Agnostic Online Learning

Dávid Pál

March 2009

UNIVERSITY OF
Waterloo

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

Online learning

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

- receive $\boldsymbol{x}_t$ e.g. an email
- predict $\hat{y}_t \in \{0, 1\}$ e.g. $\{\text{spam}, \text{not-spam}\}$
- 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 $h^* : \mathcal{X} \to \{0, 1\}$ in *fixed known* class $\mathcal{H}$
- $\hat{y}_t = h^*(\mathbf{x}_t)$ for all t
  (So called "realizable case".)
- How many mistakes do we make?
- Littlestone defined "optimal mistake bound" of $\mathcal{H}$.
  We call it $\text{Ldim}(\mathcal{H})$ – 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 $(f_1^t, f_2^t, \ldots, f_N^t) \in \{0,1\}^N$.
- $\mathbf{x}_t$'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

Valiant (1984), Haussler, Littlestone & Warmuth (1994)

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

Agnostic PAC model

Haussler (1990), Vapnik and Chervonekis (1971)

- $(\boldsymbol{x}_t, y_t)$ random drawn from a fixed (but arbitrary) probability distribution P over $\mathcal{X} \times \{0, 1\}$.
- Fixed class $\mathcal{H}$
- How many more mistakes than the best hypothesis in $\mathcal{H}$ do we make?
- $\sqrt{\mathrm{VCdim}(\mathcal{H})T}$ regret

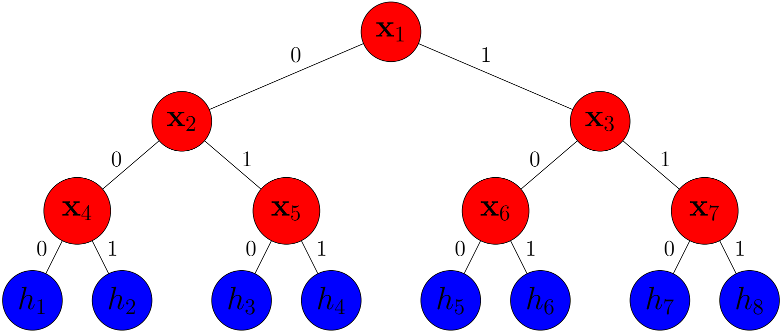
Our model: Agnostic Online Learning

- *Fixed known* class $\mathcal{H}$
- $\mathbf{x}_t$ and y_t are arbitrary
- How many more mistakes than the best hypothesis in $\mathcal{H}$ do we make?
- $\widetilde{O}\left(\sqrt{T\,\mathrm{Ldim}(\mathcal{H})}\right)$ 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.



$\mathrm{Ldim}(\mathcal{H})$ 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 $\boldsymbol{x}_t$
  for $r \in \{0, 1\}$ set $V_{t-1}^{(r)} = \{h \in V_{t-1} \ : \ h(\boldsymbol{x}_t) = r\}$
  predict $\hat{y}_t = \text{argmax}_{r \in \{0,1\}} \text{Ldim}(V_{t-1}^{(r)})$
  (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 $(\boldsymbol{x}_1, y_1), \ldots (\boldsymbol{x}_t, y_t)$
- $\text{Ldim}(V_t)$ decreases at every mistake i.e. when $\hat{y}_t \neq y_t$
- Makes at most $\text{Ldim}(\mathcal{H})$ mistakes in total

Our learning algorithm

- Create $N = O(T^{\mathrm{Ldim}(\mathcal{H})})$ experts
- Use learning with expert advice algorithm
- Total regret

$$\sqrt{T \log N} = O\left(\sqrt{\mathrm{Ldim}(H) T \log T}\right)$$

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

Experts

- Total number of experts:

$$\sum_{L=0}^{\mathrm{Ldim}(\mathcal{H})} \binom{T}{L} = O(T^{\mathrm{Ldim}(\mathcal{H})})$$

- One expert for each choice

$$\{i_1, i_2, \ldots, i_L\} \subseteq \{1, 2, \ldots, T\} \quad \text{where} \quad L \leq \mathrm{Ldim}(\mathcal{H})$$

- Expert$(i_1, \ldots, i_L)$ simulates SOA on $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T$ 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 $\mathbf{x}_t$

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

$\hat{y}_t = \mathrm{argmax}_{r \in \{0,1\}} \mathrm{Ldim}(V_{t-1}^{(r)})$

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

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

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

Else predict $f^t = y_t$

update $V_t = V_{t-1}^{(f^t)}$

Experts

Lemma

For each $h \in \mathcal{H}$ *and any sequence* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_T$ *there exists an expert,* $\text{Expert}(i_1, \ldots, i_L)$*, with the same predictions as* h*. That is,*

$$f^t = h(\mathbf{x}_t) \qquad \textit{for all } t = 1, 2, \ldots, T \ .$$

Proof.

Pretend that h is the target. Consider the predictions of SOA on $(\mathbf{x}_1, h(\mathbf{x}_1)), \ldots (\mathbf{x}_T, h(\mathbf{x}_T))$. 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 $f^t = h(\mathbf{x}_t)$. $\square$

Regret upper bound

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 $O(\sqrt{\mathrm{Ldim}(\mathcal{H}) T \log T})$.

Lower Bound

Theorem

For any $\mathcal{H}$ and any learning algorithm there exists a sequence $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T)$ such that regret to the best hypothesis in $\mathcal{H}$ is at least $\Omega(\sqrt{\mathrm{Ldim}(\mathcal{H})T})$.

Proof.

Follow a path in shattered tree. For each node $\boldsymbol{x}$ construct

$$(\boldsymbol{x}, y_1), (\boldsymbol{x}, y_2), \ldots, (\boldsymbol{x}, y_{T/\mathrm{Ldim}(\mathcal{H})})$$

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

$$\Omega\left(\mathrm{Ldim}(\mathcal{H}) \cdot \sqrt{T/\mathrm{Ldim}(H)}\right) = \Omega\left(\sqrt{\mathrm{Ldim}(H)T}\right) .$$

Conclusion

Paper:

- `www.cs.uwaterloo.ca/~dpal/papers/`
- COLT 2009
- fat-shattering and margins
- y_t's are stochastic instead of adversarial

Open problem:

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

Thanks!