COS 511: Theoretical Machine Learning

Lecturer: Rob Schapire
Scribe: Andrej Risteski

Lecture #16
April 4, 2013

1 Introduction/Recap from last time

In this lecture, we continue with a modification of the online learning scenario from last class. As a reminder, in the scenario, there is a binary event happening each day for a sequence of days. We also have a set of experts, which give predictions each day, and we guess a prediction according to all the previous information. After our prediction, we also observe what the outcome actually was.

The previous way of scoring how well we did is according to the best expert - the one who got the correct outcome most of the times. However, we could easily imagine a scenario where no one expert is good. But, if we form a "committee" of experts, they might be much better. The way we'll formalize this is as follows:

- We have N experts.
- For $t = 1, \ldots, T$ rounds, we get a $\mathbf{x}_t \in \{1, -1\}^N$ - a vector of predictions from the experts.
- In each round, the learner learns an outcome y_t after it makes its prediction $\hat{y}_t$.
- We assume that there is a "perfect committee" - i.e. weighted sum of experts, that are always right. Formally, this means that there exists a $\mathbf{u} \in R^N$, such that for all t,

$$sign(\sum_{i=1}^{N} u_i x_{t,i}) = sign(\mathbf{x}_t \cdot \mathbf{u}) = y_t$$

Geometrically, the perfect committee assumption just means that there is a separating hyperplane between the 1 and -1 points, generated by the appropriate weighted sum of the experts.

2 How to do updates

To get an algorithm for the above problem, we will do the following. We will maintain $\mathbf{w}_t$, a sort of a "guess" of the correct weighting of the experts. We will update the weighting on each round.

2.1 Perceptron

The first way to update the weights will give us an algorithm called the *perceptron*. The update rules are as follows:

- $\mathbf{w}_1 = \mathbf{0}$
- If we predict wrong on a given point (i.e. $y_t \neq \hat{y}_t$), set $\mathbf{w}_{t+1} = \mathbf{w}_t + y_t \cdot \mathbf{x}_t$. Otherwise, do nothing (i.e. $\mathbf{w}_{t+1} = \mathbf{w}_t$). This makes the algorithm *conservative*.

The intuition is that in case of a wrong answer we "shift" the weights on all the experts in the direction of the correct answer. Geometrically, this is represented in figure 1 below. We have the vector $\mathbf{w}_t$ of weights and the vector $\mathbf{x}_t$ of expert opinions. When we define $\mathbf{w}_{t+1}$, we shift $\mathbf{w}_t$ such that the angle between $\mathbf{x}_t$ and $\mathbf{w}_t$ decreases if $y_t = 1$, and such that it increases otherwise.

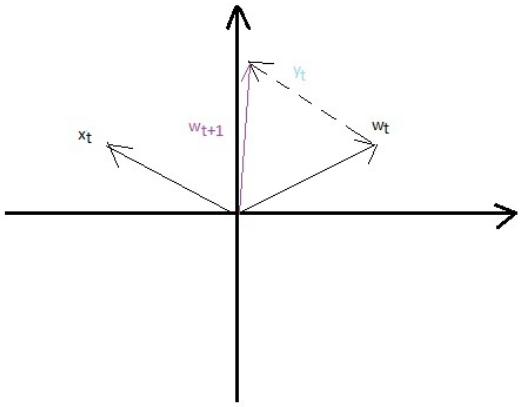

Figure 1: Perceptron geometric intuition

An alternative, more quantitative view of the algorithm is the following. Notice that $y_t(\mathbf{w}_t \cdot \mathbf{x}_t) \leq 0$ iff $y_t \neq \hat{y}_t$. Then, if we write the left hand side with $\mathbf{w}_{t+1}$ instead, we get

$$y_t(\mathbf{w}_{t+1} \cdot \mathbf{x}_t) = y_t((\mathbf{w}_t + y_t\mathbf{x}_t) \cdot \mathbf{x}_t)$$

$$= y_t\mathbf{w}_t \cdot \mathbf{x}_t + {y_t}^2\mathbf{x}_t \cdot \mathbf{x}_t$$

But now, notice that $y_t^2 = 1$, since $y_t = \pm 1$, and $\mathbf{x}_t \cdot \mathbf{x}_t = ||x_t||_2^2$, which is non-negative. What this chain of inequalities shows is that if we update the weights to $\mathbf{w}_{t+1}$ we are likely to fix the algorithm's prediction on the previously wrong point.

Now, we will state a theorem to formally analyze the performace of the perceptron algorithm. But, first, we will make a few simplifying assumptions, all without loss of generality:

- We will normalize the vector of predictions $\mathbf{x}_t$, so that $||\mathbf{x}_t||_2 \leq 1$.
- We will normalize the vector of weights for the perfect committee, so that $||\mathbf{u}||_2 = 1$. (This is fine in both cases since the sign function won't depend on multiplying the entire vector $\mathbf{u}$ or $\mathbf{x}_t$ by a constant.)
- We will assume the algorithm makes a mistake in each round. We can do this, since no algorithmic changes happen during other rounds.

Theorem 1. *Assume we are in the online learning scenario described at the beginning of these notes, along with all of the assumptions above. Furthermore, assume that something slightly stronger holds for the "committee" of best experts: for all t, $(\mathbf{u} \cdot \mathbf{x}_t)y_t \geq \delta$ for some $\delta > 0$ - i.e. the margin of correctness of the committee is bounded away from 0. Then, the number of mistakes the perceptron algorithm makes is at most $\frac{1}{\delta^2}$.*

Proof. As usual, the way we do this, is we find some quantity that depends on the state of the algorithm at time t, upper and lower bound it, and derive a bound from there. The quantity here is Φ_t, which we define as the cosine of the angle between $\mathbf{w}_t$ and $\mathbf{u}$. More formally, we put $\Phi_t = \frac{\mathbf{w}_t \cdot \mathbf{u}}{||\mathbf{w}_t||_2}$. For the upper bound, since this is a cosine, clearly, $\Phi_t \leq 1$.

Now, for the lower bound. We will prove that $\Phi_T \geq \sqrt{T}\delta$. We do this in two parts, by lower bounding the numerator of Φ_t, and upper bounding the denominator.

We prove that $\mathbf{w}_{t+1} \cdot \mathbf{u} \geq T\delta$. We have that

$$\mathbf{w}_{t+1} \cdot \mathbf{u} = (\mathbf{w}_t + y_t \mathbf{x}_t) \cdot \mathbf{u} = \mathbf{w}_t \cdot \mathbf{u} + y_t(\mathbf{u} \cdot \mathbf{x}_t) \geq \mathbf{w}_t \cdot \mathbf{u} + \delta$$

The equalities just come from the definition of $\mathbf{w}_{t+1}$. The inequality is by the margin assumption in the theorem statement. But then, keeping in mind that $\mathbf{w_0} \cdot \mathbf{u} = 0$, the above bound implies that $\vec{w_{T+1}}\mathbf{u} \geq T\delta$

Now, for the second part. We prove that $||\mathbf{w}_{t+1}||^2 \leq T$. The strategy is similar as above. We have:

$$||\mathbf{w}_{t+1}||^2 = (\mathbf{w}_t + y_t \mathbf{x}_t) \cdot (\mathbf{w}_t + y_t \mathbf{x}_t) = ||w_t||_2^2 + 2y_t(\mathbf{x}_t \cdot \mathbf{w}_t) + ||\mathbf{x}_t||_2^2$$

But now, by assumption, since we get a mistake at each round, $y_t(\mathbf{x}_t \cdot \mathbf{w}_t) \leq 0$, and again, by the normalization assumption, $||\mathbf{x}_t||_2^2 \leq 1$, so we get that $||\mathbf{w}_{t+1}||^2 \leq ||\mathbf{w}_t||_2^2 + 1$. From the fact that $\mathbf{w_0} = 0$, we get that $||\mathbf{w}_{t+1}||_2^2 \leq T$, as we need.

Now, we put everything together. We get that $1 \geq \Phi_T \geq \frac{T\delta}{\sqrt{T}}$, i.e. $T \leq \frac{1}{\delta^2}$, which is exactly what we needed to prove.

□

As a side remark, notice a simple consequence of the above: since the VC dimension of the hypothesis space certainly is upper bounded by the number of mistakes the algorithm makes, we get that the VC dimension of threshold functions with margin at least δ is at most $\frac{1}{\delta^2}$.

To motivate the next section, consider a scenario where the target $\mathbf{u}$ consists of 0's and 1's, and the number of 1's in the vector is k. (And think of k as being small compared to N, the number of experts, i.e. it's a very sparse vector.) Notice that this certainly is an instance of the problem we looked at above - the k experts are the "perfect" committee. If we normalize $\mathbf{x}_t, \mathbf{u}$, as in the proof of the theorem above, we'd get that the margin $y_t(\mathbf{u} \cdot \mathbf{x}_t) \geq \frac{1}{\sqrt{Nk}}$. So, we get that the perceptron algorithm would make at most Nk mistakes. But this is really bad - consider interpreting the experts as features, and we have millions of irrelevant features, and the committee is the important (maybe a dozen) features. We get a linear dependence on N, which is not at all desireable.

Motivated by this exact scenario, we will design another update algorithm - called the Winnow algorithm, which will get a much better bound in this case.

2.2 Winnow algorithm

Let's describe the update rules first:

- The $\mathbf{w}_t$ vectors will be normalized so that the coordinates sum to 1: one can think of them as probability distributions.
- $\mathbf{w}_1 = \mathbf{1}/N$, where $\mathbf{1}$ is the all-ones vector (i.e. we start with a uniform distribution over all experts).

- If we make a mistake, set $w_{t+1,i} = \frac{w_{t,i} \cdot e^{\eta y_i \neq x_{t,i}}}{Z_t}$. Here η is a parameter we will define later, and Z_t is a normalization factor. This is nothing more than an exponential "punishment" for the experts that are wrong. The way to see it, is that, ignoring the normalization factors, the above update is equivalent to $w_{t+1,i} = w_{t,i}e^{\eta}$, if i predicted correctly, and $w_{t+1,i} = w_{t,i}e^{-\eta}$ otherwise. Ignoring normalization again, we can interpret this as $w_{t+1,i} = w_{t,i}$, if i predicted correctly, and $w_{t+1,i} = w_{t,i}e^{-2\eta}$ otherwise. This is exactly the view we claimed.
- If we are correct, don't do anything.

Before we dive into the analysis, one thing to be noticed is that the Winnow algorithm is extremely similar to the boosting framework - and with good reason. The connection will be made more explicit in future lessons.

Before stating the formal theorem for the performance of the Winnow algorithm, we again make a few normalization assumptions without loss of generality:

- We assume $||\mathbf{x}_t||_\infty \leq 1$.
- We assume $||\mathbf{u}||_1 \leq 1$ and $u_i \geq 0$ for all indices i.
- We assume that we make a mistake at each round.

Theorem 2. *Assume the online learning with experts scenario, along with the normalization assumptions above. Furthermore, assume that there exists $\mathbf{u} \in \mathbb{R}^N$, such that for all t, $(\mathbf{u} \cdot \mathbf{x}_t)y_t \geq \delta$. Then, the winnow updates algorithm makes at most $\frac{2 \ln N}{\delta^2}$ mistakes.*

Proof. The general approach is similar as previously also - we come up with a quantity Φ_t, which we both upper and lower-bound. The quantity we will use here is $\Phi_t = RE(\mathbf{u}||\mathbf{w}_t)$, where RE is the relative entropy or KL divergence of the two distributions. (I already mentioned $\mathbf{w}_t, \mathbf{u}$ can be interpreted as distributions.)

Trivially, we have that $\Phi_t \geq 0$ for all t.

On the other direction, we make a potential sort of argument. In other words, we upper bound $\Phi_{t+1} - \Phi_t$ which by telescoping will give an upper bound on Φ_T. Let's write out the calculations:

$$\begin{aligned}\Phi_{t+1} - \Phi_t &= RE(\mathbf{u}||\mathbf{w}_{t+1}) - RE(\mathbf{u}||\mathbf{w}_t) \\ &= \sum_i u_i \ln \frac{w_{t,i}}{w_{t+1,i}} = \sum_i u_i \ln \frac{Z_t}{e^{\eta y_t x_i}} \\ &= \ln(Z_t) - \sum_i u_i \eta y_t x_{t,i}\end{aligned}$$

This follows just by taking logarithms, and the fact that $\sum_i u_i = 1$, since $\mathbf{u}$ is a distribution. Continuing onward, this equals:

$$\begin{aligned}&\ln(Z_t) - \eta y_t(\mathbf{u} \cdot \mathbf{x}_t) \\ &\leq \ln(Z_t) - \eta\delta\end{aligned}$$

This last inequality just follows from the margin property we assumed. Let's work on Z_t, now. First,

$$Z_t = \sum_i w_{t,i} e^{\eta y_t x_{t,i}} \leq$$

$$\sum_i w_{t,i} \left(\frac{1 + y_t x_{t,i}}{2} e^{\eta} + \frac{1 - y_t x_{t,i}}{2} e^{-\eta} \right)$$

The way to see this inequality is as follows. Consider the function e^x and the line through the points $(-\eta, e^{\eta})$ and (η, e^{η}). Since e^x is convex, the line will always be above the e^x curve (see Figure 2). If the line has equation $y = f(x)$, then for any $-\eta \leq x \leq \eta, e^x \leq f(x)$. The above inequality is derived exactly by comparing $e^{\eta y_t x_{t,i}}$ and $f(\eta y_t x_{t,i})$.

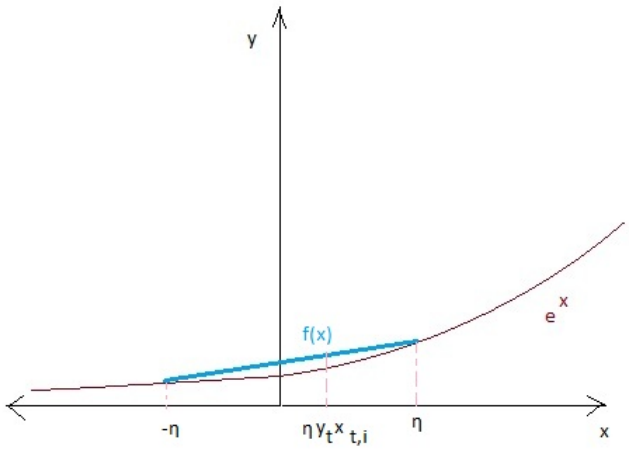

Figure 2: Upper-bounding exponential function by linear function

Then, we simplify further to get:

$$\sum_i w_{t,i} \left(\frac{1 + y_t x_{t,i}}{2} e^{\eta} + \frac{1 - y_t x_{t,i}}{2} e^{-\eta} \right)$$
$$= \frac{e^{\eta} + e^{-\eta}}{2} \sum_i w_{t,i} + \frac{e^{\eta} - e^{-\eta}}{2} y_t \sum_i w_{t,i} x_{t,i}$$

Now, $\sum_i w_{t,i} = 1$, by our assumption, $e^{\eta} - e^{-\eta} \geq 0$ since we will take $\eta > 0$ of course, and

$$y_t \sum_i w_{t,i} x_{t,i} = y_t (\mathbf{w}_t \cdot \mathbf{x}_t) \leq 0$$

by the assumption that we make a mistake at each round. Hence, $Z_t \leq \frac{e^{\eta} + e^{-\eta}}{2}$. So, we get that $\Phi_{t+1} - \Phi_t \leq \ln(\frac{e^{\eta} + e^{-\eta}}{2}) - \eta\delta$. The right hand side is a constant, and let's say it equals to $-C$.

Furthermore, $\Phi_1 = \sum_i u_i \ln(N u_i) \leq \sum_i u_i \ln N \leq \ln N$. Hence, we have that $0 \leq \Phi_{T+1} \leq \ln N - C \cdot T$. This implies that $T \leq \frac{\ln N}{C}$. We want to make C as large as possible to get the best bound. A bit of basic calculus implies that $\eta = \frac{1}{2} \ln(\frac{1+\delta}{1-\delta})$ is the value to pick, which makes

$$C = RE(\frac{1}{2} - \frac{\delta}{2} || \frac{1}{2}) \geq 2(\delta/2)^2 = \delta^2/2$$

Hence, the number of mistakes is at most $\frac{2 \ln N}{\delta^2}$, as we need.

□