Do-it Yourself Proof for Perceptron Convergence

Let $\mathbf{W}$ be a weight vector and $(\mathbf{I}, T)$ be a labeled example. Define $\mathbf{W} \cdot \mathbf{I} = \sum W_j I_j$. Let α be the learning rate. Use the following as the perceptron update rule:

```
if $\mathbf{W} \cdot \mathbf{I} < 1$ and $T = 1$
    then update the weights by:
        $W_j \leftarrow W_j + \alpha I_j$
if $\mathbf{W} \cdot \mathbf{I} > -1$ and $T = -1$
    then update the weights by:
        $W_j \leftarrow W_j - \alpha I_j$
```

Define Perceptron-Loss(T, O) as:

$$\text{Perceptron-Loss}(T, O) = \begin{cases} 0 & \text{if } T = -1 \text{ and } O \leq -1 \\ 0 & \text{if } T = 1 \text{ and } O \geq 1 \\ |T - O| & \text{otherwise} \end{cases}$$

How large is Perceptron-Loss$(T, \mathbf{W} \cdot \mathbf{I})$ if $T \neq \text{sign}(\mathbf{W} \cdot \mathbf{I})$?

Let $\mathbf{U}$ be the optimal weight vector. Let $\mathbf{W}'$ be the weight vector after updating. Define $\|\mathbf{W}\| = \sqrt{\mathbf{W} \cdot \mathbf{W}}$. For a labeled example $(\mathbf{I}, 1)$ where $\mathbf{W} \cdot \mathbf{I} < 1$, show the following:

$$\begin{aligned} &\text{Perceptron-Loss}(1, \mathbf{W} \cdot \mathbf{I}) - \text{Perceptron-Loss}(1, \mathbf{U} \cdot \mathbf{I}) \\ &\leq \mathbf{U} \cdot \mathbf{I} - \mathbf{W} \cdot \mathbf{I} \\ &= \frac{\|\mathbf{U} - \mathbf{W}\|^2 - \|\mathbf{U} - \mathbf{W}'\|^2}{2\alpha} + \frac{\alpha \|\mathbf{I}\|^2}{2} \end{aligned}$$

Similarly, for a labeled example $(\mathbf{I}, -1)$ where $\mathbf{W} \cdot \mathbf{I} > -1$, show that:

$$\begin{aligned} &\text{Perceptron-Loss}(-1, \mathbf{W} \cdot \mathbf{I}) - \text{Perceptron-Loss}(-1, \mathbf{U} \cdot \mathbf{I}) \\ &\leq \mathbf{W} \cdot \mathbf{I} - \mathbf{U} \cdot \mathbf{I} \\ &= \frac{\|\mathbf{U} - \mathbf{W}\|^2 - \|\mathbf{U} - \mathbf{W}'\|^2}{2\alpha} + \frac{\alpha \|\mathbf{I}\|^2}{2} \end{aligned}$$

Now consider a sequence of m updates on m labeled examples $(\mathbf{I}_1, T_1)$, $(\mathbf{I}_2, T_2)$, and so on. This will result in a sequence of weight vectors $\mathbf{W}_1, \mathbf{W}_2, \ldots$. Suppose that the initial weight vector $\mathbf{W}_1$ is all zeroes. Suppose $n \geq \|\mathbf{I}_k\|^2$ for all examples, and $A \geq \|\mathbf{U}\|$. Show that:

$$\begin{aligned} &\sum_{k=1}^{m} \text{Perceptron-Loss}(T_k, \mathbf{W}_k \cdot \mathbf{I}_k) \\ &\leq \left(\sum_{k=1}^{m} \text{Perceptron-Loss}(T_k, \mathbf{U} \cdot \mathbf{I}_k) \right) + \frac{A^2}{2\alpha} + \frac{m\,\alpha\,n}{2} \end{aligned}$$

Find a value for the learning rate α that minimizes the above expression. How does the loss of the perceptron algorithm compare to the loss of the optimal weights?

Do-it Yourself Proof for LMS Convergence

Let $\mathbf{W}$ be a weight vector and $(\mathbf{I}, T)$ be an example with a numeric outcome T. Define $\mathbf{W} \cdot \mathbf{I} = \sum W_j I_j$. Use the following as the update rule:

$$W_j \leftarrow W_j + \alpha(T - \mathbf{W} \cdot \mathbf{I}) I_j$$

Define $\text{Square-Loss}(T, O) = (T - O)^2$

Let $\mathbf{U}$ be the optimal weight vector. Let $\mathbf{W}'$ be the weight vector after updating. Define $\|\mathbf{W}\| = \sqrt{\mathbf{W} \cdot \mathbf{W}}$. For an example $(\mathbf{I}, T)$, show the following:

$$\begin{aligned} &\text{Square-Loss}(T, \mathbf{W} \cdot \mathbf{I}) - \text{Square-Loss}(T, \mathbf{U} \cdot \mathbf{I}) \\ &\quad \le \ 2(T - \mathbf{W} \cdot \mathbf{I})(\mathbf{U} \cdot \mathbf{I} - \mathbf{W} \cdot \mathbf{I}) \\ &\quad = \ \frac{\|\mathbf{U} - \mathbf{W}\|^2 - \|\mathbf{U} - \mathbf{W}'\|^2}{\alpha} + \alpha \|\mathbf{I}\|^2 \text{Square-Loss}(T, \mathbf{W} \cdot \mathbf{I}) \end{aligned}$$

Let $n \ge \|\mathbf{I}\|^2$. Show that the above implies:

$$\begin{aligned} &\text{Square-Loss}(T, \mathbf{W} \cdot \mathbf{I}) \\ &\quad \le \ \frac{\text{Square-Loss}(T, \mathbf{U} \cdot \mathbf{I})}{1 - \alpha\, n} + \frac{\|\mathbf{U} - \mathbf{W}\|^2 - \|\mathbf{U} - \mathbf{W}'\|^2}{\alpha(1 - \alpha\, n)} \end{aligned}$$

Now consider a sequence of m updates on m examples $(\mathbf{I}_1, T_1)$, $(\mathbf{I}_2, T_2)$, and so on. This will result in a sequence of weight vectors $\mathbf{W}_1, \mathbf{W}_2, \ldots$. Suppose that the initial weight vector $\mathbf{W}_1$ is all zeroes. Suppose $n \ge \|\mathbf{I}_k\|^2$ for all examples, and $A \ge \|\mathbf{U}\|$. Show that:

$$\begin{aligned} &\sum_{k=1}^{m} \text{Square-Loss}(T_k, \mathbf{W}_k \cdot \mathbf{I}_k) \\ &\quad \le \ \frac{\sum_{k=1}^{m} \text{Square-Loss}(T_k, \mathbf{U} \cdot \mathbf{I}_k)}{1 - \alpha n} + \frac{A^2}{\alpha(1 - \alpha n)} \end{aligned}$$

Suppose we choose $\alpha = 1/(2n)$. What does this say about the convergence of gradient descent?

References

The proof that the perceptron algorithm minimizes Perceptron-Loss comes from [1]. Tighter proofs for the LMS algorithm can be found in [2, 3].

[1] T. Bylander. Worst-case analysis of the perceptron and exponentiated update algorithms. *Artificial Intelligence*, 106:335–352, 1998.

[2] N. Cesa-Bianchi, P. M. Long, and M. K. Warmuth. Worst-case quadratic loss bounds for a generalization of the Widrow-Hoff rule. *IEEE Transactions on Neural Networks*, 7:604–619, 1996.

[3] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132:1–63, 1997.