Marginal Perceptron Convergence Proof

In this proof, the following assumptions are made

* Fixed increment, $\eta(i) = \eta$ = constant > 0
  for ease we assume $\eta = 1$ [without loss of generality] — (1)

* All the data points are considered to be linearly seperable

* Sequential Gradient Descent technique is used

* $w(0)$ is arbitary.

$w(i+1) = w(i) + z_i x_i \eta$, where $w_{(i)}^T z_i x_i \leq B$ — (2)
where $z_i x_i$ represents the reflected data points. that are cyclically ordered.

and this eqn (2) is iterated over many epochs.

from (1) as $\eta = 1$, we can re-write them as

$w(i+1) = w(i) + z_i x_i$ — (3)

Let us make another assumption,
$z^i x^i$ where $i = 0, 1, 2 \ldots c$ be the subset of training data points that are misclassified

at each epoch's iteration.

$\therefore$ We can start the Margin Perceptron algorithm as

$w(0) =$ arbitary value

$w(i+1) = w(i) + z^i x^i$ i - denotes i^{th} iteration

when $w_{(i)}^T z^i x^i \leq B \;\; \forall i$

If $\underline{\hat{w}}$ is a (weigh vector) solutions

i.e $\hat{\underline{w}}^T z_n x_n > B \;\; \forall n$

then $a\,\underline{\hat{w}}$ is also a solution

i.e $a\,\underline{\hat{w}}^T z_n x_n > B \;\; \forall n$ [iff $a > 0$] — (4)

We know that the weight vector w has some error value that keeps decreasing in each iteration. Let that error value be defined by

$$\mathcal{E}_w(i) = \| w_{(i)} - a\underline{\hat{w}} \|_2^2 \quad \text{—— (5)}$$

from equations (3) & (4), we can say that

$$w_{(i+1)} - a\,\underline{\hat{w}} = w(i) + z^i x^i - a\,\underline{\hat{w}}$$

$\Rightarrow$

$$[w(i+1) - a\hat{\underline{w}}] = [w(i) - a\hat{\underline{w}}] + z^i x^i$$

for $a > 0$.

Squaring on both sides in term of norm

$$\|w(i+1) - a\hat{\underline{w}}\|_2^2 = \|w(i) - a\hat{\underline{w}}\|^2 + 2\,(w(i) - a\hat{\underline{w}})^T z^i x^i + \|z^i x^i\|_2^2$$

$$\Rightarrow \|w(i+1) - a\hat{\underline{w}}\|_2^2 = \|w(i) - a\hat{\underline{w}}\|_2^2 + 2\,w^T z^i x^i - 2a\hat{\underline{w}} z^i x^i + \|z_i^i x^i\|_2^2$$

$\Rightarrow$

$$\|w(i+1) - a\hat{\underline{w}}\|_2^2 = \left(\|w(i) - a\hat{w}\|_2^2 - 2a\hat{\underline{w}} z^i x^i + \|z^i x^i\|_2^2\right) + 2w^T z^i x^i$$

$$\|w(i+1) - a\hat{w}\|_2^2 \le \left(\|w(i) - a\hat{w}\|_2^2 - 2a\hat{w} z^i x^i + \|z^i x^i\|_2^2\right) + 0 \quad (6)$$

Inorder to Maximize the equation at margin we need to minimize $\hat{w}^T z^i x^i$ and maximize $\|z_i^i x^i\|_2^2$ value such that

$$\hat{\underline{w}}^T z^i x^i > B \text{ and } \|z^i x^i\|_2^2 > 0$$

Lets assume variables to make eqn simpler

let $b^2 = \max_j \|x_j\|_2^2$

$c = \min_j \{ \underline{\hat{\omega}}^T z^i x^i \} > B$

$\therefore$ we can re-write eqn (6) as

$$\|w(i+1) - a\underline{\hat{\omega}}\|_2^2 \le \|w(i) - a\underline{\hat{\omega}}\|_2^2 - 2ac + b^2 \quad (7)$$

$a > 0$

from (5) wkt $\|w(i) - a\hat{\omega}\|_2^2$ is the error in weight vector.

$\therefore$ (7) $\Rightarrow$

$$\varepsilon_{e_\omega}(i+1) \le \varepsilon_{e_\omega}(i) - 2ac + b^2$$

if we choose a such that $a = \frac{b^2}{c}$

$$\Rightarrow \varepsilon_{e_\omega}(i+1) \le \varepsilon_{e_\omega}(i) - \frac{2b^2}{\cancel{c}}\cancel{c} + b^2$$

$$\Rightarrow \varepsilon_{e_\omega}(i+1) \le \varepsilon_{e_\omega}(i) - b^2$$

$\therefore$ we see that each iteration, error value is decreased by 'b^2'. $[b^2 > 0]$

$\therefore$ we can write as

$$0 \le \varepsilon_{\omega}(i+1) \le \varepsilon_{\omega}(i) - \delta^2 \quad - (8)$$

at some iteration (i_0), $\varepsilon_{\omega}(i_0) < \delta^2$

$\therefore$ eqn (8) becomes implausible

therefore the algorithm's iteration stops at $(i_0 - 1)^{th}$ iteration, and the corresponding weight vector is the converged weight vector.

3)

$$\cancel{\Delta w(i+1)} \quad \Delta w(i) = w(i+1) - w(i) \quad \text{--- (1)}$$

$$J(w) = \sum_{n=1}^{N} J_n(w) \quad \text{--- (2)}$$

$$\eta(i) = \eta = \text{const} > 0$$

a) Using Stochastic Gradient Descent, variant 2, a training data point is randomly picked with replacement and single sample update is done.

In Gradient Descent

$$w(i+1) = w(i) - \eta \sum_{n=1}^{N} \nabla J_n(w) \quad \text{--- (3)}$$

for single sample update;

$$(3) \Rightarrow \quad w(i+1) = w(i) - \eta \nabla J_n(w) \quad \text{--- (4)}$$

And the expected value of $\Delta w(i)$

$$E\{\Delta \underline{w}(i)\} = \sum_{n=1}^{N} \Delta w(i)\, p(n)$$

where $P(n)$ = probability with which points are chosen and $n = 1, 2, \ldots, N$. with uniform probability.

$$P(n) = \frac{1}{(N-1)+1} = \frac{1}{N}$$

$$\therefore E\{\Delta w(i)\} = \sum_{n=1}^{N} \Delta w(i) \times \frac{1}{N}$$

1 $\Rightarrow$

$$E\{\Delta w(i)\} = \sum_{n=1}^{N} (w(i+1) - w(i)) \times \frac{1}{N}$$

4 $\Rightarrow$

$$= \sum_{n=1}^{N} -\eta \nabla J_n(w) \times \frac{1}{N}$$

$$= -\frac{\eta}{N} \sum_{n=1}^{N} \nabla J_n(w)$$

$$E\{\Delta w(i)\} = -\frac{\eta}{N} \nabla \sum_{n=1}^{N} J_n(w)$$

(D)

2 $\Rightarrow$ $E\{\Delta w(i)\} = -\frac{\eta}{N} \nabla J(w)$

b) Find $E\left\{\sum_{n=0}^{N-1} \Delta w(i)\right\}$

where each $\Delta w(i)$ is IID.

from (a) $E\{\Delta w(i)\} = -\frac{\eta}{N} \nabla \sum_{n=1}^{N} J_n(w)$

(a)

$= -\frac{\eta}{N} \nabla J(w)$

$E\left\{\sum_{n=0}^{N-1} \Delta w(i)\right\} = \sum_{n=0}^{N-1} E\{\Delta w(i)\}$

[$E[aX+b] = aE[X]+b$]

$= (N-1)+1 \; E\{\Delta w(i)\}$

$= \cancel{N} \cdot -\frac{\eta}{\cancel{N}} \nabla J(w)$

$E\left\{\sum_{n=0}^{N-1} \Delta w(i)\right\} = -\eta \nabla J(w)$

(B)

$-\eta \nabla \sum_{n=1}^{N} J_n(w)$

c) Batch Gradient Desient :

$$w(i+1) = w(i) + \eta(i) \sum z_n x_n \quad \forall_n$$

$$= w(i) - \eta(i) \nabla_w J(w)$$

$$= w(i) - \eta(i) \nabla_w \sum_{n=1}^{N} J_n(w)$$

$$\longrightarrow (1)$$

$$w(i+1) - w(i) = -\eta(i) \nabla_w \sum_{n=1}^{N} J_n(w)$$

$$= E\left\{ \sum_{n=0}^{N-1} \Delta w(i) \right\}$$

$$[\because (b)]$$

part (c) result is similar to part (b)

i.e in batch gradient desient, we consider all the misclassified data points

but in stochastic gradient, we consider single sample and repeat for 'N' samples. It is computationally stable because it can be used for a set with more minima/ maxima