Linear Classification with a Perceptron

Anoop Sarkar

A good tutorial to refresh your memory of vectors and basic linear algebra is (Jordan, 1986).

Binary classification can be done using a function $f : x \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$. Input $x = (x_1, \ldots, x_n)$ is assigned to $+1$ if $f(x) \geq 0$ else it is assigned to -1. $f(x)$ is assumed to be a linear function. So we can write f as follows:

$$\begin{aligned} f(x) &= w \cdot x + b \\ &= \left(\sum_{i=1}^{n} w_i x_i \right) + b \end{aligned}$$

The parameters for this linear function are w and b, and (w, b) is called the hyperplane which defines a line that cuts through the points in the training data.

The *functional margin* of example (x_i, y_i) with respect to hyperplane (w, b) is defined as:

$$\gamma_i = y_i(w \cdot x_i + b)$$

If $\gamma_i > 0$ then this implies that (x_i, y_i) is correctly classified by the hyperplane.

The *functional margin distribution* of a hyperplane (w, b) wrt training set z is the distribution of margins of examples in z. The minimum of the margin distribution is the margin of the hyperplane.

The *geometric margin* measures Euclidean distance of the points from the decision boundary in the space of the examples x_i and is defined as the vector $(\frac{w}{||w||}, \frac{b}{||w||})$, where $||w||$ is the norm of the vector defined as $\sqrt{w \cdot w} = \sqrt{\sum_{i=1}^{n} w_i^2}$. The margin of a training set z is the *maximum* geometric margin over all hyperplanes on z. A hyperplane that realizes the maximum is called the maximum margin hyperplane.

The Perceptron algorithm is defined as follows:

Given training set z
Set $w_0 = \textit{zeroes}$, $b_0 = 0$ and $k = 0$
Set $R = \max_{1 \leq i \leq \ell} ||x_i||$
repeat for number of epochs
    for $i = 1, \ldots, \ell$
        if $y_i(w_k \cdot x_i + b_k) \leq 0$ then
            $w_{k+1} = w_k + y_i x_i$
            $b_{k+1} = b_k + y_i R^2$
            $k = k + 1$

We can show that the number of mistakes for the perceptron algorithm is bounded based on the properties of the data. Let z be a non-trivial training set. Suppose there exists a vector w_{opt} such that $||w_{opt}|| = 1$ and

$$y_i(w_{opt} \cdot w_i + b_{opt}) \geq \gamma \text{ for } i = 1, \ldots, \ell$$

The number of mistakes made by the perceptron on z is at most $(\frac{2R}{\gamma})^2$.

The first step in the proof is to fold in the b parameter into the weight vector using the following transformation: for each x_i we replace it with a new vector $x'_i = (x_{i_1}, \ldots, x_{i_n}, R)$ and similarly w is replaced with a new weight vector $w' = (w_1, \ldots, w_n, \frac{b}{R})$.

We start with $w'_0 = \textit{zeroes}$. Let w'_{t-1} be the weight vector just before the t^{th} mistake.

$$y_i(w'_{t-1} \cdot x'_i) = y_i(w_{t-1} \cdot x_i) + b_{t-1} \leq 0$$

So $w'_{t-1} = (w_{1_{t-1}}, \ldots, w_{n_{t-1}}, \frac{b^{t-1}}{R})$ and so:

$$
\begin{aligned}
w'_t &= (w_{1_t}, \ldots, w_{n_t}, \frac{b^t}{R}) \\
w_t &= w_{t-1} + y_i x_i \\
\frac{b_t}{R} &= \frac{b_t}{R} + y_i R \\
b_t &= b_{t-1} + y_i R^2
\end{aligned}
$$

Let us consider w_{opt} again.

$$
\begin{aligned}
w_t \cdot w_{opt} &= w_{t-1} \cdot w_{opt} + y_i (x_i \cdot w_{opt}) \\
w_t \cdot w_{opt} &\geq w_{t-1} \cdot w_{opt} + \gamma
\end{aligned}
$$

We started with w_0 initialized as zeroes, and so by induction we can see that:

$$w_t \cdot w_{opt} \geq t\gamma$$

This implies:

$$w'_t \cdot w'_{opt} \geq t\gamma$$

Similarly, we have:

$$
\begin{aligned}
||w'_t||^2 &= ||w'_{t-1}||^2 + 2y_i(w'_{t-1} \cdot x'_i) + ||x'_i||^2 \\
&\leq ||w'_{t-1}||^2 + ||x'_i||^2 \\
&\leq ||w'_{t-1}||^2 + ||x_i||^2 + R^2 \\
&\leq ||w'_{t-1}||^2 + 2R^2
\end{aligned}
$$

By induction, we get:

$$||w'_t||^2 \leq 2tR^2$$

Combining the two inequalities:

$$||w'_{opt}||\sqrt{2t}R \geq ||w'_{opt}||\ ||w'_t|| \geq w'_t \cdot w'_{opt} \geq t\gamma$$

which implies that:

$$t \leq 2\left(\frac{R}{\gamma}\right)^2 ||w'_{opt}||^2 \leq \left(\frac{2R}{\gamma}\right)^2$$

Since $b_{opt} \leq R$ (the convex hull of the points) for a non-trivial separation of the data and $||w_{opt}||^2 = 1$ hence:

$$||w'_{opt}||^2 \leq ||w_{opt}||^2 + 1 = 2$$

More details can be found in (Cristianini and Shawe-Taylor, 2000).

References

Michael Jordan 1986. An Introduction to Linear Algebra in Parallel Distributed Processing Chapter 9. In *Parallel Distributed Processing - Vol 1* ed. David Rumelhart. MIT Press.

Nello Cristianini and John Shawe-Taylor 2000. *An Introduction to Support Vector Machines: and other kernel based methods* Cambridge University Press.