Agnostic Online Learning

Dávid Pál

March 2009

joint work with
Shai Ben-David and Shai Shalev-Shwartz
Online learning

In round $t = 1, 2, \ldots, T$

- receive $x_t$ e.g. an email
- predict $\hat{y}_t \in \{0, 1\}$ e.g. {spam, not-spam}
- receive “correct” feedback $y_t \in \{0, 1\}$
- $\hat{y}_t \neq y_t$ is a mistake
Overview

Previous work:
- Littlestone’s model
- learning with expert advice
- PAC model
- agnostic PAC

Our contribution:
- Agnostic online learning

Important technicalities:
- Littlestone’s dimension
- Simulating Expert’s
Littlestone’s model

Littlestone (1988)

- *unknown* target \( h^* : \mathcal{X} \rightarrow \{0, 1\} \) in *fixed known* class \( \mathcal{H} \)
- \( \hat{y}_t = h^*(x_t) \) for all \( t \)
  (So called “realizable case”.)
- How many mistakes do we make?
- Littlestone defined “optimal mistake bound” of \( \mathcal{H} \). We call it Ldim(\( \mathcal{H} \)) – Littlestone’s dimension
Learning with Expert Advice

Littlestone & Warmuth (1994), Vovk (1990), Lugosi & Cesa-Bianchi (2006) and many others:

- $N$ experts
- in round $t$ receive expert’s advice $(f_1^t, f_2^t, \ldots, f_N^t) \in \{0, 1\}^N$.
- $x_t$’s and $y_t$’s can be arbitrary
- How many more mistakes than the best expert do we make?
- $\sqrt{T \log N}$ more (so called regret)
Valiant’s PAC model


• $x_t$ is drawn from a fixed (but arbitrary) probability distribution $P$ over $X$.
• target $h^*$ in class $\mathcal{H}$
• $\hat{y}_t = h^*(x_t)$ (realizable case)
• How many mistakes do we make?
• $\text{VCdim}(\mathcal{H}) \log T$ mistakes
Agnostic PAC model

Haussler (1990), Vapnik and Chervonekis (1971)

- \((x_t, y_t)\) random drawn from a fixed (but arbitrary) probability distribution \(P\) over \(X \times \{0, 1\}\).
- Fixed class \(\mathcal{H}\)
- How many more mistakes than the best hypothesis in \(\mathcal{H}\) do we make?
- \(\sqrt{\text{VCdim}(\mathcal{H})}T\) regret
Our model: Agnostic Online Learning

- *Fixed known* class $\mathcal{H}$
- $x_t$ and $y_t$ are arbitrary
- How many more mistakes than the best hypothesis in $\mathcal{H}$ do we make?
- $\tilde{O}\left(\sqrt{T \text{Ldim}(\mathcal{H})}\right)$ regret

(PAC $\rightarrow$ Agnostic PAC) $\sim$ (Littlestone $\rightarrow$ Agnostic Online)
**Littlestone’s dimension**

\( \mathcal{H} \) **shatters** a full binary tree iff each leaf-hypothesis is **consistent** with the path to the root.

\[ h_1 \quad h_2 \quad h_3 \quad h_4 \quad h_5 \quad h_6 \quad h_7 \quad h_8 \]

\[ x_1 \quad x_2 \quad x_3 \quad x_4 \quad x_5 \quad x_6 \quad x_7 \]

\[ 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \]

\[ 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \quad 1 \]

\[ Ldim(\mathcal{H}) \] is maximum depth of a full binary tree shattered by \( \mathcal{H} \).
Standard Optimal Algorithm (SOA)

Littlestone (1988)

**Initialize:** \( V_0 = \mathcal{H} \)

**For** \( t = 1, 2, \ldots, T \)

receive \( x_t \)

for \( r \in \{0, 1\} \) set \( V_{t-1}^{(r)} = \{h \in V_{t-1} : h(x_t) = r\} \)

predict \( \hat{y}_t = \arg\max_{r \in \{0,1\}} L_{\text{dim}}(V_{t-1}^{(r)}) \)

(if tie, then predict \( \hat{y}_t = 0 \))

receive \( y_t \)

update \( V_t = V_{t-1}^{(y_t)} \)

- \( V_t \) are hypotheses consistent with \((x_1, y_1), \ldots (x_t, y_t)\)
- \( L_{\text{dim}}(V_t) \) decreases at every mistake i.e. when \( \hat{y}_t \neq y_t \)
- Makes at most \( L_{\text{dim}}(\mathcal{H}) \) mistakes in total
Our learning algorithm

- Create \( N = O(T^{L_{\text{dim}}(H_t)}) \) experts
- Use learning with expert advice algorithm
- Total regret

\[
\sqrt{T \log N} = O \left( \sqrt{L_{\text{dim}}(H) T \log T} \right)
\]

to best expert

- Make sure that regret to the best hypothesis is at most regret to the best expert.
Experts

- Total number of experts:
  \[
  \sum_{L=0}^{\text{Ldim}(\mathcal{H})} \binom{T}{L} = \mathcal{O}(T^{\text{Ldim}(\mathcal{H})})
  \]

- One expert for each choice
  \[
  \{i_1, i_2, \ldots, i_L\} \subseteq \{1, 2, \ldots, T\} \quad \text{where} \quad L \leq \text{Ldim}(\mathcal{H})
  \]

- Expert\((i_1, \ldots, i_L)\) simulates SOA on \(x_1, x_2, \ldots, x_T\)
  assuming that it errs in rounds \(i_1, i_2, \ldots, i_L\)
Expert($i_1, \ldots, i_L$)

**Initialize:** $V_0 = \mathcal{H}$

**For** $t = 1, 2, \ldots, T$

receive $x_t$

for $r \in \{0, 1\}$ set $V_{t-1}^{(r)} = \{h \in V_{t-1} : h(x_t) = r\}$

$\hat{y}_t = \arg\max_{r \in \{0, 1\}} \text{Ldim}(V_{t-1}^{(r)})$

(if tie, then $\hat{y}_t = 0$)

**If** $t \in \{i_1, \ldots, i_L\}$

Then predict $f^t = -\hat{y}_t$

Else predict $f^t = y_t$

**update** $V_t = V_{t-1}^{(f^t)}$
Lemma
For each $h \in \mathcal{H}$ and any sequence $x_1, x_2, \ldots, x_T$ there exists an expert, $\text{Expert}(i_1, \ldots, i_L)$, with the same predictions as $h$. That is,

$$f^t = h(x_t) \quad \text{for all } t = 1, 2, \ldots, T.$$

Proof.
Pretend that $h$ is the target. Consider the predictions of SOA on $(x_1, h(x_1)), \ldots, (x_T, h(x_T))$. SOA makes mistakes in rounds $i_1, i_2, \ldots, i_L$ for some $L \leq \text{Ldim}(\mathcal{H})$. $\text{Expert}(i_1, \ldots, i_L)$ predicts $f^t = h(x_t)$. \qed
Corollary

Regret to the best hypothesis is at most the regret to the best expert.

Theorem

For any $\mathcal{H}$ there exists a learning algorithm with regret $O(\sqrt{\text{Ldim}(\mathcal{H})T \log T})$. 
Lower Bound

**Theorem**
For any $\mathcal{H}$ and any learning algorithm there exists a sequence $(x_1, y_1), \ldots, (x_T, y_T)$ such that regret to the best hypothesis in $\mathcal{H}$ is at least $\Omega(\sqrt{L\text{dim}(\mathcal{H})T})$.

**Proof.**
Follow a path in shattered tree. For each node $x$ construct

$$(x, y_1), (x, y_2), \ldots, (x, y_{T/L\text{dim}(\mathcal{H})})$$

where $y$’s are chosen independently uniformly at random. If there exists two $h, h'$ such that $h(x) = 0$ and $h'(x) = 1$, then expected regret is at least $\Omega(\sqrt{T/L\text{dim}(H)})$. Total regret is

$$\Omega \left( L\text{dim}(\mathcal{H}) \cdot \sqrt{T/L\text{dim}(H)} \right) = \Omega \left( \sqrt{L\text{dim}(H)T} \right).$$
Conclusion

Paper:
- www.cs.uwaterloo.ca/~dpal/papers/
- COLT 2009
- fat-shattering and margins
- $y_t$’s are stochastic instead of adversarial

Open problem:

$$\Omega \left( \sqrt{\text{Ldim}(\mathcal{H}) T} \right) \text{ vs. } O \left( \sqrt{\text{Ldim}(\mathcal{H}) T \log T} \right).$$

Thanks!