Perceptron Algorithm (Primal Form)

Input: a set S of tuples $\{(x_i, y_i)\}$ where x_i are vectors in $\mathbb{R}^n$ and y_i is a label, either -1 or 1, and a positive scalar η, called the learning rate.
Goal: find values of w and b so that when $w \cdot x + b > 0$ it's true that y_i is 1 and otherwise -1.
Initialize: $w = 0$, $b = 0$, $k = 0$, and $R = \max_i \|x_i\|$.
Mistake criteria: if $y_i(w \cdot x + b) \leq 0$, the current w and b would misclassify x_i, and we have encountered a mistake.
Iterate: through the inputs x_i (over and over again as needed), checking the mistake criteria. If there is a mistake, we increment $k \to k + 1$. We also update w and b as follows:

$$\begin{cases} w & \to & w + \eta y_i x_i \\ b & \to & b + \eta y_i R^2 \end{cases}$$

Stop: when there are no more mistakes.

Novikoff's Theorem

Let S be a set of tuples $\{(x_i, y_i)\}$ where x_i are vectors in $\mathbb{R}^n$ and y_i is a label, either -1 or 1. We assume that the set S is non-trivial, which means that not all x_i have the same label. Then, if the set is linearly separable, we know there exists a vector w_{opt} in $\mathbb{R}^n$, a scalar b_{opt}, and a positive scalar γ, so that $\|w_{opt}\| = 1$ and

$$y_i(w_{opt} \cdot x_i + b_{opt}) \geq \gamma$$

for all i. If this holds and we use the Perceptron Algorithm to find a value of w and b that separate the points, then the number of mistakes made by the algorithm is bounded by

$$\left(\frac{2R}{\gamma}\right)^2.$$

This theorem proves that the algorithm converges when we have a non-trivial, linearly separable set of points, and provides an upper bound on the time complexity.

Proof

Let $\hat{w}$ be the vector in $\mathbb{R}^{n+1}$ which is the vector w augmented by b/R, and $\hat{x}_i$ the vector in $\mathbb{R}^{n+1}$ which is the vector x_i augmented by R. This makes it so that $w \cdot x + b = \hat{w}_i \cdot \hat{x}_i$. One can easily check that the update rule $\hat{w} \to \hat{w} + \eta y_i \hat{x}_i$ preserves the updates to w and b in the algorithm.

Suppose we are about to make the kth correction to the parameters, meaning we have just discovered the kth mistake. We'll prove by induction that $k \leq (2R/\gamma)^2$, no matter the value of k.

Lemma 1

$$\|\hat{w}_{opt}\|^2 \leq 2.$$

Proof

Notice that $b_{opt} \leq R$, because otherwise the vectors x_i would all live on one side of the hyperplane $y = w_{opt} \cdot x + b_{opt}$. Therefore:

$$\|\hat{w}_{opt}\|^2 = \|w_{opt}\|^2 + \left(\frac{b_{opt}}{R}\right)^2 = 1 + \left(\frac{b_{opt}}{R}\right)^2 \leq 2.$$

$\square$

Lemma 2

$$\hat{w}_k \cdot \hat{w}_{opt} \geq k\eta\gamma$$

Proof

By induction:

$$\begin{aligned}
\hat{w}_k \cdot \hat{w}_{opt} &= (\hat{w}_{k-1} + \eta y_i \hat{x}_i) \cdot \hat{w}_{opt} \\
&= \hat{w}_{k-1} \cdot \hat{w}_{opt} + \eta y_i \hat{x}_i \cdot \hat{w}_{opt} \\
&\geq (k-1)\eta\gamma + \eta y_i \hat{x}_i \cdot \hat{w}_{opt} \\
&\geq (k-1)\eta\gamma + \eta\gamma \\
&= k\eta\gamma.
\end{aligned}$$

$\square$

Lemma 3

$$\|\hat{w}_k\|^2 \leq 2k\eta^2 R^2$$

Proof

By induction:

$$\begin{aligned}
\|\hat{w}_k\|^2 &= \|\hat{w}_{k-1} + \eta y_i \hat{x}_i\|^2 \\
&= \|\hat{w}_{k-1}\|^2 + 2\eta y_i \hat{w}_{k-1} \cdot \hat{x}_i + \eta^2 \|\hat{x}_i\|^2 \\
&\leq \|\hat{w}_{k-1}\|^2 + \eta^2 \|\hat{x}_i\|^2 \\
&= \|\hat{w}_{k-1}\|^2 + \eta^2 (\|x_i\|^2 + R^2) \\
&\leq 2(k-1)\eta^2 R^2 + 2\eta^2 R^2 \\
&= 2k\eta^2 R^2
\end{aligned}$$

where we have used the fact that $y_i \hat{w}_{k-1} \cdot \hat{x}_i$ is negative in going from lines (2) to (3). $\square$

Finally, to prove the main theorem we assemble these facts.

$$\begin{aligned}
k &\leq \frac{\hat{w}_k \cdot \hat{w}_{opt}}{\eta\gamma} \\
&\leq \frac{\|\hat{w}_k\| \|\hat{w}_{opt}\|}{\eta\gamma} \\
&\leq \frac{\sqrt{2}\|\hat{w}_k\|}{\eta\gamma} \\
&\leq \frac{2\sqrt{k}\eta R}{\eta\gamma} = \frac{2\sqrt{k}R}{\gamma}
\end{aligned}$$

which implies

$$\begin{aligned}
\sqrt{k} &\leq \frac{2R}{\gamma} \\
k &= \left(\frac{2R}{\gamma}\right)^2
\end{aligned}$$

$\square$