COS 511: Foundations of Machine Learning

Rob Schapire
Scribe: Jeffrey Traer Bernstein

Lecture #17
April 11, 2006

1 Proof of Widrow-Hoff continued...

Recap from last time:

$w_i = 0$

for $t = 1 \ldots T$ trials

get $\mathbf{x}_t \in \mathbb{R}^n$, a vector of expert predictions

predict $\hat{y}_t = \mathbf{w}_t \cdot \mathbf{x}_t$

observe $y_t \in \mathbb{R}$

update

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \underbrace{\eta(\overbrace{\mathbf{w}_t \cdot \mathbf{x}_t}^{\hat{y}_t} - y_t)x_t}_{\Delta_t}$$

loss is $(\hat{y}_t - y_t)^2$

We were in the middle of proving a theorem

$$\text{if} \quad \forall t \ ||\mathbf{x}_t||_2 \leq 1 \quad \text{then} \quad L_A \leq \min_{\mathbf{u} \in \mathbb{R}^n} \left[\frac{L_{\mathbf{u}}}{1-\eta} + \frac{||\mathbf{u}||_2^2}{\eta} \right]$$

where

$$L_A = \sum_{t=1}^{T} (\hat{y}_y - y_t)^2 \quad \text{and} \quad L_{\mathbf{u}} = \sum_{t=1}^{T} (\mathbf{u} \cdot \mathbf{x}_t - y_t)^2$$

with potential function $\Phi_t = ||\mathbf{w}_t - \mathbf{u}||_2^2$

$l_t = \mathbf{w}_t \cdot \mathbf{x}_t - y_t$ learner's loss is l_t^2

$g_t = \mathbf{u} \cdot \mathbf{x}_t - y_t$ u's loss is g_t^2

The main thing to prove is:

$$\Phi_{t+1} - \Phi_t \leq -\eta l_t^2 + \frac{\eta}{1-\eta} g_t^2$$

We showed last time that this is enough to prove the theorem.
To make things easier let's say $\mathbf{w} = \mathbf{w}_t$ and $\mathbf{w}' = \mathbf{w}_{t+1}$

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= ||\mathbf{w}' - \mathbf{u}||^2 - ||\mathbf{w} - \mathbf{u}||^2 && \text{just the definition of } \Phi \\
&= ||\mathbf{w} - \mathbf{u} - \Delta||^2 - ||\mathbf{w} - \mathbf{u}||^2 && \text{from the update rule above} \\
&= ||\mathbf{w} - \mathbf{u}||^2 - 2(\mathbf{w} - \mathbf{u}) \cdot \Delta + ||\Delta||^2 - ||\mathbf{w} - \mathbf{u}||^2
\end{aligned}$$

$$
\begin{aligned}
& & & \text{because } ||\mathbf{x}||^2 = \mathbf{x}\cdot\mathbf{x} \\
&= -2\underbrace{(\mathbf{w}-\mathbf{u})\cdot\Delta}_{\eta l(\mathbf{w}\cdot\mathbf{x}-\mathbf{u}\cdot\mathbf{x})} + \underbrace{||\Delta||^2}_{\eta^2 l^2 ||\mathbf{x}||^2 \le \eta^2 l^2} & & \\
&\le -2\eta l(l-g) + \eta^2 l^2 & & ||\mathbf{x}||^2 \le 1 \text{ by assumption} \\
& & & \text{and } \underbrace{\mathbf{w}\cdot\mathbf{x}-y}_{l} + \underbrace{y-\mathbf{u}\cdot\mathbf{x}}_{-g} \\
&= (\eta^2-2\eta)l^2 + 2\eta \underbrace{lg}_{\le \frac{1}{2}[\frac{g^2}{1-\eta}+l^2(1-\eta)]\star} & & \text{regroup a little} \\
&\le -\eta l_t^2 + \tfrac{\eta}{1-\eta} g_t^2 & & \text{all cleaned up}
\end{aligned}
$$

$\star$ For the last step we used a little trick: $\sqrt{ab} \le \frac{a+b}{2}$. The easy thing to do is just pick $a = g^2$ and $b = l^2$ but we are going to add in a constant that disappears when we multiply in the $\sqrt{ab}$ part so we can make the end result cleaner. So we pick $a = \frac{g^2}{1-\eta}$ and $b = l^2(1-\eta)$. To figure these out you could just plug in a generic constant and then solve afterwards for what yields the best result.

2 How do you derive an update rule?

Where did that update rule come from? We can derive an update by minimizing an expression of the form:

$$\eta\Big(\text{loss of } \mathbf{w}_{t+1} \text{ on } \mathbf{x}_t \text{ and } y_t\Big) + \Big(\text{distance between weight vectors } \mathbf{w}_t \text{ and } \mathbf{w}_{t+1}\Big)$$

You minimize this with respect to $\mathbf{w}_{t+1}$ to get a new update rule for some other distance measure. You can also use this distance function in the analysis. We used the euclidian distance squared:

$$\eta L(\mathbf{w}_{t+1}, \mathbf{x}_t, y_t) + \underbrace{||\mathbf{w}_{t+1}-\mathbf{w}_t||_2^2}_{\text{EUCLIDIAN DISTANCE}^2}$$

Minimizing this expression gives us a gradient-descent update:

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \underbrace{\nabla L(\mathbf{w}_{t+1}, \mathbf{x}_t, y_t)}_{\text{GRADIENT}}$$

You may notice the $\mathbf{w}_{t+1}$ on both sides of the equation. To make our lives easier we can approximate $\mathbf{w}_{t+1}$ on the right hand side with $\mathbf{w}_t$ in practice.
Now, we could, for example, substitute relative entropy as our distance function:

$$\eta L(\mathbf{w}_{t+1}, \mathbf{x}_t, y_t) + \text{RE}(\mathbf{w}_t || \mathbf{w}_{t+1})$$

If we now minimize this expression, we get an update rule of the form:

$$w_{t+1,i} = \frac{w_{t,i}\exp(-\eta \frac{\partial L}{\partial w_i}(\mathbf{w}_{t+1}, \mathbf{x}_t, y_t))}{\underbrace{Z_t}_{\text{normalize to make a distribution}}}$$

This algorithm is called EG for Exponentiated Gradient. For square loss we can prove that if $||\mathbf{x}_t||_\infty \le 1$ and $||\mathbf{u}||_1 \le 1$ then $L_{EG} \le \min_{\mathbf{u}} \Big[a_\eta L_{\mathbf{u}} + b_\eta \ln n\Big]$

3 Can we apply online learning to batch?

Wouldn't it be nice to apply the power and majesty of online learning to the batch learning we talked about before? We're going to look at regression but the technique is general and can be applied to classification and other kinds of learning.
As usual for batch we are given

$$S = \langle (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \rangle \qquad \begin{array}{ll} (\mathbf{x}_i, y_i) \sim D & \text{training} \\ (\mathbf{x}, y) \sim D & \text{test} \end{array}$$

Now how do we apply Widrow-Hoff to this? Our goal is to find $\mathbf{v}$ such that $R_{\mathbf{v}} = E_{(\mathbf{x},y)\sim D}[(\mathbf{v}\cdot\mathbf{x} - y)^2]$ is small relative to $\min_{\mathbf{u}} R_{\mathbf{u}}$ where $\mathbf{u}$ is the best weight vector.
A simple plan:

1. run Widrow-Hoff on our training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)$
   get back weight vectors $\mathbf{w}_1, \ldots, \mathbf{w}_m$
2. Output $\mathbf{v} = \frac{1}{m}\sum_t^m \mathbf{w}_t$, i.e. take the average of all the weight vectors

Now we can prove something about this:

$$E_S[R_{\mathbf{v}}] \leq \min_{\mathbf{u}} \Big[\frac{R_{\mathbf{u}}}{1-\eta} + \underbrace{\frac{||\mathbf{u}||_2^2}{\eta m}}_{\text{as } m \to \infty \text{ this term} \to 0} \Big]$$

So the expected risk of this is going to be just a little more risk than $\frac{1}{1-\eta} R_{\mathbf{u}}$ where $R_{\mathbf{u}}$ can be thought of as the best possible risk.

First, we make three observations. In what follows, unless otherwise noted, expectations are over S and (x, y).

1. Look at the loss

$$\begin{aligned} (\mathbf{v}\cdot\mathbf{x} - y)^2 &= \Big((\frac{1}{m}\sum_t \mathbf{w}_t)\cdot\mathbf{x} - y \Big)^2 \\ &= \Big(\frac{1}{m}\sum_t (\mathbf{w}_t\cdot\mathbf{x} - y) \Big)^2 && \text{square of averages} \\ &\leq \frac{1}{m}\sum_{t=1}^m (\mathbf{w}_t\cdot\mathbf{x} - y)^2 && \text{average of squares} \end{aligned}$$

   The last step follows from the definition of convexity and since $f(x) = x^2$ is convex.

2. $E[(\mathbf{u}\cdot\mathbf{x}_t - y_t)^2] = E[(\mathbf{u}\cdot\mathbf{x} - y)^2]$

   This is because $(\mathbf{x}_t, y_t)$ and $(\mathbf{x}, y)$ come from identical distributions.

3. $E[(\mathbf{w}_t\cdot\mathbf{x}_t - y_t)^2] = E[(\mathbf{w}_t\cdot\mathbf{x} - y)^2]$

   This is because $\mathbf{w}_t$ is chosen before $(\mathbf{x}_t, y_t)$ and $(\mathbf{x}, y)$, which are therefore identically distributed given $\mathbf{w}_t$.

Now given these three observations...

$$
\begin{aligned}
E_S[R_{\mathbf{v}}] &= E[(\mathbf{v}\cdot\mathbf{x}-y)^2] \\
&\leq E\Big[\frac{1}{m}\sum_{t=1}^{m}(\mathbf{w}_t\cdot\mathbf{x}-y)^2\Big] && \text{from (1)} \\
&= \frac{1}{m}\sum_t E[(\mathbf{w}_t\cdot\mathbf{x}-y)^2] && \text{move E in by linearity of expectations} \\
&= \frac{1}{m}\sum_t E[(\mathbf{w}_t\cdot\mathbf{x}_t-y_t)^2] && \text{by (3)} \\
&= \frac{1}{m}E\Big[\sum_t(\mathbf{w}_t\cdot\mathbf{x}_t-y_t)^2\Big] && \text{move E back out}
\end{aligned}
$$

We are using Widrow-Hoff and we have a bound for this!

$$
\begin{aligned}
&\leq \frac{1}{m}E\Big[\frac{\sum_t(\mathbf{u}\cdot\mathbf{x}_t-y_t)^2}{1-\eta}+\frac{||\mathbf{u}||_2^2}{\eta}\Big] && \text{plug in the W-H bound} \\
&\leq \frac{1}{m}\Big[\frac{\sum_t E[(\mathbf{u}\cdot\mathbf{x}_t-y_t)^2]}{1-\eta}+\underbrace{\frac{||\mathbf{u}||_2^2}{\eta}}_{\text{CONSTANT}}\Big] && \text{move E} \\
&\leq \frac{1}{m}\Big[\frac{\sum_t E[(\mathbf{u}\cdot\mathbf{x}-y)^2]}{1-\eta}+\frac{||\mathbf{u}||_2^2}{\eta}\Big] && \text{get rid of the } t\text{'s, from (2)} \\
&= \frac{1}{m}\Big[\frac{\sum_t R_{\mathbf{u}}}{1-\eta}+\frac{||\mathbf{u}||_2^2}{\eta}\Big] \\
&= \frac{R_{\mathbf{u}}}{1-\eta}+\frac{||\mathbf{u}||_2^2}{\eta m} && \text{and we're done!}
\end{aligned}
$$