Computational Learning Theory

- Brief Overview of Machine Learning
- Consistency Model
- Probably Approximately Correct Learning
- Sample Complexity and Occam's Razor
- Dealing with Noises and Inconsistent Hypotheses
- **Online Learning and Learing with Expert Advice**
- ...

Relaxing Some Assumptions from PAC

In PAC and the Inconsistent Hypothesis Models, we assumed

- Examples are given in a batch
- There's an "underlying distribution" to learn from and measure output quality

Suppose we relax both of these assumptions: we get *Online Learning*

- Examples are given one at a time, in T steps
- At step t, we're given $\mathbf{x} \in \Omega$, we predict $\mathbf{x}$'s label
- Then, $\mathbf{x}$'s true label is revealed

Main Question: how to measure learner's quality?

Mistake Bound Model

Suppose there's some concept $c \in \mathcal{C}$ from which the "true" labels came

- *Quality of learner* is measured by the number of mistakes M it made in T steps
- **Example**: if $\mathcal{C}$ is the class of *boolean disjunctions*, i.e. target concept c has the form $c = x_{i_1} \vee x_{i_2} \vee \cdots \vee x_{i_q}$, then there's an algorithm learning $\mathcal{C}$ in the mistake bound model with at most n mistakes (n is the number of boolean variables)
- Easy to design a learner making $\leq \log_2 |\mathcal{C}|$ mistakes
  - Take majority vote over all (remaining) consistent $h \in \mathcal{C}$
  - This is called the *halving algorithm*, because if learner makes a mistake then at least half the experts are removed
- (We can do better than the halving algorithm)

But, what if there's no $c \in \mathcal{C}$ consistent with examples?

Learning from Expert Advices

- In this model, think of each $h \in \mathcal{C}$ as an expert.
- At each time step, given $\mathbf{x}$, we get advices from experts on the label of $\mathbf{x}$
- There might not be a "perfect" expert (i.e. consistent with examples)
- Want learner to be as close to the best expert as possible!

"Halving algorithm" is no longer good because the best expert might err in the beginning.
What is learning from expert advices good for?

- In practice, we have many "prediction" algorithms to choose from, but don't know which one is best
- Nice connection to game theory

Weighted Majority Algorithm (WMA)

Idea: trust an expert less if he makes a mistake

- Assign the ith expert a *trustworthiness weight* w_i
- Let $\alpha \in [0, 1]$ be a fixed parameter.

WMA

- Initially, $w_i = 1$ for all $i \in [n]$ (there are n experts)
- At time t,
    - let W_t^v be the total weight of experts who predict value v ($\in \{0, 1\}$)
    - Learner predicts 0 if $W_t^0 \geq W_t^1$ and vice versa
    - After getting the true label, for each i, set $w_i = \alpha w_i$ if he was wrong

(If $\alpha = 0$, we get back the halving algorithm!)

WMA: Analysis

- Suppose WMA makes M mistakes, best expert i_0 makes m mistakes
- For any t, let W_t be total weight at time t.
- Say, WMA makes a mistake at time t. Let W_t^{right} and W_t^{wrong} be total weights of experts who are right and wrong, respectively. Then,

$$
\begin{aligned}
W_t^{\text{wrong}} &\geq \frac{1}{2} W_t \\
W_{t+1} &= \alpha W_t^{\text{wrong}} + W_t^{\text{right}} \leq \left(\frac{1+\alpha}{2}\right) W_t
\end{aligned}
$$

$$
\text{weight of } i_0 \text{ at T} = \alpha^m \leq W_T \leq \left(\frac{1+\alpha}{2}\right)^M W_0
$$

- Since $W_0 = n$,

$$
M \leq \frac{\ln(1/\alpha)}{\ln\left(\frac{2}{1+\alpha}\right)} m + \frac{1}{\ln\left(\frac{2}{1+\alpha}\right)} \ln n
$$

- For example, $\alpha = 1/2$, then $M \leq 2.41m + 3.48 \ln n$.

Randomized WMA

- We want $M \approx m$, but
$$\frac{\ln(1/\alpha)}{\ln\left(\frac{2}{1+\alpha}\right)} \geq 2.$$
(the function is decreasing for $\alpha \in (0,1)$, and the limit as $\alpha \to 1$ is 2)
- Thus, if best expert has 25% error rate, then (the bound for) WMA is only as good as random guessing

Randomized Weighted Majority Algorithm

- Initially, $w_i = 1$ for all $i \in [n]$
- At time t,
  - Learner predicts 0 (1) with probability $\frac{W_t^0}{W_t}$ $(\frac{W_t^1}{W_t})$
  - After getting the true label, for each i, set $w_i = \alpha w_i$ if he was wrong

RWMA: Analysis

- $p_t = \frac{W_t^{\text{wrong}}}{W_t}$ is the probability RWMA guessed wrong at time t
- M is now a random variable, $\mathsf{E}[M] = \sum_t p_t$

$$W_{t+1} = \alpha W_t^{\text{wrong}} + W_t^{\text{right}} = W_t \left(1 - (1-\alpha) p_t\right)$$

- Thus,

$$\begin{aligned} \alpha^m \leq W_T &= W_0 \prod_{t=1}^{T-1} \left(1 - (1-\alpha)p_t\right) \\ &\leq n \prod_{t=1}^{T-1} e^{-(1-\alpha)p_t} \\ &= n e^{-(1-\alpha)\mathsf{E}[M]} \end{aligned}$$

- Hence,

$$\mathsf{E}[M] \leq \frac{\ln(1/\alpha)}{1-\alpha} m + \frac{1}{1-\alpha} \ln n.$$

RWMA: Observations

- When $\alpha = 1/2$, $\mathsf{E}[M] \le 1.39m + 2\ln n$ (much better than WMA)
- $\frac{\ln(1/\alpha)}{1-\alpha} m \ge m$ and $\to m$ as $\alpha \to 1$, but, $\frac{1}{1-\alpha}\ln n \to \infty$ as $\alpha \to 1$
  Need to choose α to balance these two.
- First, since $\ln(1-x) > -x - x^2$ when $x > -1$, we have

  $$\ln(1/\alpha) = -\ln(1-(1-\alpha)) < (1-\alpha) + (1-\alpha)^2$$

  implying

  $$\mathsf{E}[M] < m + (1-\alpha)m + \frac{1}{1-\alpha}\ln n.$$

- Suppose we know $m \le \bar{m}$. WLOG, assume $\bar{m} \ge \ln n$.
- Choose $1 - \alpha = \sqrt{\frac{\ln n}{\bar{m}}}$ to balance things out:

  $$\mathsf{E}[M] < m + 2\sqrt{\bar{m}\ln n}.$$

RWMA: Observations

- If best expert makes at most a constant fraction r of errrors over time, i.e. $m, \bar{m} \approx rT$, then

$$\lim_{T\to\infty} \frac{\mathsf{E}[M]}{T} \leq \lim_{T\to\infty} \left(r + 2\sqrt{r\frac{\ln n}{T}} \right) = r$$

  So the algorithm RWMA converges to optimality with rate $O\left(1/\sqrt{T}\right)$

A Slightly Different View of Learning from Experts

- At time t, the n experts give advices $\mathbf{x} = (x_1, \dots, x_n)$, $x_i \in \{-1, 1\}$ (instead of $\{0, 1\}$, for mathematical convenience)
- We try to find a "expert weight function" $\mathbf{w} \in \mathbb{R}^n$ such that our prediction is

$$\text{sign}(\mathbf{w}^T \mathbf{x}) = \text{sign}(w_1 x_1 + \cdots + w_n x_n).$$

  ($\text{sign}(\alpha) = 1$ if $\alpha > 0$ and $\text{sign}(\alpha) = -1$ if $\alpha \le 0$.)
- The problem is the same as finding a hyperplane separating T n-dimensional data points into the $+1$-class and the -1-class.

Rosenblatt's Perceptron Algorithm

WLOG, we will assume that $\|\mathbf{x}\| = 1$ for each (advice) vector $\mathbf{x}$, since normalizing $\mathbf{x}$ does not change the side of the hyperplane $\mathbf{x}$ is on.

- Set $\mathbf{w}_0 = 0$
- At time t,
  - Given (advices) $\mathbf{x}$, predict $+1$ iff $\mathbf{w}_t^T\mathbf{x} > 0$
  - Suppose the true label is y_t ($\in \{1, -1\}$)
  - If we predicted $+1$ but $y_t = -1$, set $\mathbf{w}_{t+1} = \mathbf{w}_t + \mathbf{x}$
  - If we predicted -1 but $y_t = +1$, set $\mathbf{w}_{t+1} = \mathbf{w}_t - \mathbf{x}$

Why is it reasonable?

- If we predicted $+1$ but $y_t = -1$, then

$$\mathbf{w}_{t+1}^T\mathbf{x} = (\mathbf{w}_t + \mathbf{x})^T\mathbf{x} = \mathbf{w}_t^T\mathbf{x} + 1$$

- If we predicted -1 but $y_t = +1$, then

$$\mathbf{w}_{t+1}^T\mathbf{x} = (\mathbf{w}_t - \mathbf{x})^T\mathbf{x} = \mathbf{w}_t^T\mathbf{x} - 1$$

- Either way, $\mathbf{w}_{t+1}^T\mathbf{x}$ moves in the right direction

Analysis

Theorem

Let S be a set of labeled examples. Suppose there exists a good separating hyperplane, i.e. there exists a unit-length $\mathbf{w}^* \in \mathbb{R}^n$ such that $\langle \mathbf{w}^*, \mathbf{x} \rangle > 0$ for all positive examples and $\langle \mathbf{w}^*, \mathbf{x} \rangle < 0$ for all negative examples. Then, the number of mistakes M made by the Perceptron algorithm is at most $(1/\delta)^2$, where

$$\delta = \min_{\mathbf{x} \in S} |\langle \mathbf{w}^*, \mathbf{x} \rangle|.$$

(Recall that $\|x\| = 1$ for all examples $\mathbf{x}$.)

Proof of the Theorem

Fact 1: if we made a mistake at time t, then

$$\langle \mathbf{w}_{t+1}, \mathbf{w}^* \rangle \geq \langle \mathbf{w}_t, \mathbf{w}^* \rangle + \delta.$$

(That is, in some sense the angle between $\mathbf{w}_{t+1}$ and $\mathbf{w}^*$ is smaller, unless $\mathbf{w}_{t+1}$ gets really long compared to $\mathbf{w}_t$. However, the next fact says that it won't be too long compared to $\mathbf{w}_t$.)

Fact 2: if we made a mistake at time t, then

$$\|\mathbf{w}_{t+1}\|^2 \leq \|\mathbf{w}_t\|^2 + 1.$$

- Thus, after M mistakes, by Fact 1 we know $\langle \mathbf{w}_T, \mathbf{w}^* \rangle \geq \delta M$; and by Fact 2 we conclude $\|\mathbf{w}_T\| \leq \sqrt{M}$.
- Thus, $\delta M \leq \langle \mathbf{w}_T, \mathbf{w}^* \rangle \leq \|\mathbf{w}_T\| \leq \sqrt{M}$. Done!

Later