:
  \[ \text{RE}(u \| v^1) - \text{RE}(u \| v^{T+1}) = L_A - u \cdot \sum_{t=1}^T \ell^t \]
- \( L_A \leq \min_u \left( u \cdot \sum_{t=1}^T \ell^t + \text{RE}(u \| v^1) \right) \)
- For the special case \( u = \langle 0, \ldots, 0, 1, 0, \ldots, 0 \rangle \) and \( v^1 = \langle 1/N, \ldots, 1/N \rangle \) we get the old bound:
  \( L_A \leq \min_i L_i + \log N \)
Specialists bounding cumulative loss using relative entropy

Visual intuition

\[ \text{RE}(u || v^t) - \text{RE}(u || v^{t+1}) = \ell^t_A - u \cdot \ell^t \]

\( v^{t+1} \) is chosen to minimize \( \text{RE}(v^{t+1} || v^t) + v^{t+1} \cdot \ell^t \)

Last line is confusing! I don’t understand it!
But Manfred Warmuth does!
Proof of Lemma

\[
\begin{align*}
\text{RE}(u||v^t) - \text{RE}(u||v^{t+1}) &= \ell^t_A - u \cdot \ell^t \\
\text{RE}(u||v^t) - \text{RE}(u||v^{t+1}) &= \sum_i u_i \log \frac{u_i}{v_i^t} - \sum_i u_i \log \frac{u_i}{v_i^{t+1}} = \sum_i u_i \log \frac{v_i^{t+1}}{v_i^t} \\
&= \sum_i u_i \log \left( \frac{W^t}{W^{t+1}} \frac{w_i^{t+1}}{w_i^t} \right) \\
&= \log \frac{W^t}{W^{t+1}} + \sum_i u_i \log e^{-\ell_i^t} = \ell^T_A - \sum_i u_i \ell_i^t
\end{align*}
\]
Suppose that loss is \((a, c)\)-achievable.

Achievable with Vovk algorithm, learning rate \(\eta = \frac{a}{c}\).

Let \(u\) be an arbitrary distribution vector over experts.

Lemma: \(\text{RE}(u||v^t) - \text{RE}(u||v^{t+1}) \geq \frac{1}{c} \ell_A - \frac{a}{c} u \cdot \ell^t\)

Summing over \(t = 1, \ldots, T\) we get:
\[
\text{RE}(u||v^1) - \text{RE}(u||v^{T+1}) = \frac{1}{c} L_A - \frac{a}{c} u \cdot \sum_{t=1}^{T} \ell^t
\]

\(L_A \leq \min_u \left( a u \cdot \sum_{t=1}^{T} \ell^t + c \text{RE}(u||v^1) \right)\)

For any mixable loss, \(a = 1\), using 
\(u = \langle 0, \ldots, 0, 1, 0, \ldots, 0 \rangle\) and 
\(v^1 = \langle 1/N, \ldots, 1/N \rangle\) we get the old bound: 
\(L_A \leq \min_i L_i + c \log N\)
Example Application

- Consider the context algorithm.
- Let each node in the tree be a specialist.
- Gives an inferior algorithm (regret bound is twice as large)
- But much easier to generalize.
Generic Example

- Partition the input space. Assign each part to a specialist.
- Use several partitions, of different fineness.
- Can partition time in addition to space.
- Parts do not have to be disjoint.
- Partitions can adapt to data.
- Your idea here...