The Perceptron Algorithm

An online algorithm for learning threshold functions.

$$f : \mathbb{R}^n \longrightarrow \{+1, -1\}$$

$\vec{x} \in \mathbb{R}^n$
$$f_{\vec{\omega},b}(\vec{x}) = \begin{cases} +1 \text{ if } \vec{\omega}\cdot\vec{x} \geq b \\ -1 \text{ if } \vec{\omega}\cdot\vec{x} < b \end{cases}$$

This defines a hyperplane in $\mathbb{R}^n$ with $\vec{\omega}$ as the normal to the hyperplane. The goal is to learn which hyperplane separates the positive labelled points from the negative ~~~~ labelled points. Without loss of generality we can set $b=0$. We can redefine the problem in $\mathbb{R}^{n+1}$

$$\bar{f} : \mathbb{R}^{n+1} \longrightarrow \{+1,-1\}$$

and change $\vec{x} \in \mathbb{R}^n$ to $\vec{x}' = \begin{pmatrix} \vec{x} \\ 1 \end{pmatrix} \in \mathbb{R}^{n+1}$

and $\vec{\omega}$ to $\vec{\omega}' \in \mathbb{R}^{n+1}$ where $\vec{\omega}' = \begin{pmatrix} \vec{\omega} \\ -b \end{pmatrix}$

Now $\vec{x}'\cdot\vec{\omega}' \geq 0 \Leftrightarrow \vec{x}\cdot\vec{\omega} \geq b$.

We want to <u>learn</u> $\vec{\omega}$ from examples. The examples are presented to us in online fashion. So we have to make a prediction and then the true label is revealed. We seek <u>bounds</u> on our mistakes.

BASIC PERCEPTRON ALGORITHM:

Init: $\vec{\omega}_1 = 0$

LOOP $t = 1 \ldots T$
- Obtain $\vec{x}_t$
- (PREDICT) if $\vec{\omega}\cdot\vec{x}_t \geq 0$ then $\hat{y}_t = +1$ else $\hat{y}_t = -1$
- Obtain y_t
- (UPDATE) if $\hat{y}_t \neq y_t$ then $\vec{\omega}_{t+1} \leftarrow \vec{\omega}_t + y_t\vec{x}_t$ else $\vec{\omega}_{t+1} = \vec{\omega}_t$

When it makes a mistake on a positive example it shifts the weight vector towards that point if it makes a mistake on a negative example it shifts the weight vector away.

Crucial concept: How close is a point to the hyperplane
Let f be a threshold function determined by a hyperplane passing through the origin with unit normal $\vec{u}$.

Def The margin of f at a labelled point $(\vec{x}, y)$ is defined to be $y(\vec{x} \cdot \vec{u})$.

Remark If the margin is positive then f classifies $(\vec{x}, y)$ correctly and if it is negative then f classifies the point incorrectly. The numerical value of the margin gives the distance to the decision boundary. The margin of a finite set of points is the minimum margin of any point in the set. The margin measures robustness of the classification.

Thm Suppose $(\vec{x}_1, y_1), (\vec{x}_2, y_2) \cdots (\vec{x}_T, y_T)$ are such that $\|\vec{x}_t\| \leq D$ for some $D > 0$ and all $t \in \{1, \dots T\}$. Suppose $\exists\, \vec{u}$ a unit vector & $\gamma > 0$ such that $\forall t \quad y_t(\vec{x}_t \cdot \vec{u}) \geq \gamma$. Then the perceptron algorithm makes at most $(D/\gamma)^2$ mistakes.

Proof Let m_t be the number of mistakes just before round t. Thus $m_1 = 0$. We proceed by breaking the proof into 2 lemmas.

Lemma 1 $\forall t \in \{1, \dots, T\} \quad \vec{w}_t \cdot \vec{u} \geq m_t \gamma$

Proof By induction on t.

Base case $t = 1$, $m_1 = 0$ so clearly $\vec{w}_1 \cdot \vec{u} = 0 = m_1 \gamma$

Induction step: suppose there is a mistake in round t. Then we have

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

$$\geq \underbrace{m_t \gamma}_{\text{Ind Hyp}} + \underbrace{\gamma}_{\text{Margin assumption}} = m_{t+1} \gamma$$

So lemma 1 shows that as mistakes are made the projection of $\vec{w}_t$ on $\vec{u}$ gets longer.

<u>Lemma 2</u> $\|\vec{w}_t\|^2 \leq m_t D^2$

<u>Proof</u> Induction on t; base case is trivial as is the case when there is no mistake. Suppose there is a mistake in round t. Then we have

$$\|\vec{w}_{t+1}\|^2 = \|\vec{w}_t + y_t \vec{x}_t\|^2$$
$$= \|\vec{w}_t\|^2 + \|\vec{x}_t\|^2 + \underbrace{2 y_t (\vec{w}_t \cdot \vec{x}_t)}_{\text{Negative}}$$
$$\leq \|\vec{w}_t\|^2 + \|\vec{x}_t\|^2$$
$$\leq m_t D^2 + D^2 = m_{t+1} D^2$$

<u>Proof of Thm</u> $\|w_t\|^2 \leq m_t D^2$ (Lemma 2)

or $D\sqrt{m_t} \geq \|\vec{w}_t\|$
$$= \|\vec{w}_t\| \, \|\vec{u}\|$$
$$\geq \vec{w}_t \cdot \vec{u} \quad [\text{Cauchy-Schwartz}]$$
$$\geq m_t \gamma$$

or $D/\gamma \geq \sqrt{m_t}$

or $m_t \leq (D/\gamma)^2$ ∎

<u>APPLICATION</u> Suppose there are n financial analysts who predict every day whether the market will go up or down. We represent each prediction as a vector in $\{+1,-1\}^n$. We would like to use perceptron to figure out whom to follow. Suppose there are k "experts" within the group s.t. a majority vote among these k always gives the right answer.

Define $\vec{u} = \frac{1}{\sqrt{k}}(0, \cdot\, 0, 1, 0, 1, 1, 0, \ldots 0, 1 \cdots)$ where the entry is 1 if it corresponds to one of the k experts.

Then $y_t(\vec{x}_t \cdot \vec{u}) \geq 1/\sqrt{k}$ since this subset is always right.

Thus $\|x_t\| \leq \sqrt{n} = D$ & $\gamma = 1/\sqrt{k}$. So the mistake bound is nk.