Perceptrons

Prof. Dan A. Simovici

UMB

1 On-Line vs. Off-Line Learning

2 The Algorithm

- For support vector machines the training set part of the data set to be classified is presented to the algorithm, the classification function is inferred, and then the algorithm is tested on the test set part of the data set.
- The perceptron constructs the classification function contemporaneously with the analysis of the training set; this exemplifies the paradigm of *on-line learning*.

- A *training set* is a sequence of pairs $S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$, where $(\mathbf{x}_i, y_i) \in \mathbb{R}^n \times \{-1, 1\}$ for $1 \leqslant i \leqslant n$. If $y = 1$, $\mathbf{x}$ is a *positive example*; if $y = -1$, $\mathbf{x}$ is a *negative example*.
- The sequence S is *linearly separable* if there exists a hyperplane $\mathbf{w}_*'\mathbf{x} + b_* = 0$ such that $\mathbf{w}_*'\mathbf{x}_i + b_* \geqslant 0$ if $y_i = 1$ and $\mathbf{w}_*'\mathbf{x}_i + b_* < 0$ if $y_i = -1$. Both cases are captured by the inequality $\gamma_i = y_i(\mathbf{w}_*'\mathbf{x}_i + b_*) \geqslant 0$. The number γ_i is the *functional margin* of $(\mathbf{x}_i, y_i)$.
- If $\gamma_i > 0$ then $(\mathbf{x}_i, y_i)$ is classified correctly; otherwise, it is incorrectly classified and we say that a mistake occurred. Without loss of generality we may assume that

$$\left\| \mathbf{w}_* \right\| = 1;$$

  if this is not the case, the coefficients of the hyperplane $\mathbf{w}_*'\mathbf{x} + b_* = 0$ may be rescaled to make $\left\| \mathbf{w}_* \right\| = 1$.

Terminology

- The vector $\mathbf{w}$ (or $\mathbf{w}_*$) is the *weight vector*.
- The number b (or b_*) is the *bias*.

Also, we may assume that there exists $\gamma > 0$ such that

$$y_i(\mathbf{w}_*'\mathbf{x}_i + b_*) \geqslant \gamma. \tag{1}$$

The algorithm builds a sequence of weight vectors $(\mathbf{w}_k)$ and a sequence of bias values (b_k).

How does it work?

- Upon examining the first $m - 1$ training examples

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{m-1}, y_{m-1})$$

  and making the predictions $y_1, \ldots, y_{m-1}$ (some of which may be erroneous, in which cases modification are applied to parameters maintained by the algorithm), the algorithm is presented with the input x_m.
- Asumming that at that moment the parameters of the algorithm are $\mathbf{w}_k$ and b_k, an error is committed if $y_i(\mathbf{w}'_k \mathbf{x}_i + b_k) < 0$. In this case, a correction of the parameters of the algorithm is applied; otherwise, the algorithm continues by analyzing the next example.
- The processing of the sequence of pairs $((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$ that occurs between two consecutive errors is referred to as an *epoch* of the algorithm.

Let R be the minimum radius of a closed ball centered in $\mathbf{0}$, that contains all points $\mathbf{x}_i$, that is,

$$R = \max\{\parallel \mathbf{x}_i \parallel \mid 1 \leqslant i \leqslant \ell\}$$

and let η be a parameter called the *learning rate*.
If a correction is applied, the new weight vector is defined as:

$$\mathbf{w}_{k+1} = \mathbf{w}_k + \eta y_i \mathbf{x}_i,$$

while the new bias value will be

$$b_{k+1} = b_k + \eta y_i R^2.$$

In other words, the correction of the weight vector is

$$\Delta \mathbf{w}_k = \mathbf{w}_{k+1} - \mathbf{w}_k = \eta y_i \mathbf{x}_i$$

and the correction of the bias is

$$\Delta b_k = \eta y_i R^2.$$

The Perceptron Algorithm

input: labelled training sequence S of length ℓ and learning rate η;
output: weight vector $\mathbf{w}$ and parameter b defining classifier

```
initialize $\mathbf{w}_0 = \mathbf{0}$, $b_0 = 0$, $k = 0$, $R = \max\{\| \mathbf{x}_i \| \mid 1 \leqslant i \leqslant \ell\}$, errors $= 0$;
repeat
        errors $\leftarrow 0$
        for $(i = 1$ to $\ell)$ do {
            if $(y_i(\mathbf{w}'_k \mathbf{x}_i + b_k) < 0)$ {
                $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k + \eta y_i \mathbf{x}_i$;
                $b_{k+1} \leftarrow b_k + \eta y_i R^2$;
                $k \leftarrow k + 1$;
                errors $\leftarrow 1$;
            }
        }
until (errors $==$ 0); # (no new errors occur in the current epoch)
return $k$, $\mathbf{w}_* = \mathbf{w}_k$ and $b_* = b_k$ where $k$ is the number of mistakes
```

Theorem

$S = ((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_\ell, y_\ell))$ be a training sequence that is linearly separable, and let $R = \max\{\| \mathbf{x}_i \| \mid 1 \leqslant i \leqslant \ell\}$. Suppose there exists a weight vector $\mathbf{w}_*$ and a bias b_* such that

$$\| \mathbf{w}_* \| = 1 \text{ and } y_i(\mathbf{w}_*'\mathbf{x}_i + b_*) \geqslant \gamma,$$

for $1 \leqslant i \leqslant \ell$. Then, the number of mistakes made by the algorithm is at most

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

Proof

As we noted before, we may assume that $\| \mathbf{w}_* \| = 1$.
Let k be the update counter and let $\mathbf{w}_k$ be the weight vector when the algorithm makes error k on example $\mathbf{x}_i$. Then,

$$\begin{aligned} \mathbf{w}_{k+1} &= \mathbf{w}_k + \eta y_i \mathbf{x}_i \\ b_{k+1} &= b_k + \eta y_i R^2. \end{aligned}$$

Let $\tilde{\mathbf{w}}_k = \begin{pmatrix} \mathbf{w}_k \\ \frac{b_k}{R} \end{pmatrix}$, and $\tilde{\mathbf{w}}_* = \begin{pmatrix} \mathbf{w}_* \\ \frac{b_*}{R} \end{pmatrix}$, and $\tilde{\mathbf{x}}_i = \begin{pmatrix} \mathbf{x}_i \\ R \end{pmatrix}$.
Observe that $\| \tilde{\mathbf{x}}_i \|^2 = \| \mathbf{x} \|^2 + R^2$.

Proof (cont'd)

Note that

$$\begin{aligned}
\tilde{\mathbf{w}}_{k+1} &= \begin{pmatrix} \mathbf{w}_{k+1} \\ \frac{b_{k+1}}{R} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{w}_k + \eta y_i \mathbf{x}_i \\ \frac{b_k + \eta y_i R^2}{R} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{w}_k \\ \frac{b_k}{R} \end{pmatrix} + \eta y_i \begin{pmatrix} x_i \\ R \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{w}_k \\ \frac{b_k}{R} \end{pmatrix} + \tilde{\mathbf{x}}_i \\
&= \tilde{\mathbf{w}}_k + \eta y_i \tilde{\mathbf{x}}_i .
\end{aligned}$$

Proof (cont'd)

Since $y_i\tilde{\mathbf{w}}'_*\tilde{\mathbf{x}}_i = y_i(\mathbf{w}'_*\mathbf{x}_i + b_*) \geqslant \gamma$, and $\tilde{\mathbf{w}}_{k+1} = \tilde{\mathbf{w}}_k + \eta y_i\tilde{\mathbf{x}}_i$ it follows that:

$$\tilde{\mathbf{w}}_*\tilde{\mathbf{w}}_{k+1} = \tilde{\mathbf{w}}_*\tilde{\mathbf{w}}_k + \eta y_i\tilde{\mathbf{w}}'_*\mathbf{w}_k \geqslant \tilde{\mathbf{w}}'_*\tilde{\mathbf{w}}_k + \eta\gamma.$$

because $y_i\tilde{w}'_*\tilde{x}_i \geqslant \gamma$.
By repeated application of the above inequality we obtain

$$\tilde{\mathbf{w}}_*\mathbf{w}_k \geqslant k\eta\gamma.$$

Proof (cont'd)

If the $k^{\rm th}$ error occurs on input $\mathbf{x}_i$ we have $\tilde{\mathbf{w}}_{k+1} = \tilde{\mathbf{w}}_k + \eta y_i \tilde{\mathbf{x}}_i$. This implies

$$\begin{aligned}
\parallel \tilde{\mathbf{w}}_{k+1} \parallel^2 &= \tilde{\mathbf{w}}'_{k+1}\tilde{\mathbf{w}}_{k+1} = (\tilde{\mathbf{w}}'_k + \eta y_i \tilde{\mathbf{x}}'_i)(\tilde{\mathbf{w}}_k + \eta y_i \tilde{\mathbf{x}}_i) \\
&= \parallel \tilde{\mathbf{w}}_k \parallel^2 + 2\eta y_i \tilde{\mathbf{w}}'_k \tilde{\mathbf{x}}_i + \eta^2 \parallel \tilde{\mathbf{x}}_i \parallel^2 \\
& \quad \text{(because } y_i \tilde{\mathbf{w}}'_k \mathbf{x}_i < 0 \text{ when an error occurs and } y_i^2 = 1) \\
&\leqslant \parallel \tilde{\mathbf{w}}_k \parallel^2 + \eta^2 \parallel \tilde{\mathbf{x}}_i \parallel^2 \\
&\leqslant \parallel \tilde{\mathbf{w}}_k \parallel^2 + \eta^2 (\parallel \mathbf{x}_i \parallel^2 + R^2) \\
&\leqslant \parallel \tilde{\mathbf{w}}_k \parallel^2 + 2\eta^2 R^2,
\end{aligned}$$

where we took into account that $\parallel \tilde{\mathbf{x}}_i \parallel^2 = \parallel \mathbf{x} \parallel^2 + R^2$.

Proof (cont'd)

Therefore, $\parallel \tilde{\mathbf{w}}_k \parallel^2 \leqslant 2k\eta^2 R^2$, hence $\parallel \tilde{\mathbf{w}}_k \parallel \leqslant \eta R\sqrt{2k}$. By combining the equalities

$$\tilde{\mathbf{w}}_* \mathbf{w}_k \geqslant k\eta\gamma \text{ and } \parallel \tilde{\mathbf{w}}_k \parallel \leqslant \eta R\sqrt{2k}$$

we obtain

$$\parallel \tilde{\mathbf{w}}_* \parallel \eta R\sqrt{2k} \geqslant \parallel \tilde{\mathbf{w}}_* \parallel \cdot \parallel \tilde{\mathbf{w}}_k \parallel \geqslant \tilde{\mathbf{w}}'_* \tilde{\mathbf{w}}_k \geqslant k\eta\gamma,$$

which imply

$$k \leqslant 2\left(\frac{R}{\gamma}^2\right) \parallel \tilde{\mathbf{w}}_* \parallel^2 \leqslant \left(\frac{2R}{\gamma}\right)^2$$

because $b_* \leqslant R$, hence $\parallel \tilde{\mathbf{w}}_* \parallel^2 \leqslant \parallel \mathbf{w}_* \parallel^2 + 1 = 2$.