LINEAR CLASSIFIERS

• **Γραμμικές Συναρτήσεις Διάκρισης και Υπερεπίπεδα Απόφασης**

ΥΠΟΘΕΣΕΙΣ:

→ ΔΥΑΔΙΚΟ ΠΡΟΒΛΗΜΑ ΤΑΞΙΝΟΜΗΣΗΣ. (***)

→ ΣΥΝΑΡΤΗΣΕΙΣ ΔΙΑΚΡΙΣΗΣ ΜΕΤΑΞΥ ΚΛΑΣΕΩΝ ΓΡΑΜΜΙΚΕΣ. ***

→ ΥΠΕΡΕΠΙΦΑΝΕΙΕΣ ΑΠΟΦΑΣΗΣ (DECISION HYPER-PLANES) ΣΤΟΝ ℓ-διάστατο ΧΩΡΟ ΤΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΕΚΦΡΑΖΟΝΤΑΙ ΩΣ ΥΠΕΡΕΠΙΠΕΔΑ. (***)

$$\boxed{g(\underline{x}) = \underline{w}^T \cdot \underline{x} + w_0 = 0 \quad [1] \quad \text{όπου } \underline{w}, \underline{x} \in \mathbb{R}^{\ell} \text{ και } w_0 \in \mathbb{R}}$$

• Το διάνυσμα των βαρών $\underline{w}$ (weight vector) και το κατώφλι (w_0) (threshold / bias term) αποτελούν τις βασικές παραμέτρους του υπερ-επιπέδου απόφασης.

• Αν δύο διανύσματα $\underline{x}_1, \underline{x}_2 \in \mathbb{R}^{\ell}$ βρίσκονται επάνω στο υπερ-επίπεδο απόφασης θα ισχύει ότι:

$$0 = \underline{w}^T \underline{x}_1 + w_0 = \underline{w}^T \underline{x}_2 + w_0 \Longrightarrow$$

$$\boxed{\underline{w}^T(\underline{x}_1 - \underline{x}_2) = 0 \quad [2]}$$

Συμπεράσματα:

(A) ⇒ Το διάνυσμα της διαφοράς των διανυσμάτων $(\underline{x}_1 - \underline{x}_2)$ βρίσκεται επάνω στο υπερ-επίπεδο απόφασης καθώς επαληθεύει την εξίσωση [1].

(B) ⇒ Η εξίσωση [2] υποδεικνύει πως το διάνυσμα $\underline{w}$ είναι ορθογώνιο στο υπερεπίπεδο απόφασης.

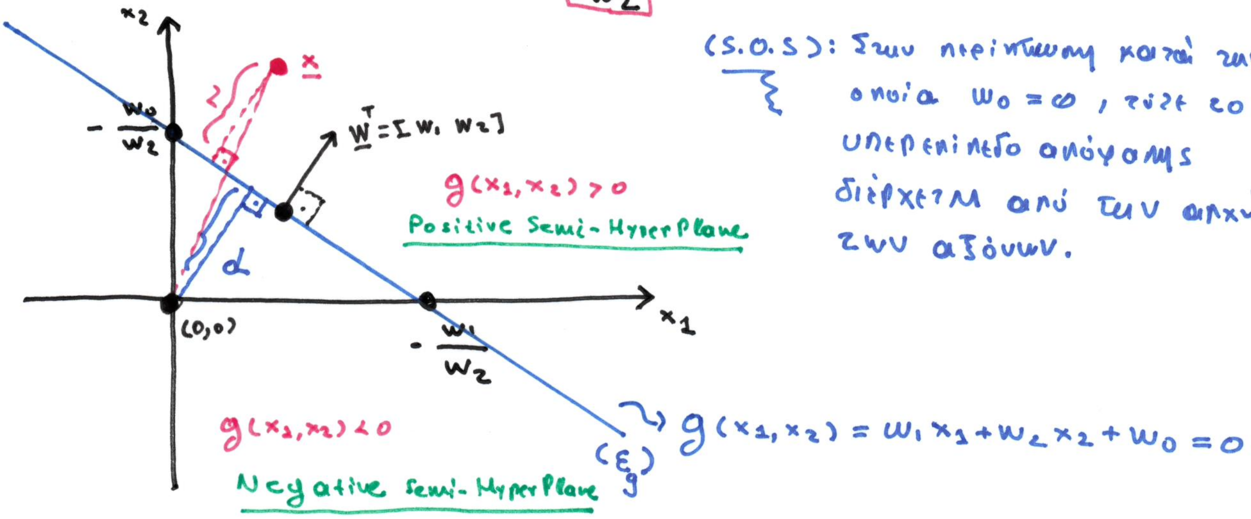
ΣΧΗΜΑΤΙΚΗ ΑΝΑΠΑΡΑΣΤΑΣΗ (Χώρος χαρακτηριστικών με $\ell = 2$)

- Έστω ότι: $\boxed{w_1, w_2 > 0}$ και $\boxed{w_0 < 0}$.
- Έχουμε ότι:

(α): $\boxed{g(x_1, x_2) = w_1 x_1 + w_2 x_2 + w_0 = 0}$

(β): $\boxed{\text{Για } x_1 = 0 \Rightarrow \hat{x}_2 = -\frac{w_0}{w_2}}$

(γ): $\boxed{\text{Για } x_2 = 0 \Rightarrow \hat{x}_1 = -\frac{w_1}{w_2}}$



(S.O.S): Στην περίπτωση κατά την οποία $w_0 = 0$, τότε το υπερεπίπεδο απόφασης διέρχεται από την αρχή των αξόνων.

- Ο συντελεστής διεύθυνσης της ευθείας $g(x_1, x_2) = 0$ (στην ειδική περίπτωση όπου $\ell = 2$) μπορεί να υπολογισθεί ως εξής:

$$\boxed{x_2 = -\frac{w_1}{w_2} \cdot x_1 - \frac{w_0}{w_2} \quad [3]}$$

Από την σχέση [3] έχουμε ότι: $\boxed{\lambda_g = -\frac{w_1}{w_2} \quad [4]}$

- Ο συντελεστής διεύθυνσης του διανύσματος $\underline{w} = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix}$ θα δίνεται προφανώς από την σχέση: $\boxed{\lambda_{\underline{w}} = \frac{w_2}{w_1} \quad [5]}$

- Με βάση τα παραπάνω, θα έχουμε ότι:

$$\boxed{\lambda_{\underline{w}} \cdot \lambda_g = -1 \quad [6]}$$

Που σημαίνει πως το διάνυσμα $\underline{w}$ είναι κάθετο στην ευθεία g:

$$\boxed{\underline{w} \perp \varepsilon_g \quad [7]}$$

ΑΠΟΣΤΑΣΗ ΣΗΜΕΙΟΥ ΑΠΟ ΥΠΕΡ-ΕΠΙΠΕΔΟ:

* Θεωρούμε το υπερ-επίπεδο $g(\underline{x}) = \underline{w}^T\underline{x} + w_0 = 0$.

* Θέλουμε να υπολογίσουμε την απόσταση δοσμένου σημείου $\underline{x_0}$ από το υπερεπίπεδο.

* Να διατυπωθεί το συγκεκριμένο πρόβλημα ως πρόβλημα βελτιστοποίησης.

* **ΔΙΑΤΥΠΩΣΗ ΠΡΟΒΛΗΜΑΤΟΣ:** Η απόσταση $d(\underline{x_0}, g(\underline{x}))$ του δοσμένου σημείου $\underline{x_0}$ από το υπερεπίπεδο που δίνεται από την σχέση $g(\underline{x}) = 0$ αντιστοιχεί στην ελάχιστη απόσταση που μπορεί να έχει ένα σημείο του υπερεπιπέδου $\underline{x}$ από το σημείο $\underline{x_0}$.

* **OPTIMIZATION PROBLEM:**

$$\min_{\underline{x} \in \mathbb{R}^\ell} \|\underline{x} - \underline{x_0}\|^2 \quad [8]$$

$$\text{s.t.} \quad g(\underline{x}) = 0$$

* Let $f(\underline{x}, \underline{x_0}) = \|\underline{x} - \underline{x_0}\|^2$ [9] be the objective/cost function of the optimization problem.

* The required distance value is given as:

$$\left\{ \underline{x}^* = \arg\min_{\substack{\underline{x} \in \mathbb{R}^\ell \\ \text{s.t. } g(\underline{x}) = 0}} \|\underline{x} - \underline{x_0}\|^2 \quad \text{and} \quad d(\underline{x_0}, g(\underline{x})) = \|\underline{x}^* - \underline{x_0}\| \right\}$$

* **Lagrangian Function**:

$$\boxed{\mathcal{L}(\underline{x}; \lambda) = f(\underline{x}, \underline{x_0}) + \lambda \cdot g(\underline{x}) \quad [10]}$$

* **Karush-Kuhn-Tucker (KKT) Conditions:**

(α): **Stationarity Conditions**

$$\nabla_{\underline{x}} \mathcal{L}(\underline{x}; \lambda) = \underline{0}_{[\ell \times 1]} \quad [11]$$

(β): **Primal Feasibility Conditions**

$$g(\underline{x}) = 0 \in \mathbb{R} \quad [12]$$

(γ): **Dual Feasibility Conditions**

$$\lambda \geq 0$$

⊛ Solve the system of Eqs. (11), (12):

$$\begin{cases} \dfrac{\partial}{\partial \underline{x}} \left\{ \| \underline{x} - \underline{x}_0 \|^2 + \lambda (\underline{w}^T \underline{x} + w_0) \right\} = \underline{0} & [13] \\ \underline{w}^T \underline{x} + w_0 = 0 & [14] \end{cases} \Longrightarrow$$

$$\begin{cases} \dfrac{\partial}{\partial \underline{x}} \left\{ \underline{x}^T \underline{x} - 2 \underline{x}^T \underline{x}_0 + \underline{x}_0^T \underline{x}_0 + \lambda \underline{w}^T \underline{x} + \lambda w_0 \right\} = \underline{0} & [15] \\ \underline{w}^T \underline{x} + w_0 = 0 & [16] \end{cases} \Longrightarrow$$

$$\begin{cases} 2 \underline{x} - 2 \underline{x}_0 + \lambda \underline{w} = \underline{0} & [17] \\ \underline{w}^T \underline{x} + w_0 = 0 & [18] \end{cases}$$

⊛ Eq. (17) yields : $2 (\underline{x} - \underline{x}_0) = - \lambda \underline{w} \Rightarrow \underline{x} - \underline{x}_0 = - \frac{\lambda}{2} \underline{w} \Rightarrow$

$$\boxed{\underline{x}^* = \underline{x}_0 - \frac{\lambda}{2} \underline{w} \quad [19]}$$

⊛ The next step is to derive the optimal value for λ, λ^*:

Eq. (18) yields : $\underline{w}^T \cdot \underline{x}^* + w_0 = 0 \Rightarrow$

$$\underline{w}^T \left(\underline{x}_0 - \frac{\lambda}{2} \underline{w} \right) + w_0 = 0 \Rightarrow$$

$$\underline{w}^T \underline{x}_0 - \frac{\lambda}{2} \underline{w}^T \underline{w} + w_0 = 0 \Rightarrow$$

$$\frac{\lambda}{2} \underline{w}^T \underline{w} = \underline{w}^T \underline{x}_0 + w_0 \Rightarrow$$

$$\frac{\lambda}{2} \| \underline{w} \|^2 = \underline{w}^T \underline{x}_0 + w_0 \Rightarrow$$

$$\boxed{\lambda^* = \frac{2}{\| \underline{w} \|^2} \cdot \left(\underline{w}^T \underline{x}_0 + w_0 \right) \quad (20)}$$

(*) Remember, however, that our ultimate goal is to compute the quantity:

$$\boxed{d(\underline{x}_0 ; g(\underline{x})) = d_{min} = \| \underline{x}^* - \underline{x}_0 \| = \sqrt{\| \underline{x}^* - \underline{x}_0 \|^2} \quad (21)}$$

(*) Taking into consideration Eq. (19), yields:

$$\boxed{d_{min} = \sqrt{\| - \lambda^*/2 \cdot \underline{w} \|^2} = \frac{1}{2} \cdot |\lambda^*| \cdot \| \underline{w} \| \quad (22)}$$

(*) Substituting λ^* from Eq. (20) into (22) yields:

$$d_{min} = \frac{1}{\cancel{2}} \cdot \left| \frac{\cancel{2}}{\|\underline{w}\|^{\cancel{2}}} \cdot (\underline{w}^T \cdot \underline{x}_0 + w_0) \right| \cdot \cancel{\| \underline{w} \|} \Rightarrow$$

$$\boxed{d_{min} = \frac{|\underline{w}^T \underline{x}_0 + w_0|}{\| \underline{w} \|} \quad [23]}$$

or equivalently:

$$\boxed{d_{min} \equiv d = \frac{|g(\underline{x}_0)|}{\| \underline{w} \|} \quad [24]}$$

ΑΛΓΟΡΙΘΜΟΣ PERCEPTRON

ΣΤΟΧΟΣ: Υπολογισμός των άγνωστων παραμέτρων του υπερεπιπέδου απόφασης $\{w_i, i \in \{0,1,2,\ldots,l\}\}$ μεταξύ των κλάσεων ω_1 και ω_2 που δίνεται από την σχέση $\boxed{g(\underline{x}) = \underline{w}^T\underline{x} + w_0 = 0}$ (1), υποθέτοντας πως οι δύο κλάσεις είναι γραμμικά διαχωρίσιμες.

Η γραμμική διαχωρισιμότητα των δύο κλάσεων προϋποθέτει πως είναι δυνατός ο προσδιορισμός των παραμέτρων $(\underline{w}, w_0)$ έτσι ώστε:

$$\boxed{\begin{cases} \forall\, \underline{x} \in \omega_1, & \underline{w}^T \cdot \underline{x} + w_0 > 0 \\ \forall\, \underline{x} \in \omega_2, & \underline{w}^T \cdot \underline{x} + w_0 < 0 \end{cases} \quad (2)}$$

★ Το υπερεπίπεδο απόφασης που ορίζεται μέσω της σχέσης (1) μπορεί να διατυπωθεί στην πιο συνοπτική μορφή ως εξής:

$$\boxed{g(\underline{x}') = \underline{w}'^T \cdot \underline{x}' = 0 \quad (3)}$$

ορίζοντας τα εκτεταμένα διανύσματα βαρών $(\underline{w}')$ και χαρακτηριστικών (x') ως εξής:

$$\boxed{\begin{aligned} \underline{x}' &= [\underline{x}^T, 1]^T \quad (4) \\ \underline{w}' &= [\underline{w}^T, w_0]^T \quad (5) \end{aligned}}$$

★ Από το σημείο αυτό και μετά θεωρούμε πως $\underline{w} \equiv \underline{w}'$ και $\underline{x} = \underline{x}'$.

Το πρόβλημα υπολογισμού των παραμέτρων ($\underline{w}'$) του ζητούμενου υπερεπιπέδου απόφασης μπορεί να διατυπωθεί ως πρόβλημα βελτιστοποίησης. Για τον σκοπό αυτό είναι απαραίτητο να υιοθετήσουμε:

(A): Ένα κατάλληλο συναρτησιοειδές κόστους

(B): Αλγόριθμο βελτιστοποίησης

COST FUNCTIONAL:

$$\boxed{J(\underline{w}) = \sum_{\underline{x} \in Y} \delta_{\underline{x}} \cdot g(\underline{x}) \quad (6)}$$

Ισοδύναμα μπορούμε να γράψουμε ότι:

$$\boxed{J(\underline{w}) = \sum_{\underline{x} \in Y} \delta_{\underline{x}} \cdot (\underline{w}^T \cdot \underline{x}) \quad (7)}$$

Το σύνολο Y περιλαμβάνει τα διανύσματα των χαρακτηριστικών εκπαίδευσης που έχουν ταξινομηθεί εσφαλμένα από το υπερεπίπεδο απόφασης που καθορίζεται από το διάνυσμα βαρών $\underline{w}$.

Η μεταβλητή $\delta_{\underline{x}}$ ορίζεται ως εξής:

$$\boxed{\delta_{\underline{x}} = \begin{cases} -1, & \underline{x} \in \omega_1; \\ +1, & \underline{x} \in \omega_2. \end{cases} \quad (8)}$$

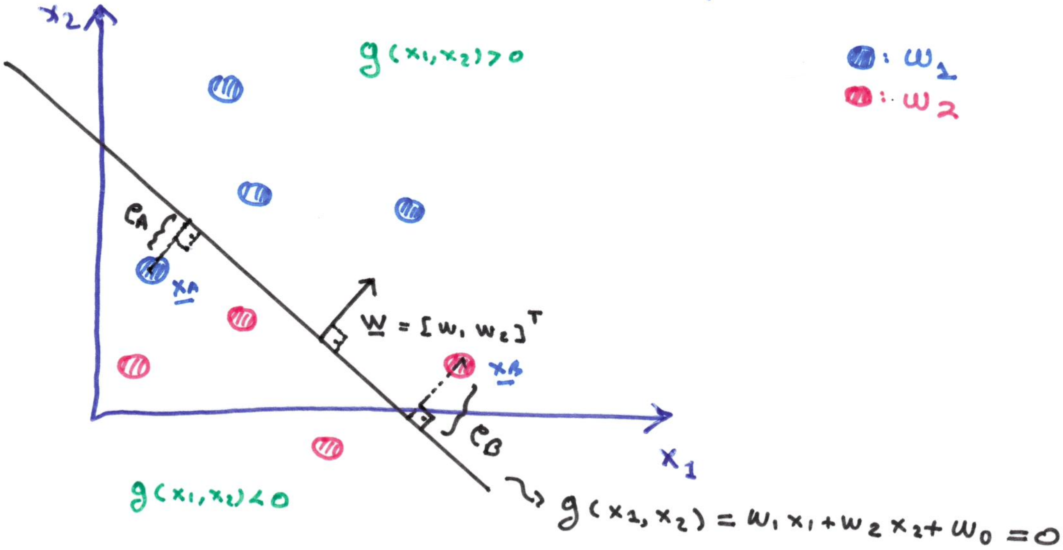
Είναι προφανές πως η ποσότητα $J(\underline{w})$ θα πρέπει να παραμένει θετική. Πράγματι έχουμε ότι:

$$\left\{ \forall \underline{x} \in Y, \ \text{sign}[g(\underline{x})] = \begin{cases} -1, & \underline{x} \in \omega_1; \\ +1, & \underline{x} \in \omega_2. \end{cases} \right\} \Rightarrow \boxed{\forall \underline{x} \in Y, \ \delta_{\underline{x}} \cdot \text{sign}[g(\underline{x})] = +1.} \quad (9)$$

Ισοδύναμα θα μπορούσαμε να γράψουμε ότι:

$$\boxed{\forall \underline{x} \in Y, \ \delta_{\underline{x}} \cdot g(\underline{x}) = |g(\underline{x})| \quad (10)}$$

⊛ Θεωρείστε το παρακάτω στιγμιότυπο ενός δυαδικού προβλήματος ταξινόμησης στον χώρο των χαρακτηριστικών $\mathcal{X} = \mathbb{R}^2$.



⊛ Γνωρίζουμε πως
$$\begin{cases} d(\underline{x_A}; g(\underline{x})) = \dfrac{|g(\underline{x_A})|}{\|\underline{w}\|} & (11) \\[2ex] d(\underline{x_B}; g(\underline{x})) = \dfrac{|g(\underline{x_B})|}{\|\underline{w}\|} & (12) \end{cases}$$

⊛ Σύμφωνα με την προηγούμενη ανάλυση, η συνάρτηση κόστους $J(\underline{w})$ συντίθεται υπολογίζοντας το στοιχειώδες σφάλμα για κάθε ένα από τα διανύσματα χαρακτηριστικών εκπαίδευσης. Μπορούμε να γράψουμε ότι:

$$\boxed{\forall \underline{x} \in \mathcal{Y},\quad e(\underline{x}) = \delta_{\underline{x}} \cdot g(\underline{x}) = |g(\underline{x})| = \|\underline{w}\| \cdot d(\underline{x}, g(\underline{x}))} \quad (13)$$

$$\boxed{e(\underline{x}) = \|\underline{w}\| \cdot d(\underline{x}, g(\underline{x})) \quad (14)}$$ ή διαφορετικά

$$\boxed{d(\underline{x}, g(\underline{x})) \propto e(\underline{x}),\ \forall \underline{x} \in \mathcal{Y} \quad (15)}$$

* Η συνάρτηση κόστους $J(\underline{w})$ λαμβάνει την ελάχιστη τιμή της, που συμβαίνει να είναι το μηδέν ($\min_{\underline{w}} J(\underline{w}) = 0$), στην περίπτωση κατά την οποία δεν υπάρχει κανένα διάνυσμα χαρακτηριστικών που να ταξινομείται λάθος.

* Το συναρτησιοειδές κόστους $J(\underline{w})$ αποτελεί μια γραμμική κατά τμήματα συνεχής συνάρτηση. {(linear piecewise continuous function)}

* Πράγματι, αν το διάνυσμα των βαρών υποστεί μια ομαλή μεταβολή, η συνάρτηση κόστους $J(\underline{w})$ θα μεταβληθεί γραμμικά μέχρι το σημείο όπου το σύνολο Y αλλάξει.

* Το διάνυσμα βαθμίδας (gradient vector $\nabla_{\underline{w}} J(\underline{w})$) στο συγκεκριμένο σημείο δεν ορίζεται και η παράγωγος της συνάρτησης κόστους $\frac{\partial J(\underline{w})}{\partial \underline{w}}$ δεν είναι συνεχής.

▶ Ο αλγόριθμος ελαχιστοποίησης της συνάρτησης κόστους που σχετίζεται με το Perceptron ακολουθεί το πνεύμα της μεθόδου της βαθμιαίας κατάβασης (gradient descent) και διατυπώνεται ως εξής:

$$\boxed{\underline{w}(t+1) = \underline{w}(t) - \rho_t \cdot \left.\frac{\partial J(\underline{w})}{\partial \underline{w}}\right|_{\underline{w}=\underline{w}(t)} \qquad (16)}$$

όπου $\underline{w}(t)$ είναι η εκτίμηση του διανύσματος των βαρών κατά την t-οστή επανάληψη και ρ_t μια ακολουθία θετικών πραγματικών αριθμών.

▶ Λαμβάνοντας υπόψιν το γεγονός πως η σχέση (16) ορίζεται εκτός των προαναφερθέντων σημείων ασυνέχειας, τότε για κάθε έγκυρο σημείο μπορούμε να γράψουμε ότι:

$$\frac{\partial J(\underline{w})}{\partial \underline{w}} = \sum_{\underline{x}\in Y} \frac{\partial}{\partial \underline{w}}\left\{\delta_{\underline{x}}\cdot(\underline{w}^T\cdot\underline{x})\right\} \Rightarrow$$

$$\frac{\partial J(\underline{w})}{\partial \underline{w}} = \sum_{\underline{x}\in Y} \delta_{\underline{x}}\cdot\frac{\partial}{\partial \underline{w}}\left\{\underline{w}^T\cdot\underline{x}\right\} \Rightarrow$$

$$\boxed{\frac{\partial J(\underline{w})}{\partial \underline{w}} = \sum_{\underline{x}\in Y} \delta_{\underline{x}}\cdot\underline{x} \quad [17]}$$

▶ Επομένως, μπορούμε να γράψουμε ότι:

$$\boxed{\underline{w}(t+1) = \underline{w}(t) - \rho_t\cdot\sum_{\underline{x}\in Y} \delta_{\underline{x}}\cdot\underline{x} \quad [18]}$$

★ Ο αλγόριθμος εκπαίδευσης του Perceptron αρχικοποιείται αποδίδοντας μια αυθαίρετη αρχική τιμή στο διάνυσμα των βαρών $\underline{w}(0)$ ενώ σε κάθε επανάληψη διαμορφώνεται ένα διάνυσμα διόρθωσης (correction vector) το οποίο βασίζεται στο σύνολο των διανυσμάτων χαρακτηριστικών που ταξινομούνται εσφαλμένα κατά την τρέχουσα επανάληψη:

$$\boxed{\underline{e} = \sum_{\underline{x}\in Y} \underline{e}_{\underline{x}} \quad (19)} \quad \text{όπου} \quad \boxed{\underline{e}_{\underline{x}} = \delta_{\underline{x}}\cdot\underline{x} \quad (20)}$$

★ Η διαδικασία ενημέρωσης των βαρών θα επαναλαμβάνεται μέχρι το σύνολο $Y_t = \{\emptyset\}$.

Perceptron Algorithm

Assuming $X = \{\underline{x}_1, \underline{x}_2, \dots, \underline{x}_N\}$ from known classes ω_1 and ω_2.

1: Choose $\underline{W}(0)$ randomly

2: Choose ρ_0

3: Initialize $t = 0$

4: Repeat

4.1: Initialize $Y = \{\emptyset\}$

4.2: For $i = 1$ to N

if ($\delta_{\underline{x}_i} \cdot \underline{W}^T(t) \cdot \underline{x}_i \geq 0$) then $Y = Y \cup \{\underline{x}_i\}$

End {For}

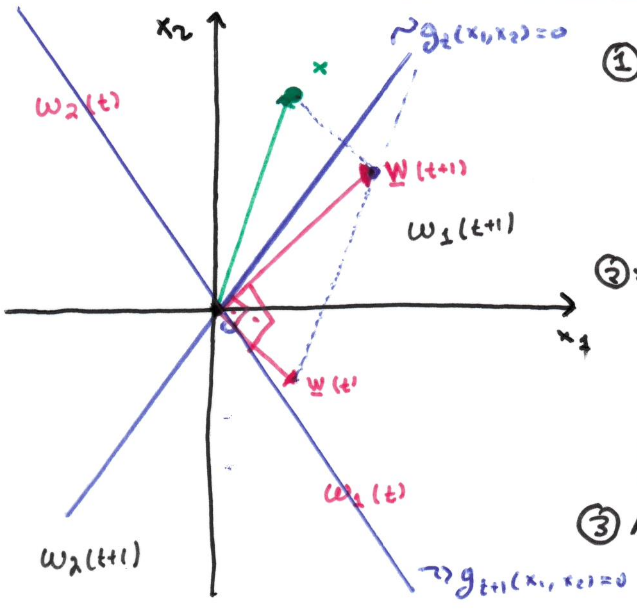
4.3: $\underline{W}(t+1) = \underline{W}(t) - \rho_t \cdot \sum_{\underline{x} \in Y} \delta_{\underline{x}} \cdot \underline{x}$

4.4: Adjust ρ_t

4.5: Increase t: $t = t+1$

Until: $Y = \{\emptyset\}$

Σχηματική Αναπαράσταση της Δράσης του Perceptron



① : Υποθέτουμε πως κατά την εκτέλεση του t-οστού βήματος του αλγορίθμου perceptron η τρέχουσα τιμή του διανύσματος των βαρών ($\underline{W}(t)$) είναι αυτή που φαίνεται στο σχήμα.

② : Αν το διάνυσμα $\underline{x}$ που φαίνεται στο σχήμα είναι το μοναδικό διάνυσμα χαρακτηριστικών που είναι εσφαλμένα ταξινομημένο με βάση την ευθεία $g_t(x_1, x_2) = 0$ (που προσδιορίζεται από το διάνυσμα $\underline{W}(t)$), τότε θα έχουμε ότι: $\boxed{Y_t = \{\underline{x}\}}$. Έστω ότι $\boxed{\rho_t = 1}$.

③ Με βάση την τρέχουσα διαμόρφωση του σχήματος, το διάνυσμα $\underline{x}$ ανατίθεται εσφαλμένα στην κλάση ω_2 και κατά συνέπεια ανήκει πραγματικά στην ω_1, άρα $\boxed{\underline{x} \in \omega_1}$. Αυτό σημαίνει πως $\boxed{\delta_{\underline{x}} = -1}$.

④ Κατά συνέπεια, το διάνυσμα της διόρθωσης του διανύσματος των βαρών $\underline{c}$ θα δίνεται από την σχέση: $\boxed{\underline{c} = \delta_{\underline{x}} \cdot \underline{x}} \Rightarrow \boxed{\underline{c} = -\underline{x}}$. Επομένως, το ανανεωμένο διάνυσμα των βαρών κατά την επανάληψη $(t+1)$ θα δίνεται ως: $\boxed{\underline{W}(t+1) = \underline{W}(t) - \underline{c}_{\underline{x}}} \Rightarrow$

$\boxed{\underline{W}(t+1) = \underline{W}(t) - (-\underline{x}) \Rightarrow}$ $\boxed{\underline{W}(t+1) = \underline{W}(t) + \underline{x}}$.