COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Eric Goldlust

Lecture #17
April 9, 2008

Bound on the Loss of the Widrow-Hoff algorithm

Last time, we were analyzing the performance of the Widrow-Hoff algorithm, specified as follows:

```
$\mathbf{w}_1 = \mathbf{0}$
for $t = 1$ to $T$ do
    get example $\mathbf{x}_t \in \mathbb{R}^n$
    predict response $\hat{y}_t = \mathbf{w}_t \cdot \mathbf{x}_t$
    observe response $y_t \in \mathbb{R}$
    set $\mathbf{w}_{t+1} = \mathbf{w}_t - \eta(\mathbf{w}_t \cdot \mathbf{x}_t - y_t)\mathbf{x}_t$
end for
```

We defined the loss under Widrow-Hoff to be: $L_{\text{WH}} = \sum_{t=1}^{T}(\hat{y}_t - y_t)^2$ And the loss under a *fixed* vector $\mathbf{u}$ to be $L_{\mathbf{u}} = \sum_{t=1}^{T}(\mathbf{u} \cdot \mathbf{x}_t - y_t)^2$ and partially proved the following theorem about L_{WH}, which does not rely on any distributional assumptions about the data.

Theorem. *Assume that for every t, we know that $||\mathbf{x}_t||_2 \leq 1$. Then:*

$$L_{WH} \leq \min_{\mathbf{u}} \left[\frac{L_{\mathbf{u}}}{1-\eta} + \frac{||\mathbf{u}||_2^2}{\eta} \right].$$

In the previous lecture, we reduced the proof of this theorem to the proof of the following lemma, which we now prove:

Lemma. *Let $||\mathbf{x}_t||_2 \leq 1$. Define the potential function $\Phi_t = ||\mathbf{w}_t - \mathbf{u}||_2^2$. Define the time-$t$ signed error of Widrow-Hoff by $\ell_t = \mathbf{w}_t \cdot \mathbf{x}_t - y_t$ and the signed error from a fixed $\mathbf{u}$ by $g_t = \mathbf{u} \cdot \mathbf{x}_t - y_t$. Then, for every t and $\mathbf{u}$:*

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

Proof. Let $\mathbf{\Delta}_t = \eta \ell_t \mathbf{x}_t$

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &= ||\mathbf{w}_{t+1} - \mathbf{u}||_2^2 - ||\mathbf{w}_t - \mathbf{u}||_2^2 \\
&= ||(\mathbf{w}_t - \mathbf{u}) - \mathbf{\Delta}_t||_2^2 - ||\mathbf{w}_t - \mathbf{u}||_2^2 \\
&= ||\mathbf{\Delta}_t||_2^2 - 2(\mathbf{w}_t - \mathbf{u}) \cdot \mathbf{\Delta}_t \\
&= \eta^2 \ell_t^2 \underbrace{||\mathbf{x}_t||_2^2}_{\leq 1} - 2\eta\ell_t \underbrace{\mathbf{x}_t \cdot (\mathbf{w}_t - \mathbf{u})}_{=\ell_t - g_t} \\
&\leq \eta^2 \ell_t^2 - 2\eta \ell_t^2 + 2\eta \ell_t g_t \\
&= \eta^2 \ell_t^2 - 2\eta \ell_t^2 + 2\eta \left[\left(\ell_t \sqrt{1-\eta} \right) \left(\frac{g_t}{\sqrt{1-\eta}} \right) \right].
\end{aligned}$$

We next use the real algebraic inequality[1] $ab \leq (a^2 + b^2)/2$ for the case where $a = \ell_t\sqrt{1-\eta}$ and $b = \frac{g_t}{\sqrt{1-\eta}}$. This gives

$$\begin{aligned}
\Phi_{t+1} - \Phi_t &\leq \eta^2\ell_t^2 - 2\eta\ell_t^2 + \eta(1-\eta)\ell_t^2 + \frac{\eta}{1-\eta}g_t^2 \\
&= [\eta^2 - 2\eta + \eta(1-\eta)]\ell_t^2 + \frac{\eta}{1-\eta}g_t^2 \\
&= -\eta\ell_t^2 + \frac{\eta}{1-\eta}g_t^2
\end{aligned}$$

which completes the proof of the lemma, and in turn the theorem. □

Generalization: Varying the Loss function and the Norm

When we originally derived the Widrow-Hoff update rule, we tried to find a value of $\mathbf{w}_{t+1}$ that minimized a linear combination of the loss of $\mathbf{w}_{t+1}$ on $(\mathbf{x}_t, y_t)$ and the norm $||\mathbf{w}_{t+1} - \mathbf{w}_t||_2^2$. Specifically, we wanted to minimize $\eta(\mathbf{w}_{t+1} \cdot \mathbf{x}_t - y_t)^2 + ||\mathbf{w}_{t+1} - \mathbf{w}_t||_2^2$. We can try to generalize this objective function by replacing either the loss term or the norm term with a more general function.

General loss function, L_2 distance

If our objective function is given by:

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

Then we get the "Gradient Descent" (GD) update rule:

$$\begin{aligned}
\mathbf{w}_{t+1} &= \mathbf{w}_t - \eta\nabla_{\mathbf{w}}L(\mathbf{w}_{t+1}, \mathbf{x}_t, y_t) \\
&\approx \mathbf{w}_t - \eta\nabla_{\mathbf{w}}L(\mathbf{w}_t, \mathbf{x}_t, y_t).
\end{aligned}$$

Square loss, Relative Entropy distance

If our objective function is given by:

$$\eta(\mathbf{w}_t \cdot \mathbf{x}_t - y_t)^2 + RE(\mathbf{w}_t||\mathbf{w}_{t+1})$$

then we find the following update rule, which is specified component-wise:

$$\begin{aligned}
\omega_{t+1,i} &= \frac{\omega_{t,i}\exp\{-\eta(\mathbf{w}_{t+1} \cdot \mathbf{x}_t - y_t)x_{t,i}\}}{Z_t} \\
&\approx \frac{\omega_{t,i}\exp\{-\eta(\mathbf{w}_t \cdot \mathbf{x}_t - y_t)x_{t,i}\}}{Z_t}
\end{aligned}$$

where Z_t are normalization factors. Note the parallel with the original Widrow-Hoff rule: In that case, we added $-\mathbf{\Delta}_t$ to $\mathbf{w}_t$. In this case, we multiply componentwise with the exponentiation of the components of $-\mathbf{\Delta}_t$ and then normalize.

[1] Follows from $a^2 - 2ab + b^2 = (a-b)^2 \geq 0$.

General loss function, Relative Entropy distance

If our objective function is given by:

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

then we get the "Exponentiated Gradient" (EG) update rule:

$$\begin{aligned} \omega_{t+1,i} &= \frac{\omega_{t,i} \exp\left\{-\eta \frac{\partial L}{\partial \omega_i}(\mathbf{w}_{t+1}, \mathbf{x}_t, y_t)\right\}}{Z_t} \\ &\approx \frac{\omega_{t,i} \exp\left\{-\eta \frac{\partial L}{\partial \omega_i}(\mathbf{w}_t, \mathbf{x}_t, y_t)\right\}}{Z_t}. \end{aligned}$$

There is a performance bound for the EG update rule that looks similar to the one we proved about WH. In order to make an apples-to-apples comparison, we first rewrite the original WH bound as follows:

$$||\mathbf{x}_t||_2 \leq 1 \Rightarrow L_{\mathrm{WH}} \leq \min_{\mathbf{u}:||\mathbf{u}||_2=1} [aL_{\mathbf{u}} + b] \text{ for some } a,\ b.$$

For EG with square loss, there is the following similar-looking bound:

$$||\mathbf{x}_t||_\infty \leq 1 \Rightarrow L_{EG} \leq \min_{\mathbf{u}:||\mathbf{u}||_1=1} [aL_{\mathbf{u}} + b \ln N] \text{ for some } a,\ b.$$

We can now add an element to our recurring dichotomy between additive and multiplicative algorithms:

additive updates	multiplicative updates
L_2/L_2	L_∞/L_1
Support Vector Machines	AdaBoost
Perceptrons	Winnow
Gradient Descent / Widrow-Hoff	Exponentiated Gradient

Using Online Algorithms in a Batch Setting

We have analyzed both the batch setting and the online setting. In some sense, the results in the online setting are stronger because they do not rely on statistical assumptions about the data (for example that the examples are i.i.d). We now analyze the use of these online algorithms on batch data where these statistical assumptions are assumed to hold. The result will be simple and fast algorithms with generalization bounds that come for free from the analysis we did in the online setting.

The Batch Setting

Given $\mathcal{S} = \langle(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\rangle$, assume that for any i, $(\mathbf{x}_i, y_i) \sim D$. Let there be a new test point $(\mathbf{x}, y)$ that is also distributed according to D. We want to find a linear predictor $\mathbf{v}$ with a low expected loss (also called "risk" or "true risk"), defined by:

$$R_{\mathbf{v}} = \mathbb{E}_{(\mathbf{x},y)\sim D}\left[(\mathbf{v} \cdot \mathbf{x} - y)^2\right].$$

When we say that we want the expected loss to be low, we mean this in relation to that of the best possible value for any $\mathbf{u}$, i.e. it should be low relative to $\min_{\mathbf{u}} R_{\mathbf{u}}$.

One reasonable way to accomplish this is to use Widrow-Hoff on the dataset, treating it as though the examples were arriving online. This would yield a sequence of weight vectors $\mathbf{w}_1 = \mathbf{0}, \mathbf{w}_2, \ldots, \mathbf{w}_m$. We could then return the final hypothesis $\mathbf{v} = \mathbf{w}_m$. Unfortunately, if we return $\mathbf{w}_m$, the analysis turns out to be too difficult. We can, however, analyze the performance if we return $\mathbf{v} = \frac{1}{m}\sum_{t=1}^{m} \mathbf{w}_t$. In this case, we can prove the following theorem:

Theorem.

$$\mathbb{E}_{\mathcal{S}}\left[R_{\mathbf{v}}\right] \leq \min_{\mathbf{u}} \left[\frac{R_{\mathbf{u}}}{1-\eta} + \underbrace{\frac{||\mathbf{u}||_2^2}{\eta m}}_{\to 0 \textit{ as } m \to \infty}\right]$$

Here, the expectation is over the random training set $\mathcal{S}$. We proceed in steps, starting with three lemmas.

Lemma (1).

$$(\mathbf{v} \cdot \mathbf{x} - y)^2 \leq \frac{1}{m}\sum_{t=1}^{m}(\mathbf{w}_t \cdot \mathbf{x} - y)^2.$$

Proof.

$$\begin{aligned}
(\mathbf{v} \cdot \mathbf{x} - y)^2 &= \left(\left(\frac{1}{m}\sum_{t=1}^{m} \mathbf{w}_t\right) \cdot \mathbf{x} - y\right)^2 \\
&= \left(\frac{1}{m}\sum_{t=1}^{m}\left(\mathbf{w}_t \cdot \mathbf{x} - y\right)\right)^2 \\
&\leq \frac{1}{m}\sum_{t=1}^{m}\left(\mathbf{w}_t \cdot \mathbf{x} - y\right)^2 \text{ (by convexity of } f(z) = z^2\text{).}
\end{aligned}$$

□

Lemma (2).

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

Proof. This statement is true because $(\mathbf{x}_t, y_t)$ and $(\mathbf{x}, y)$ are identically distributed. □

Lemma (3).

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

Proof. $\mathbf{w}_t$ is chosen before $(\mathbf{x}_t, y_t)$, so $(\mathbf{x}_t, y_t)$ and $(\mathbf{x}, y)$ are identically distributed given $\mathbf{w}_t$. Note that this is not true, for example, given $\mathbf{w}_{t+1}$. □

We are now ready to prove the theorem.

Proof. For any fixed $\mathbf{u}$, we have:

$$\begin{aligned}
\mathbb{E}_{\mathcal{S}}\left[R_{\mathbf{v}}\right] &= \mathbb{E}\left[(\mathbf{v}\cdot\mathbf{x}-y)^2\right] \\
&\leq \mathbb{E}\left[\frac{1}{m}\sum_{t=1}^{m}(\mathbf{w}_t\cdot\mathbf{x}-y)^2\right] \text{ (by lemma 1)} \\
&= \frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\left[(\mathbf{w}_t\cdot\mathbf{x}-y)^2\right] \\
&= \frac{1}{m}\sum_{t=1}^{m}\mathbb{E}\left[(\mathbf{w}_t\cdot\mathbf{x}_t-y_t)^2\right] \text{ (by lemma 3)} \\
&= \frac{1}{m}\mathbb{E}\left[\sum_{t=1}^{m}(\mathbf{w}_t\cdot\mathbf{x}_t-y_t)^2\right] \\
&= \frac{1}{m}\mathbb{E}\left[L_{\mathrm{WH}}\right] \\
&\leq \frac{1}{m}\mathbb{E}\left[\frac{\sum_{t=1}^{m}(\mathbf{u}\cdot\mathbf{x}_t-y_t)^2}{1-\eta}+\frac{||\mathbf{u}||_2^2}{\eta}\right] \text{ (previously shown in WH analysis)} \\
&= \frac{1}{m}\left[\frac{\sum_{t=1}^{m}\mathbb{E}\left[(\mathbf{u}\cdot\mathbf{x}_t-y_t)^2\right]}{1-\eta}+\frac{||\mathbf{u}||_2^2}{\eta}\right] \\
&= \frac{1}{m}\left[\frac{\sum_{t=1}^{m}\mathbb{E}\left[(\mathbf{u}\cdot\mathbf{x}-y)^2\right]}{1-\eta}+\frac{||\mathbf{u}||_2^2}{\eta}\right] \text{ (by lemma 2)} \\
&= \left[\frac{R_{\mathbf{u}}}{1-\eta}+\frac{||\mathbf{u}||_2^2}{\eta m}\right].
\end{aligned}$$

Since this holds for any $\mathbf{u}$, the theorem follows. □

Probability Modeling

So far, we have analyzed situations where we deal with $(\mathbf{x}, y)$ pairs, where y could be real or categorical. We now consider the situation where we receive only $\mathbf{x}$ and our goal is to model its distribution. Let $\mathbf{x} \sim P$. The goal is to estimate P. This task is called "Probability Modeling" or "Density Estimation".

One example of where this can be useful is in speech recognition. In order to decide if a speaker has just said "I sat on a chair" or "I fat on a chair," a system could use prior estimates of the relative likelihood of these two phrases in English in order to decide that "I sat on a chair" was more likely.

We might also want to perform density estimation for classification problems. If we wanted to estimate human gender based on height, we could build separate density estimates (possibly just by estimating gaussian parameters) for men and women and then use Bayes' rule to decide which gender was more likely under that probability model given the height of the test person. This is called the **generative approach**.

By contrast, the **discriminative approach** makes no attempt to model the distribution of the data, but rather just tries to find an effective classification rule. In this case, we would just estimate a threshold height above which to classify a test person as male.

An advantage of the discriminative approach is that it is direct and simple and does not require assumptions about the distribution of the data. The generative approach, however, has the advantage that expert knowledge of the distribution can sometimes lead to higher performance with less training data.

One example where the generative approach is effective is in the detection of fraudulent calling-card phone calls. One could build, for every customer, a probability distribution over phone calls (frequency, duration, location, etc), and then build similar models for fraudsters. Given a test phone call, one could use the probabilities under these two distributions to estimate the probability that the call is fraudulent. In this case, the generative approach can work well.