CS260: Machine Learning Theory
Lecture 8: The Perceptron Algorithm

October 19, 2011

Lecturer: Jennifer Wortman Vaughan

1 Preliminaries

In this lecture, we will analyze the Perceptron algorithm for learning n-dimensional linear threshold functions or n-dimensional linear separators. (We will use these terms interchangeably.) For today's class, we will use the label set $\{-1, +1\}$ instead of $\{0, 1\}$. This notational change doesn't make any difference in terms of the meaning of the learning problem, but it will make some of our derivations easier. In these notes, we will use bold letters to represent vectors, and $\|\mathbf{w}\|$ denotes the length of the vector $\mathbf{w}$.

We can map any n-dimensional linear separator that passes through the origin to an n-dimensional weight vector $\mathbf{w}$ such that $\mathbf{w} \cdot \mathbf{x} \geq 0$ for all positive points $\mathbf{x}$, and $\mathbf{w} \cdot \mathbf{x} < 0$ for all negative points. This vector $\mathbf{w}$ is any normal vector of the decision boundary. For this lecture, we will restrict our attention to linear separators that pass through the origin. However, this restriction is without loss of generality as any n-dimensional linear separator can be represented as an $(n + 1)$-dimensional linear separator that passes through the origin by adding a "dummy feature" that is always equal to 1 to each input point $\mathbf{x}$. (Exercise: Work out the details and convince yourself this is true.)

2 The Margin

Suppose that we run a learning algorithm on a data set and it outputs a linear threshold function. Intuitively speaking, we can probably be relatively confident that points that are far from the decision boundary are labeled correctly, while we may be less confident about points that are very close to the decision boundary, since a small change to the boundary would result in different labels for these points. It would be nice if we could find a decision boundary such that no points are too close. We formalize this idea by introducing the notion of a margin.

Definition 1. *Given a linear separator represented by its normal vector* $\mathbf{w}$*, the* margin γ *of a point* $\mathbf{x}$ *with label* $y \in \{\,-1, +1\,\}$ *is the distance between* $\mathbf{x}$ *and the decision boundary. That is,*

$$\gamma \;=\; y\left(\frac{\mathbf{w}}{\|\mathbf{w}\|} \cdot \mathbf{x}\right).$$

Let's verify that this expression does indeed give the distance between $\mathbf{x}$ and the decision boundary.

If $\mathbf{x}$ lies on the decision boundary, then $\mathbf{x}$ must be orthogonal to $\mathbf{w}$, and so we get $\gamma = 0$ as desired.

All CS260 lecture notes build on the scribes' notes written by UCLA students in the Fall 2010 offering of this course. Although they have been carefully reviewed, it is entirely possible that some of them contain errors. If you spot an error, please email Jenn.

Suppose $\mathbf{x}$ does not lie on the boundary. Let $\mathbf{z}$ be the projection of $\mathbf{x}$ onto the decision boundary, i.e., $\mathbf{z}$ is the closest point to $\mathbf{x}$ that lies on the boundary. Note that $\mathbf{x_i} - \mathbf{z_i}$ is parallel to $\mathbf{w}$. If $y = +1$, we have that

$$\begin{aligned}
\mathbf{x} - \mathbf{z} &= \gamma \frac{\mathbf{w}}{\|\mathbf{w}\|} \\
\mathbf{z} &= \mathbf{x} - \gamma \frac{\mathbf{w}}{\|\mathbf{w}\|} \\
\mathbf{z} \cdot \mathbf{w} &= \mathbf{x} \cdot \mathbf{w} - \gamma \frac{\mathbf{w} \cdot \mathbf{w}}{\|\mathbf{w}\|} \\
0 &= \mathbf{x} \cdot \mathbf{w} - \gamma \|\mathbf{w}\| \\
\gamma &= \mathbf{x} \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|} = y \left(\mathbf{x} \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|} \right)
\end{aligned}$$

If $y = -1$, the derivation is similar, except we start with

$$\mathbf{x} - \mathbf{z} = -\gamma \frac{\mathbf{w}}{\|\mathbf{w}\|}$$

which leads us to

$$\gamma = -\mathbf{x} \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|} = y \left(\mathbf{x} \cdot \frac{\mathbf{w}}{\|\mathbf{w}\|} \right).$$

These ideas are illustrated in the figure below.

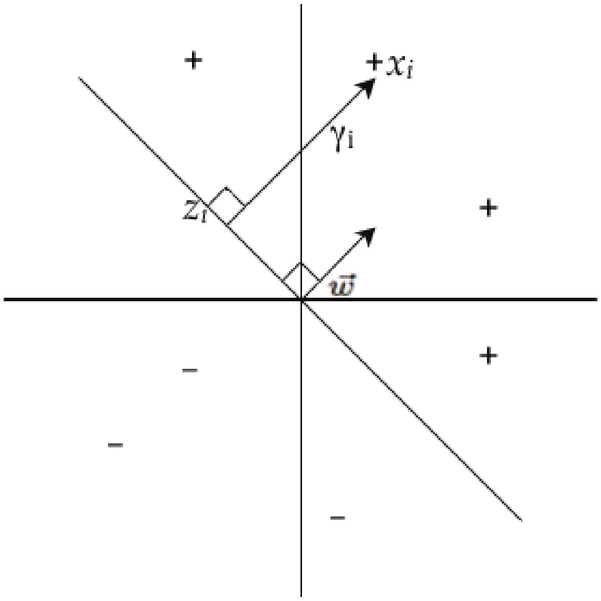

Figure 2: γ_i is the distance between the point $\mathbf{x_i}$ and the separator.

3 The Perceptron Algorithm

With these ideas in place, we are ready to introduce the Perceptron algorithm.

```
PERCEPTRON ALGORITHM
Initialize $\mathbf{w_1} = \mathbf{0}$
At each round $t \in \{1, 2, \cdots\}$
    • Receive input $\mathbf{x_t}$
    • If $\mathbf{w_t} \cdot \mathbf{x_t} \geq 0$, predict $+1$, else predict $-1$
    • If there is a mistake (i.e., if $y_t(\mathbf{w_t} \cdot \mathbf{x_t}) < 0$), set $\mathbf{w}_{t+1} \leftarrow \mathbf{w_t} + y_t \cdot \mathbf{x_t}$, else set $\mathbf{w}_{t+1} \leftarrow \mathbf{w_t}$.
```

The general intuition behind the algorithm is that every time it makes a mistake on a positive example, it shifts the weight vector toward the input point, whereas every time it makes a mistake on a negative point, it shifts the weight vector away from that point.

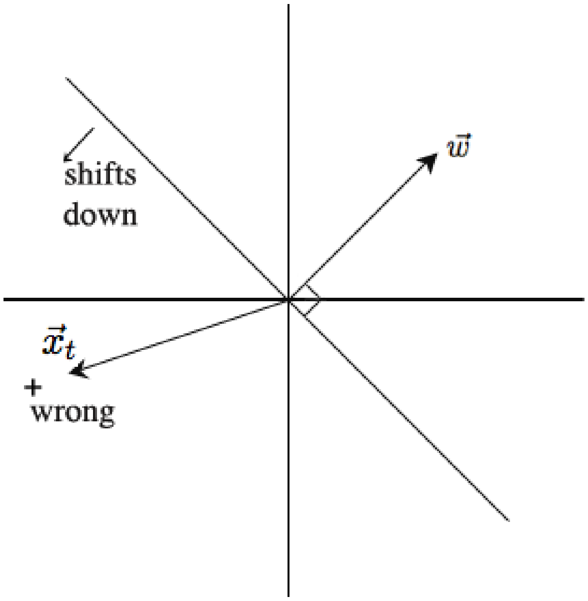

Figure 3

Now we prove a mistake bound for the above algorithm. We assume that the perfect target function is represented by a normal vector $\mathbf{u}$ of unit length; this is without loss of generality since normalizing the weight vector doesn't change the decision boundary. However, we also make a stronger assumption that the perfect target function we consider has a margin of at least γ. (Exercise: Show how an adversary could force any algorithm to make an unbounded number of mistakes if we didn't have a margin assumption like this.)

Theorem 1. *Suppose there exists a $\mathbf{u}$ of unit length and values $\gamma > 0$ and $D > 0$ such that $\forall t \;\; y_t(\mathbf{x}_t \cdot \mathbf{u}) \geq \gamma$ and $\|\mathbf{x}_t\| \leq D$. Then, the number of mistakes made by the Perceptron algorithm is no more than $(D/\gamma)^2$.*

Let $m(i)$ be the round in which the ith mistake is made. Define $m(0) = 0$.

Lemma 1. *For all mistakes k, $\mathbf{w}_{m(k)+1} \cdot \mathbf{u} \geq k\gamma$.*

Proof: We prove by induction on the number of mistakes k. For the base case, $k = 0$, note that since the initial weight vector $\mathbf{w_1}$ is all 0s, we have $\mathbf{w}_{m(0)+1} \cdot \mathbf{u} = \mathbf{w}_1 \cdot \mathbf{u} = 0$.

For the induction hypothesis, assume that the above statement holds true for all $k < i$.

For the induction step, consider $\mathbf{w}_{m(i)+1}$. We have

$$\begin{aligned} \mathbf{w}_{m(i)+1} \cdot \mathbf{u} &= (\mathbf{w}_{m(i)} + y_{m(i)}\mathbf{x}_{m(i)}) \cdot \mathbf{u} \\ &= \mathbf{w}_{m(i)} \cdot \mathbf{u} + y_{m(i)}(\mathbf{x}_{m(i)} \cdot \mathbf{u}). \end{aligned}$$

The first equality comes from the Perceptron update rule. We did make a mistake on round $m(i)$, so the weights at round $m(i) + 1$ can be computed by applying the update rule to the weights at round $m(i)$. Now, we know that we did not make a mistake between round $m(i-1)+1$ and round $m(i)$. Since the Perceptron only updates weights when there is a mistake, we have $\mathbf{w}_{m(i)} \cdot \mathbf{u} = \mathbf{w}_{m(i-1)+1} \cdot \mathbf{u}$. We also have $y_{m(i)}(\mathbf{x}_{m(i)} \cdot \mathbf{u}) \geq \gamma$, by the margin requirement in the statement of the theorem. Thus, we have,

$$\begin{aligned} \mathbf{w}_{m(i)+1} \cdot \mathbf{u} &\geq \mathbf{w}_{m(i-1)+1} \cdot \mathbf{u} + \gamma \\ &\geq i\gamma, \end{aligned}$$

where the last inequality follows from the induction hypothesis. □

Lemma 2. *For all mistakes k, $\|\mathbf{w}_{m(k)+1}\|^2 \leq kD^2$.*

Proof: We again prove this lemma by induction on the number of mistakes k. For the base case, $k = 0$, we have $\|\mathbf{w}_{m(0)+1}\|^2 = \|\mathbf{w}_1\|^2 = 0$.

Now, let us assume that the statement is true for all $k < i$.

For the induction step, note that

$$\begin{aligned} \|\mathbf{w}_{m(i)+1}\|^2 &= \|\mathbf{w}_{m(i)} + y_{m(i)}\mathbf{x}_{m(i)}\|^2 \\ &= \|\mathbf{w}_{m(i)}\|^2 + \|\mathbf{x}_{m(i)}\|^2 + 2y_{m(i)}(\mathbf{x}_{m(i)} \cdot \mathbf{w}_{m(i)}), \end{aligned}$$

where the first equality holds for the same reason as in Lemma 1 above. Now, as above, we have $\|\mathbf{w}_{m(i)}\|^2 = \|\mathbf{w}_{m(i-1)+1}\|^2$. Further, by the bound on the lengths of vectors in the theorem statement, we have $\|\mathbf{x}_{m(i)}\|^2 \leq D^2$. For the third term in the expression above, note that as there was a mistake in round $m(i)$, our prediction of the label did not match with the correct label. Thus, $y_{m(i)}(\mathbf{x}_{m(i)} \cdot \mathbf{w}_{m(i)}) < 0$. Therefore, we have,

$$\begin{aligned} \|\mathbf{w}_{m(i)+1}\|^2 &\leq \|\mathbf{w}_{m(i-1)+1}\|^2 + D^2 \\ &\leq iD^2. \end{aligned}$$

Here, the last inequality follows from the induction hypothesis. This proves the lemma. □

To complete the proof of Theorem 1, we first recall the following simple fact from linear algebra: For any vectors $\mathbf{z}$ and $\mathbf{u}$, if θ is the angle between $\mathbf{z}$ and $\mathbf{u}$, then

$$\mathbf{z} \cdot \mathbf{u} = \|\mathbf{z}\|\|\mathbf{u}\| \cos(\theta) \leq \|\mathbf{z}\|\|\mathbf{u}\|.$$

Using this fact and the two lemmas above, we have that for any mistake k,

$$\begin{aligned} D\sqrt{k} &\geq \|\mathbf{w}_{m(k)+1}\| \\ &= \|\mathbf{w}_{m(k)+1}\|\|\mathbf{u}\| \\ &\geq \mathbf{w}_{m(k)+1} \cdot \mathbf{u} \\ &\geq k\gamma. \end{aligned}$$

Thus it must be the case that $k \leq (D/\gamma)^2$, and the Perceptron cannot make more than k mistakes.