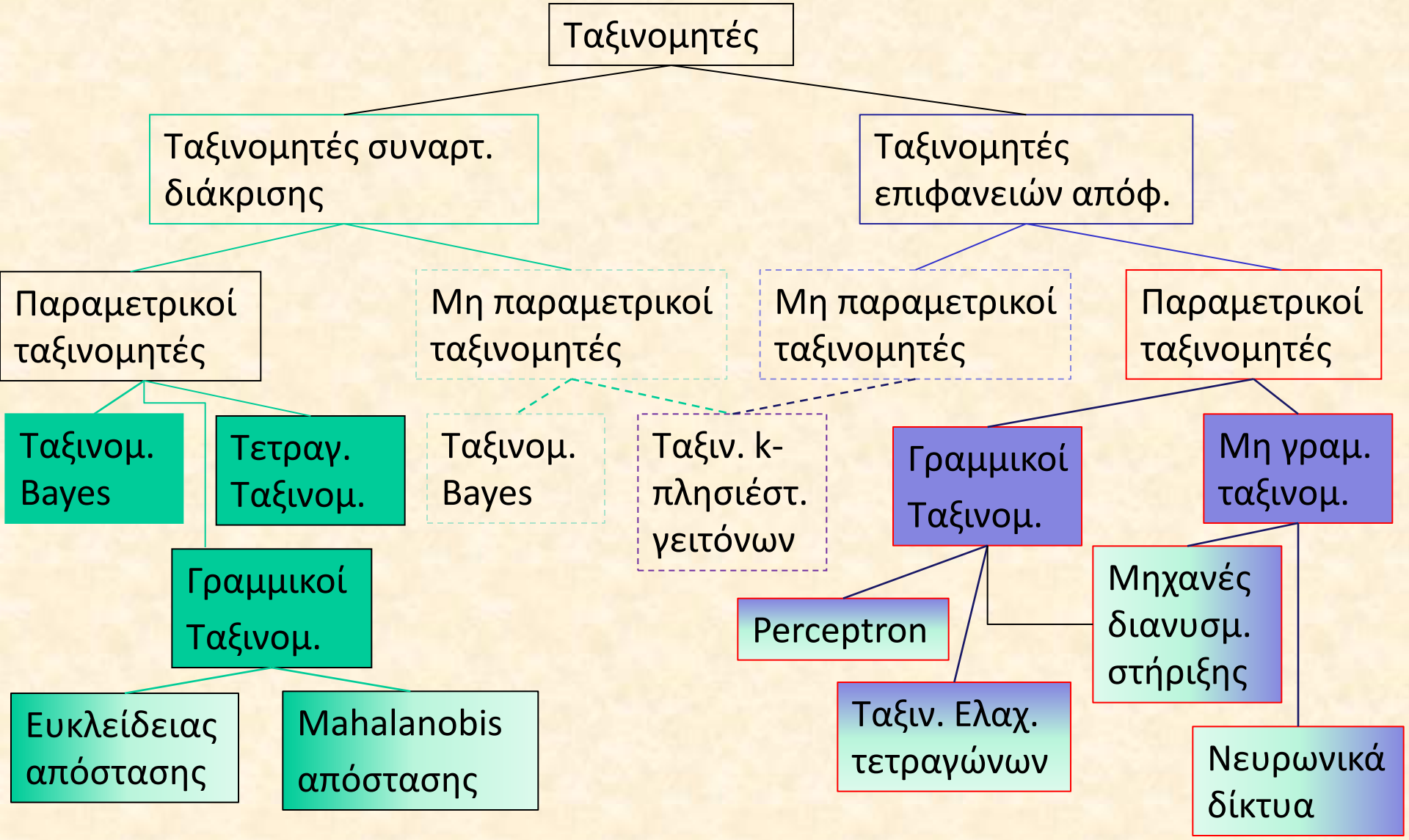
❖ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ
❖(PATTERN RECOGNITION)

Σέργιος Θεοδωρίδης

Κωνσταντίνος Κουτρούμπας

“ΧΑΡΤΟΓΡΑΦΗΣΗ” ΤΟΥ ΧΩΡΟΥ ΤΩΝ ΤΑΞΙΝΟΜΗΤΩΝ



"ΧΑΡΤΟΓΡΑΦΗΣΗ" ΤΟΥ ΧΩΡΟΥ ΤΩΝ ΤΑΞΙΝΟΜΗΤΩΝ

Υπενθ.: X είναι το σύνολο των δεδομένων σημείων όλων των κλάσεων

X_j είναι το υποσύνολο του X που περιέχει τα διανύσματα της κλάσης ω_j,

$X=X_1 \cup \ldots \cup X_M$

Ταξινομητές με βάση τις συναρτήσεις διάκρισης

Ταξινομ. Bayes

- $g_j(x) = f(P(\omega_j)p(x|\omega_j))$
- Εκτιμ. $p(x|\omega_j) \approx \hat{p}(x|\omega_j;\vartheta_j)$
- Εκτιμ. ϑ_j, με βάση το X_j

$$(ML,\ EM): X_j \to \hat{\vartheta}_j$$

- $x_i \to p(x_i|\omega_j) \approx \hat{p}(x_i|\omega_j;\hat{\vartheta}_j)$

Τετραγωνικός ταξινομητής

- $g_j(x) = (x-\mu_j)^T \Sigma_j^{-1}(x-\mu_j)$
- Υπόθεση : $N(\mu_j, \Sigma_j)$
- Εκτιμ. μ_j, Σ_j, με βάση το X_j

$$(ML): X_j \to \hat{\mu}_j, \hat{\Sigma}_j$$

- $x_i \to g_j(x_i) = (x_i-\hat{\mu}_j)^T \hat{\Sigma}_j^{-1}(x_i-\hat{\mu}_j)$

Γραμμικός Ευκλείδειος ταξινομητής

- $g_j(x) = (x-\mu_j)^T(x-\mu_j)$
- Υπόθεση : $N(\mu_j, I)$
- Εκτίμ. μ_j, με βάση το X_j

$$(ML): X_j \to \hat{\mu}_j$$

- $x_i \to g_j(x_i) = (x_i-\hat{\mu}_j)^T(x_i-\hat{\mu}_j)$

Γραμμικός Mahalanobis ταξινομητής

- $g_j(x) = (x-\mu_j)^T \Sigma^{-1}(x-\mu_j)$
- Υπόθεση : $N(\mu_j, \Sigma)$
- Εκτίμ. μ_j, Σ, με βάση το X_j

$$(ML): X_j \to \hat{\mu}_j, \hat{\Sigma}$$

- $x_i \to g_j(x_i) = (x_i-\hat{\mu}_j)^T \hat{\Sigma}^{-1}(x_i-\hat{\mu}_j)$

"ΧΑΡΤΟΓΡΑΦΗΣΗ" ΤΟΥ ΧΩΡΟΥ ΤΩΝ ΤΑΞΙΝΟΜΗΤΩΝ

Υπενθ.: X είναι το σύνολο των δεδομένων σημείων όλων των κλάσεων

X_j είναι το υποσύνολο του X που περιέχει τα διανύσματα της κλάσης ω_j,

$X=X_1 \cup \ldots \cup X_M$

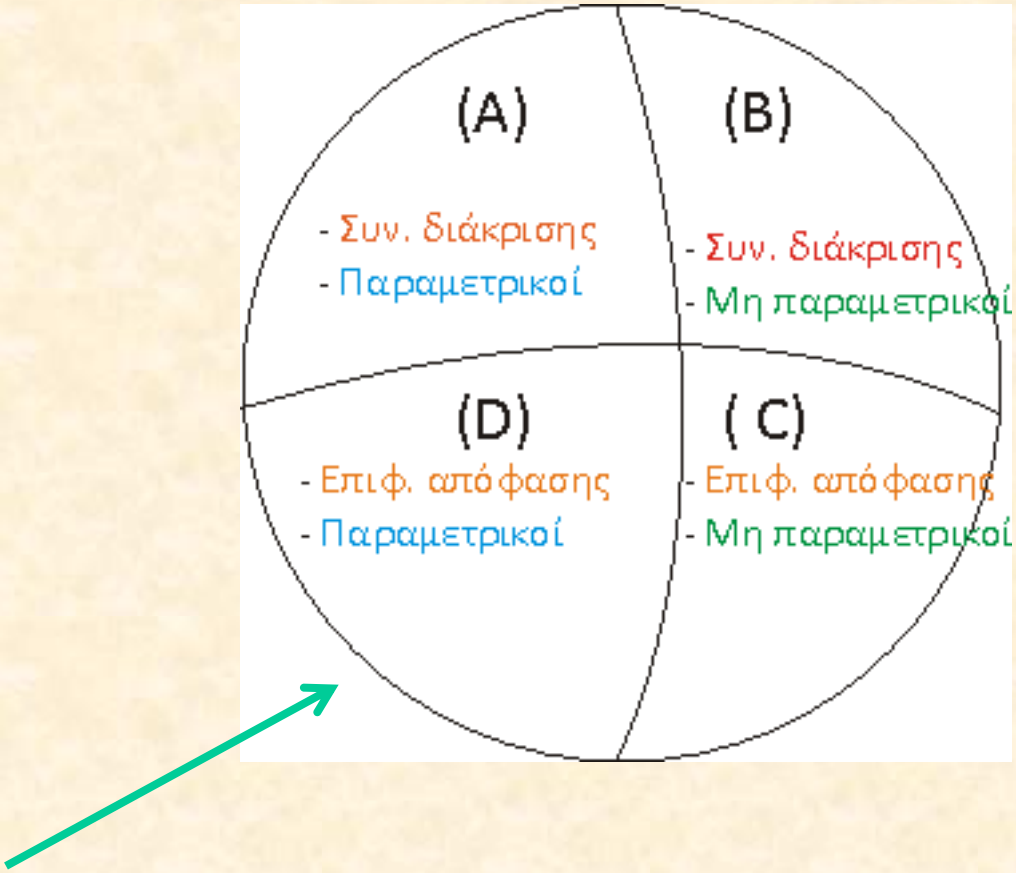
Μη παραμετρικοί ταξινομητές με βάση τις συναρτήσεις διάκρισης

Ταξινομητής Bayes

- $g_j(x) = f(P(\omega_j)p(x|\omega_j))$
- $x_i \rightarrow p(x_i|\omega_j) \approx \hat{p}(x_i|\omega_j; X_j)$
  $(παράθυρα\ Parzen,$
  $εκτίμ.\ πυκν.\ βάσει\ των\ k-πλησ.\ γειτ.$

Ταξινομητής k-πλησιέστερων γειτόνων

- $x_i \rightarrow g_j(x_i) = k_i^j$


(A)
- Συν. διάκρισης
- Παραμετρικοί
(B)
- Συν. διάκρισης
- Μη παραμετρικοί
(D)
- Επιφ. απόφασης
- Παραμετρικοί
(C)
- Επιφ. απόφασης
- Μη παραμετρικοί

ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Μερικά προκαταρτικά:

- Στην πράξη έχουμε στη διάθεσή μας ένα σύνολο δεδομένων (σύν. εκπαίδευσης)

$$X = \{(x_i, d_i), x_i \in R^l, d_i \in \{1,2,...,M\}, i = 1,...N\}$$

όπου

$\boldsymbol{x}_i$ είναι η *l*-διάστατη αναπαράσταση του *i*-στού από τις *N* οντότητες (διάνυσμα εκπαίδευσης)

d_i είναι η ετικέτα της κλάσης όπου ανήκει το $\boldsymbol{x}_i$ (1 για την ω_1, 2 για την ω_2,…).

- Εστιάζουμε κυρίως στην περίπτωση των δύο κλάσεων.

- Δεν υιοθετούμε καμία υπόθεση σχετικά με τις pdfs που μοντελοποιούν τις διάφορες κλάσεις.

- Αναζητούμε την επιφάνεια (γραμμική ή μη γραμμική) που επιτυγχάνει τον "βέλτιστο" διαχωρισμό των (διανυσμάτων δεδομένων των) κλάσεων.

ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Επιφάνεια απόφασης (*C*): Περιγράφεται από εξίσωση της μορφής $h(\boldsymbol{x})=0$, ή $h(\boldsymbol{x};\boldsymbol{w})=0$, όπου $\boldsymbol{w}$ είναι το διάνυσμα των παραμέτρων που ορίζουν την επιφάνεια (*C*).

Παραδείγματα:

1. Αν η (*C*) είναι **καμπύλη 1ου βαθμού** (υπερεπίπεδο – γραμμικός διαχωρισμός), τότε

$$h(x,w) = w_1\, x_1 + \ldots + w_l\, x_l + w_0 = \sum_{k=1}^{l} w_k x_k + w_0 = w^T x + w_0 = 0$$

όπου $\boldsymbol{w}=[w_1,\ldots,w_l]^T$, $\boldsymbol{x}=[x_1,\ldots,x_l]^T$. Οι παράμετροι είναι το διάνυσμα $\boldsymbol{w}$ και το w_0.

2. Αν η (*C*) είναι **καμπύλη 2ου βαθμού** (π.χ. Υπερέλλειψη – μη γραμμικός διαχωρισμός), τότε

$$h(x,w) = \sum_{k=1}^{l}\sum_{q=k}^{l} w_{kq} x_k x_q + \sum_{k=1}^{l} w_k x_k + w_0 = 0$$

όπου $\boldsymbol{w}=[w_0, w_1,\ldots,w_l, w_{11},\ldots,w_{1l},w_{22},\ldots,w_{2l},\ldots,w_{ll}]^T$

Σημείωση: Σε αρκετούς μη γραμμικούς ταξινομητές (π.χ. στα νευρωνικά δίκτυα) η μορφή της (*C*) δεν μπορεί να εκφραστεί άμεσα (explicitly).

ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Υπόθεση: Αν δεν ορίζεται διαφορετικά, θεωρούμε την περίπτωση των δύο κλάσεων, δηλ., ω_1 (+1) και ω_2 (-1).

Ορισμός προβλήματος: Δοθέντος ενός συνόλου δεδομένων *X* προσδιόρισε μία επιφάνεια που επιτυγχάνει το "βέλτιστο" δυνατό διαχωρισμό των (διανυσμάτων των) δύο κλάσεων.

Στρατηγική αντιμετώπισης του προβλήματος:

1. **Υιοθέτησε** μια συγκεκριμένη ("άμεση" ή "έμμεση") παραμετρική μορφή για την επιφάνεια $h(\boldsymbol{x};\boldsymbol{w})=0$.
2. **Όρισε** κατάλληλη συνάρτηση (συνάρτηση κόστους - cost function) του $\boldsymbol{w}$, $J(\boldsymbol{w})$, η οποία περιλαμβάνει επίσης τα διανύσματα του *X*, έτσι ώστε **τα βέλτιστά της (ελάχιστα ή μέγιστα) να αντιστοιχούν στις καλύτερες δυνατές επιφάνειες για το υπό μελέτη πρόβλημα**.
3. **Βελτιστοποίησε** την $J(\boldsymbol{w})$ ως προς το $\boldsymbol{w}$. **Η θέση του βέλτιστου αυτής ορίζει την επιφάνεια απόφασης**.

Σημαντική παρατήρηση: Το νόημα της φράσης "βέλτιστη δυνατή καμπύλη" διαφέρει για διαφορετικές επιλογές της $J(\boldsymbol{w})$.

ΒΑΣΙΚΑ ΣΤΟΙΧΕΙΑ ΘΕΩΡΙΑΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ

Έστω $J(\boldsymbol{w})$ συνεχής συνάρτηση του $\boldsymbol{w}$.

Πρόβλημα (P1): Προσδιόρισε τη θέση $\boldsymbol{w}^*$ όπου η συνάρτηση $J(\boldsymbol{w})$ λαμβάνει την ελάχιστη τιμή της.

Μια απλή μέθοδος για την επίλυση του **(P1)** είναι αυτή της οξύτερης καθόδου (gradient descent - GD).

-**Αρχικοποίησε** $\boldsymbol{w}=\boldsymbol{w}(0)$

-t=0

-**Επανέλαβε**

- $w(t+1) = w(t) - \mu \frac{\partial J(w)}{\partial w}|_{w=w(t)}$
- t=t+1

-**Έως ότου** επιτευχθεί σύγκλιση

ΒΑΣΙΚΑ ΣΤΟΙΧΕΙΑ ΘΕΩΡΙΑΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ

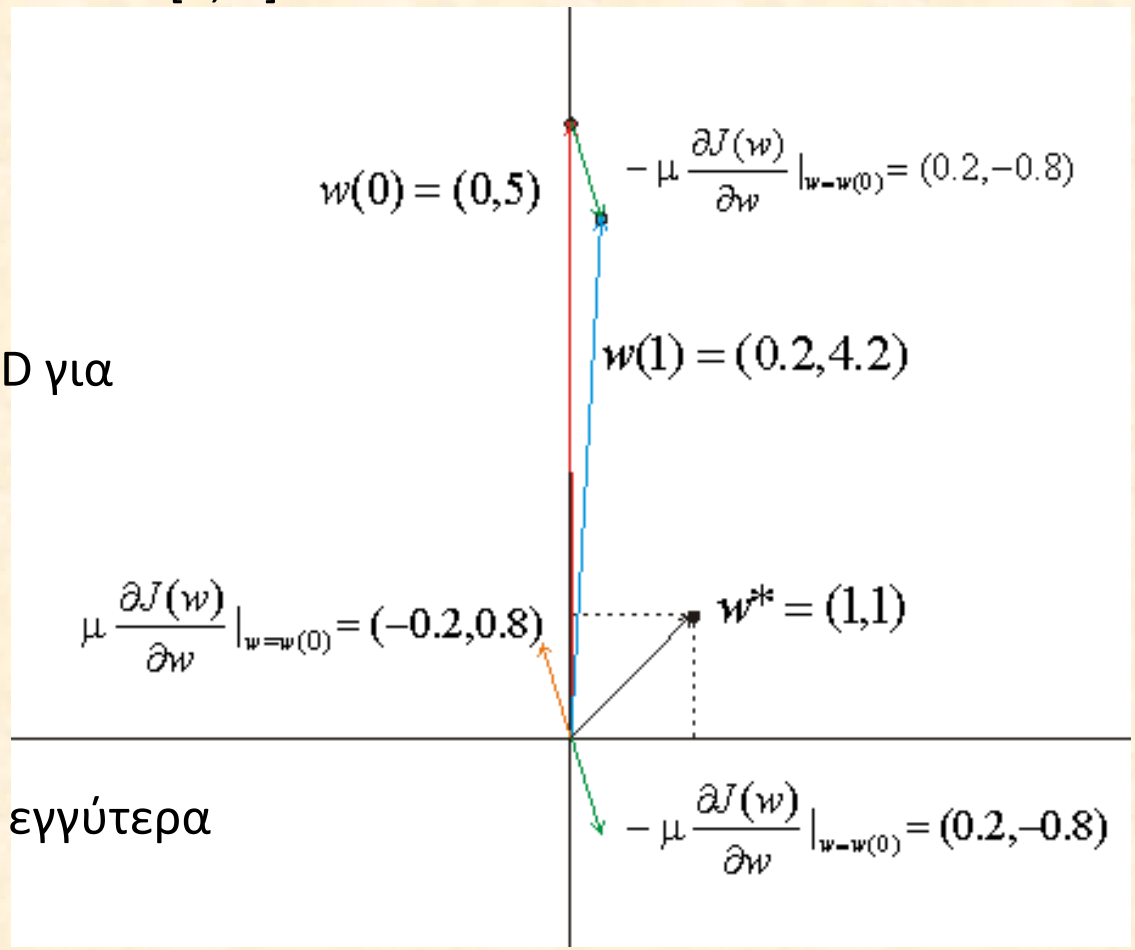
-**Ένα παράδειγμα:** Έστω $\boldsymbol{w}=[w_1, w_2]^T$ και $J(\boldsymbol{w})=(w_1-1)^2+(w_1-1)^2$. Είναι σαφές ότι η $J(\boldsymbol{w})$ λαμβάνει την ελάχιστη τιμή της στο $\boldsymbol{w}^*=[1, 1]^T$.

-Είναι $\dfrac{\partial J(w)}{\partial w}=\begin{bmatrix}2w_1-2\\ 2w_2-2\end{bmatrix}$

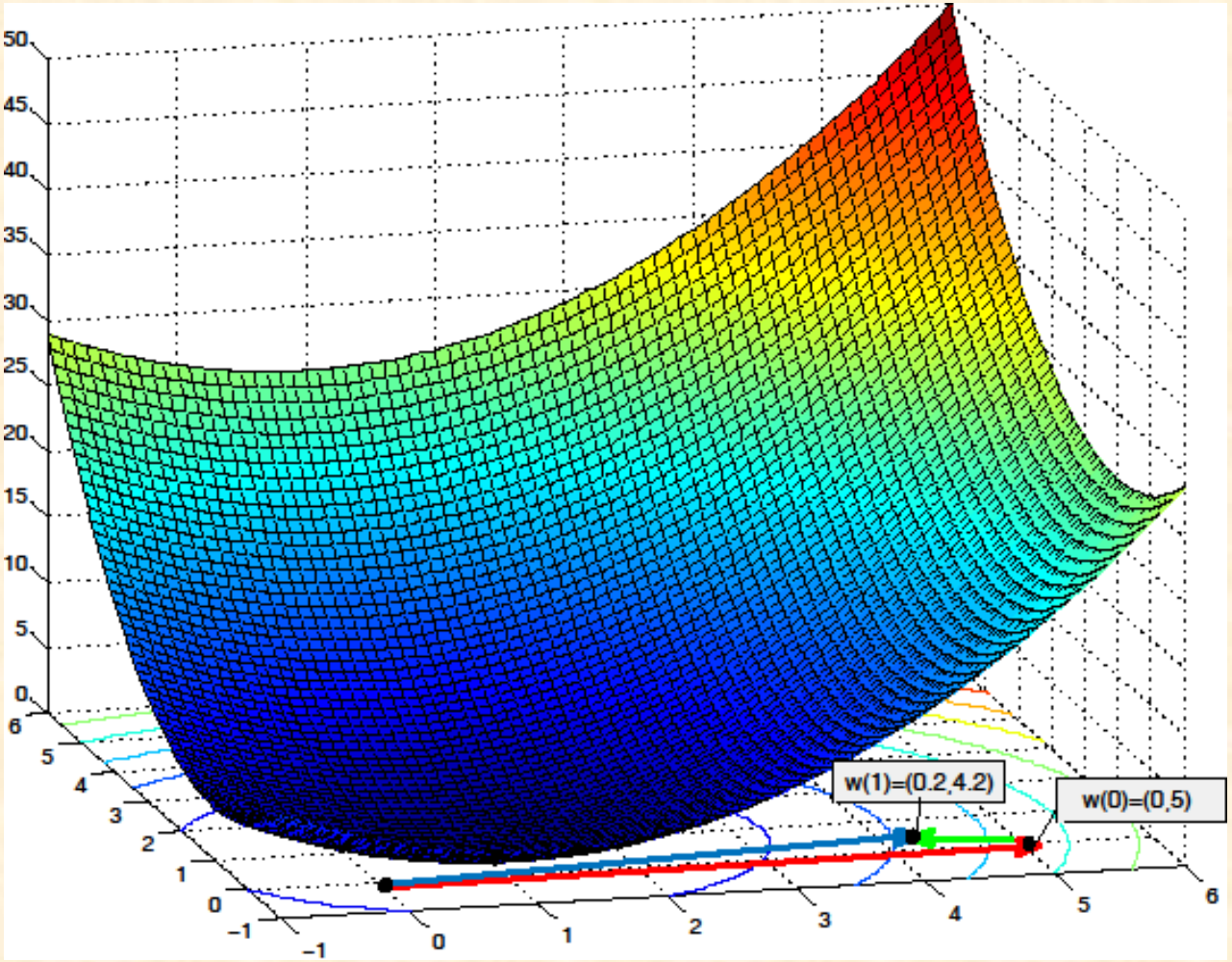
-Εφαρμόζοντας τον αλγόριθμο GD για $\boldsymbol{w}(0)=[0, 5]^T$, και $\mu=0.1$, έχουμε

$$w(1)=\begin{bmatrix}0\\5\end{bmatrix}-0.1\begin{bmatrix}-2\\8\end{bmatrix}=\begin{bmatrix}0.2\\4.2\end{bmatrix}$$

-Βλέπουμε ότι το $\boldsymbol{w}(1)$ βρίσκεται εγγύτερα στο $\boldsymbol{w}^*$.



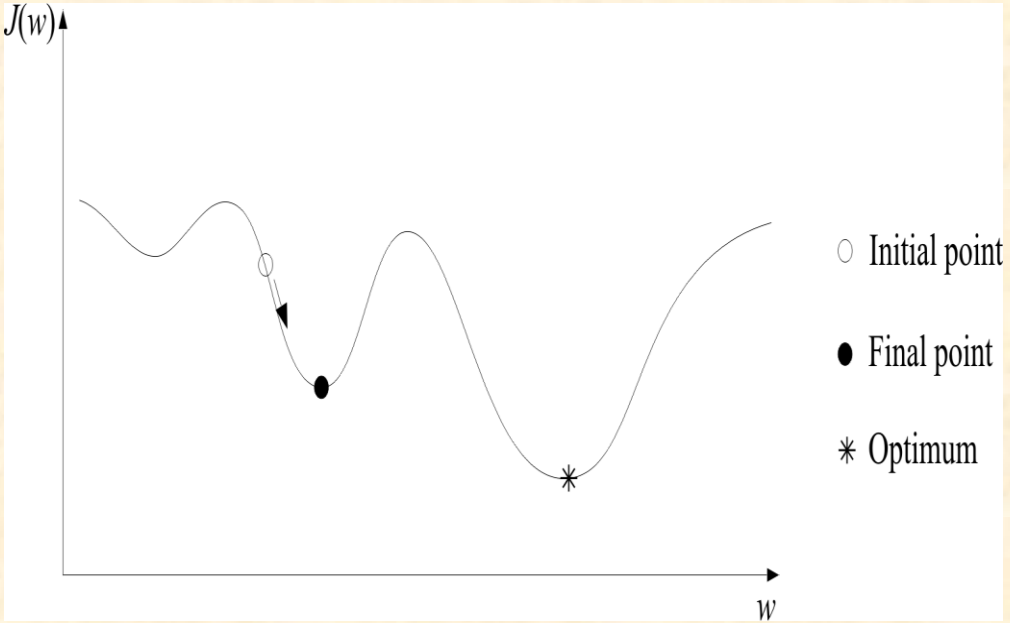
ΒΑΣΙΚΑ ΣΤΟΙΧΕΙΑ ΘΕΩΡΙΑΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ



ΒΑΣΙΚΑ ΣΤΟΙΧΕΙΑ ΘΕΩΡΙΑΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ

Σχόλια για τον αλγόριθμο GD:

- Η τιμή του μ πρέπει να είναι όχι πολύ μεγάλη, ώστε να αποφεύγονται ταλαντώσεις γύρω από το ελάχιστο της συνάρτησης και όχι πολύ μικρή, ώστε να αποφεύγονται άσκοπες καθυστερήσεις στη σύγκλιση.
- Αν η $J(\boldsymbol{w})$ έχει περισσότερα από ένα ελάχιστα, ο αλγόριθμος GD θα συγκλίνει (γενικά) σε εκείνο που βρίσκεται εγγύτερα στο $\boldsymbol{w}(0)$.
- Αν ο αλγόριθμος παγιδευτεί σε ένα τοπικό ελάχιστο *που αντιστοιχεί σε μια "κακή" λύση*, ο μόνος τρόπος για να το αποφύγουμε είναι να επανεκκινήσουμε τον αλγόριθμο από μια διαφορετική αρχική θέση.
- Μπορεί να αποδειχθεί ότι, κάτω από κάποιες συνθήκες, ο αλγόριθμος συγκλίνει ασυμπτωτικά σε ένα τοπικό ελάχιστο της $J(\boldsymbol{w})$.



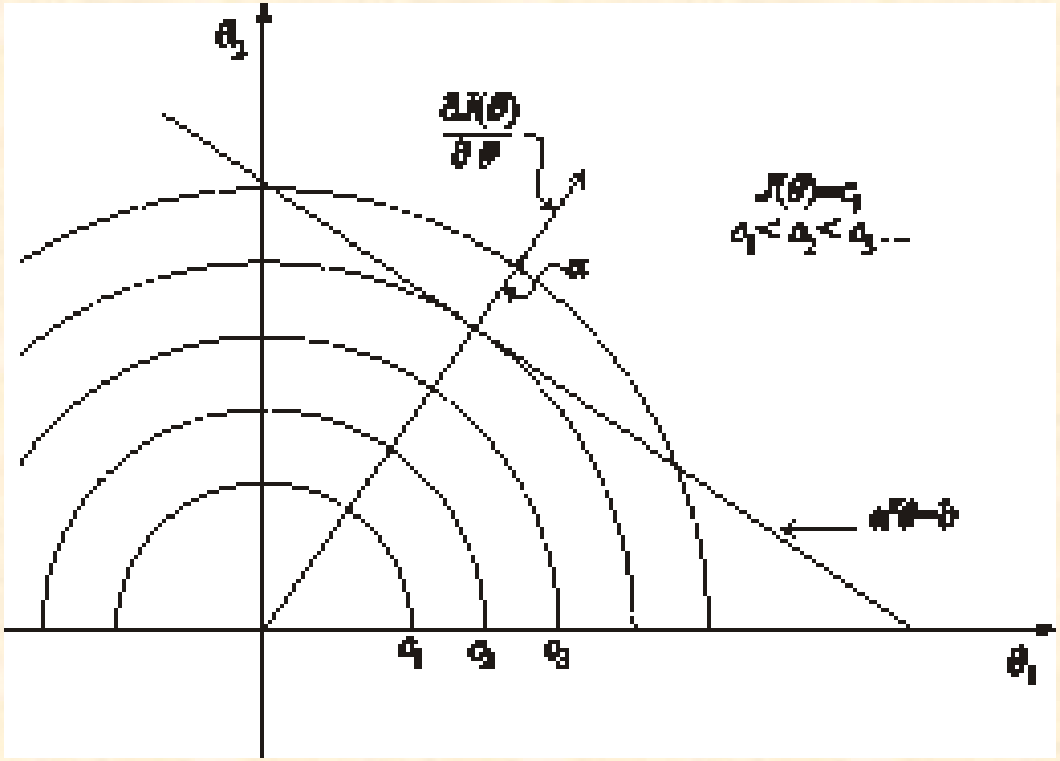
ΒΑΣΙΚΑ ΣΤΟΙΧΕΙΑ ΘΕΩΡΙΑΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ

Έστω $J(\boldsymbol{w})$ μια συνεχής συνάρτηση του $\boldsymbol{w}$.

Πρόβλημα (P2): Προσδιόρισε τη θέση $\boldsymbol{w}^*$ όπου η συνάρτηση $J(\boldsymbol{w})$ λαμβάνει την ελάχιστη τιμή της, υπό την προϋπόθεση ότι το $\boldsymbol{w}$ ικανοποιεί μερικούς περιορισμούς ισότητας (equality constraints).

Για γραμμικούς περιορισμούς ισότητας, το πρόβλημα διατυπώνεται ως εξής

- Ελαχιστοποίησε την $J(\boldsymbol{w})$
- Υπό τους περιορισμούς $A\boldsymbol{w}=\boldsymbol{b}$, όπου Α ένας mxl πίνακας και $\boldsymbol{b}$ ένα m-διάστατο διάνυσμα.



Λύση: Πολ/στές Lagrange

Ελαχιστοποίησε την

- $L(\boldsymbol{w})=J(\boldsymbol{w})+\boldsymbol{\lambda}^T(A\boldsymbol{w}-\boldsymbol{b})$
- $\boldsymbol{\lambda}$ είναι ένα m-διάστατο διάνυσμα που εκτιμάται μέσω των περιορισμών $A\boldsymbol{w}=\boldsymbol{b}$

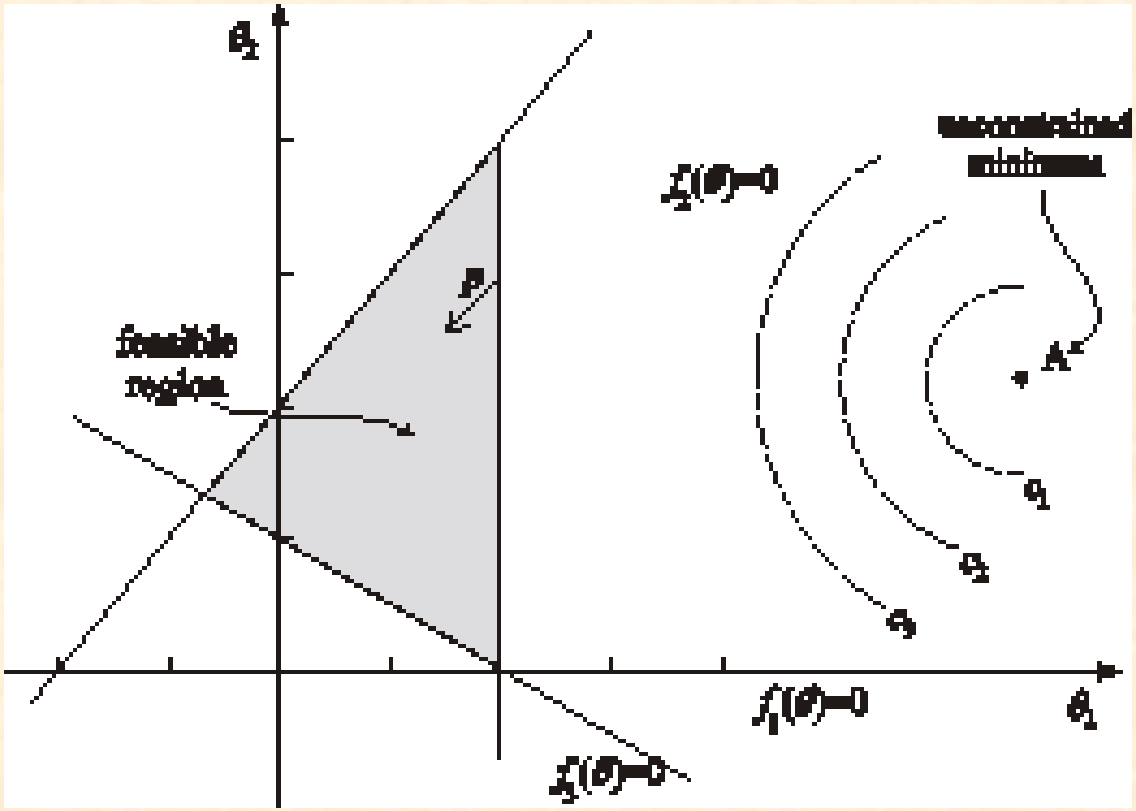
ΒΑΣΙΚΑ ΣΤΟΙΧΕΙΑ ΘΕΩΡΙΑΣ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ

Έστω $J(\boldsymbol{w})$ μια συνεχής συνάρτηση του $\boldsymbol{w}$.

Πρόβλημα (P3): Προσδιόρισε τη θέση $\boldsymbol{w}^*$ όπου η συνάρτηση $J(\boldsymbol{w})$ λαμβάνει την ελάχιστη τιμή της, υπό τις την προϋπόθεση ότι το $\boldsymbol{w}$ ικανοποιεί μερικούς περιορισμούς ανισότητας.

Για γραμμικούς περιορισμούς ανισότητας, το πρόβλημα διατυπώνεται ως εξής

- Ελαχιστοποίησε την $J(\boldsymbol{w})$
- Υπό τους περιορισμούς $A\boldsymbol{w} \geq \boldsymbol{b}$, όπου Α ένας mxl πίνακας και $\boldsymbol{b}$ ένα m-διάστατο διάνυσμα.



ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

-Στην περίπτωση αυτή η επιφάνεια είναι ένα υπερεπίπεδο (*H*) της μορφής

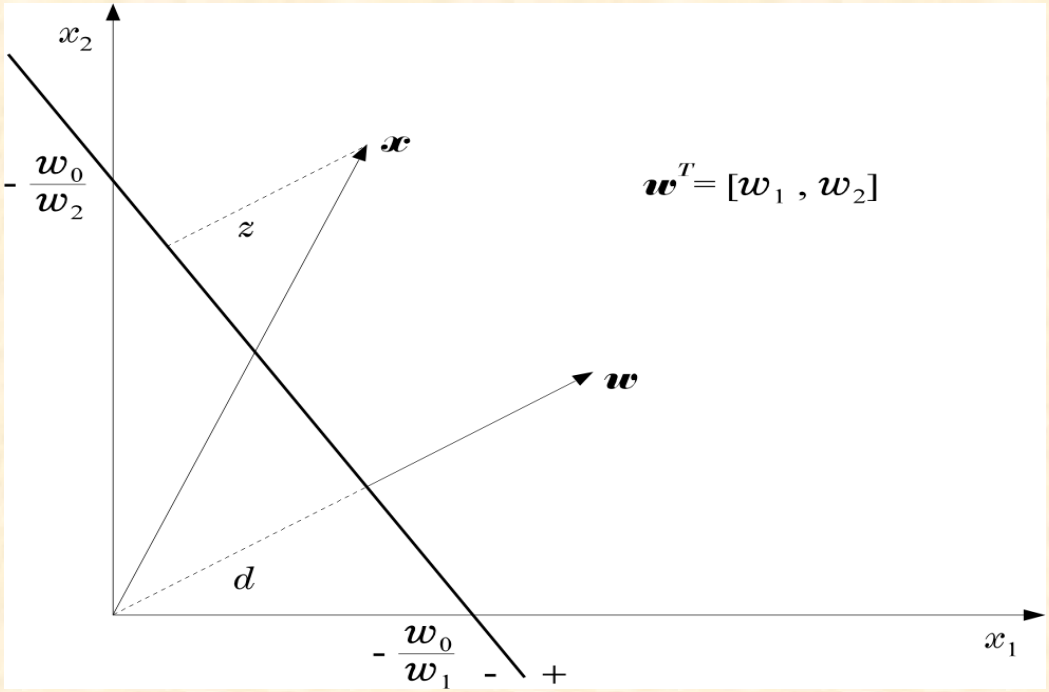
$$(H): h(x,w) = w_1\, x_1 + \ldots + w_l\, x_l + w_0 = \sum_{k=1}^{l} w_k x_k + w_0 = w^T x + w_0 = 0$$

όπου $\boldsymbol{w}=[w_1,\ldots,w_l]^T$, $\boldsymbol{x}=[x_1,\ldots,x_l]^T$

- Το (*H*) **ορίζεται πλήρως** από τα ***w*** και w_0.

- Μερικά στοιχεία από τη Γεωμετρία:
  - Το διάνυσμα ***w*** είναι κάθετο στο (*H*).
  - Το ***w*** ορίζει τον προσανατολισμό (orientation) του (*H*).
  - Όλα τα υπερεπίπεδα που είναι παράλληλα με το (*H*) έχουν τον ίδιο προσανατολισμό με αυτό.
  - Το ***w*** είναι κάθετο σε όλα τα υπερεπίπεδα που είναι παράλληλα με το (*H*).
  - Το w_0 είναι η παράμετρος που ταυτοποιεί μοναδικά ένα συγκεκριμένο υπερεπίπεδο, ανάμεσα από υπερεπίπεδα ίδιου προσανατολισμού.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ



$$d = \frac{|w_0|}{\sqrt{w_1^2 + w_2^2}}, \quad z = \frac{|g(\underline{x})|}{\sqrt{w_1^2 + w_2^2}}$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

ΜΙΑ ΣΥΜΒΑΣΗ: Προκειμένου να καταστήσουμε πιο συμπαγή το συμβολισμό

$$(H): w^T x + w_0 = 0$$

ορίζουμε

$\boldsymbol{w}^* = [\boldsymbol{w}^T, w_0]^T$ and $\boldsymbol{x}^* = [\boldsymbol{x}^T\ 1]^T$.

’Ετσι έχουμε

$$(H): w^{*T}\ x^* = 0$$

ΣΗΜΕΙΩΣΗ: Στη συνέχεια, εκτός αν ορίζεται διαφορετικά, θα γράφουμε $\boldsymbol{w}$ και $\boldsymbol{x}$, εννοώντας τα $\boldsymbol{w}^*$ και $\boldsymbol{x}^*$, αντίστοιχα.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Ο αλγόριθμος perceptron

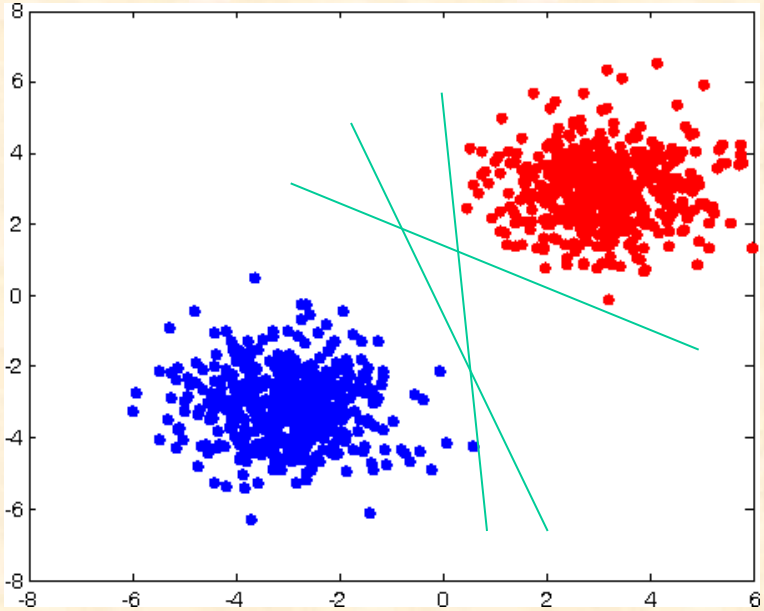
Υπόθεση: (Τα διανύσματα των) δύο κλάσεων ω_1 και ω_2 είναι γραμμικώς διαχωρίσιμα.

Έτσι, υπάρχει ένα υπερεπίπεδο (H^*): $\boldsymbol{w}^{*T}\boldsymbol{x}=0$, ώστε

$$w^{*T}\ x > 0,\ \ \forall x \in \omega_1$$
$$w^{*T}\ x < 0,\ \ \forall x \in \omega_2$$

Το κριτήριο στο perceptron: Προσδιόρισε ένα υπερεπίπεδο το οποίο διαχωρίζει τέλεια τα διανύσματα από τις δύο κλάσεις.

ΣΗΜΕΙΩΣΗ: Αν οι κλάσεις είναι γραμμικώς διαχωρίσιμες, υπάρχουν άπειρα υπερεπίπεδα που ικανοποιούν το κριτήριο perceptron.



ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Ο αλγόριθμος perceptron

Η συνάρτηση κόστους:

$$J(\underline{w}) = \sum_{\underline{x} \in Y(w)} (\delta_x w^T x)$$

όπου

- $Y(\boldsymbol{w})$ το σύνολο των λανθασμένα ταξινομημένων διανυσμάτων από το $\boldsymbol{w}$ και

- $\delta_x = -1 \text{ if } x \in \omega_1$
  $\delta_x = +1 \text{ if } x \in \omega_2$

Σημειώσεις:

- Είναι $J(\boldsymbol{w}) \geq 0$.

  Πράγματι, αν $\boldsymbol{x} \in Y(\boldsymbol{w})$ και $\boldsymbol{x} \in \omega_1$ τότε $\boldsymbol{w}^T\boldsymbol{x} < 0$ και $\delta_x = -1$. Επομένως $\delta_x \boldsymbol{w}^T\boldsymbol{x} > 0$.

  Επίσης, αν $\boldsymbol{x} \in Y(\boldsymbol{w})$ και $\boldsymbol{x} \in \omega_2$ τότε $\boldsymbol{w}^T\boldsymbol{x} > 0$ και $\delta_x = +1$. Επομένως $\delta_x \boldsymbol{w}^T\boldsymbol{x} > 0$.

- Είναι $J(\boldsymbol{w}) = 0$ μόνον όταν $Y = \varnothing$, πράγμα που σημαίνει ότι προσδιορίστηκε **μία λύση που ικανοποιεί το κριτήριο perceptron**.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Ο αλγόριθμος perceptron

Η συνάρτηση κόστους:

H J(***w***) είναι τμηματικά γραμμική (piecewise linear) (γιατί;)

Η ελαχιστοποίηση της *J*(***w***) επιτυγχάνεται μέσω διαδικασίας που ακολουθεί τη φιλοσοφία της GD.

Όπου είναι έγκυρο, ορίζουμε $$\frac{\partial J(w)}{\partial w} = \frac{\partial}{\partial w}\left(\sum_{x \in Y(w)} \delta_x w^T x\right) = \sum_{x \in Y(w)} \delta_x x$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Ο αλγόριθμος perceptron

-**Αρχικοποίησε** $\boldsymbol{w}=\boldsymbol{w}(0)$

-t=0

-**Επανέλαβε**

-Προσδιόρισε τα λανθασμένα ταξινομημένα διανύσματα του X από το $\boldsymbol{w}(t)$ και συσσώρευσέ τα στο $Y(\boldsymbol{w}_t)$.

-$w(t+1)=w(t)-\rho_t\sum_{x\in Y(w_t)}\delta_x x$

- t=t+1

-**Έως όπου** επιτευχθεί σύγκλιση

Σημειώσεις:

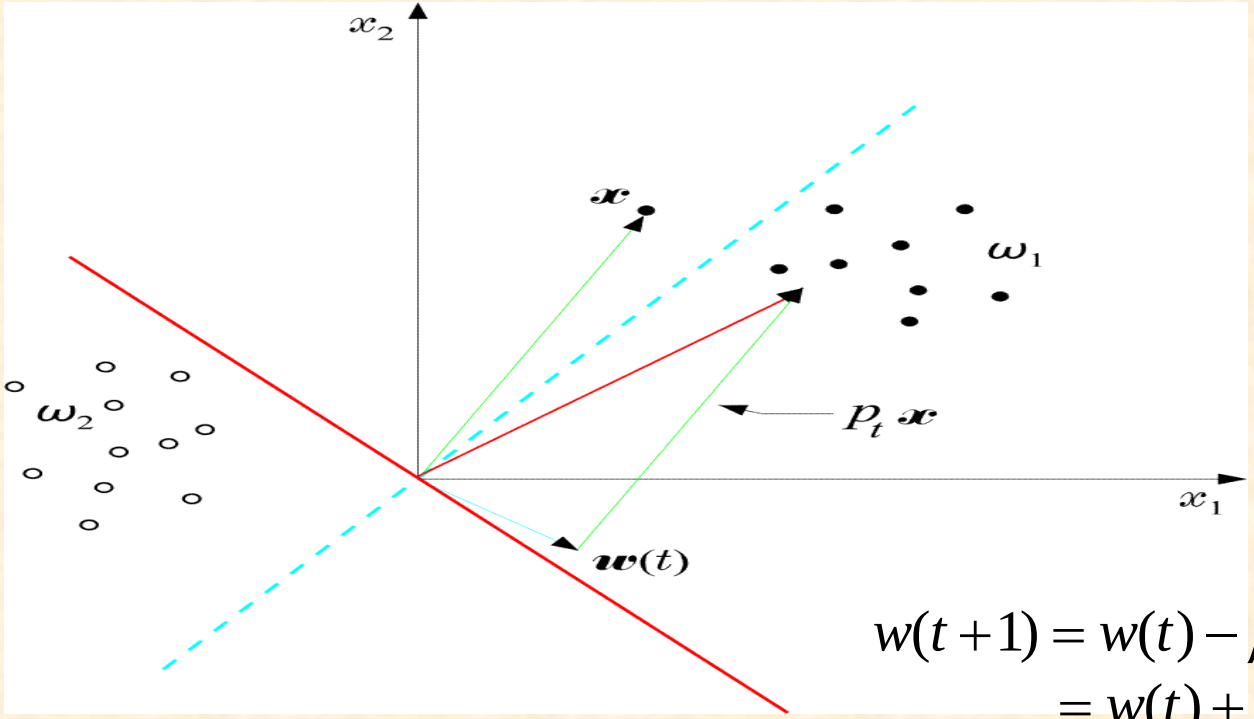
- Πρόκειται για αλγόριθμο επεξεργασίας κατά συρροή (batch algorithm), όπου η ενημέρωση των παραμέτρων σε κάθε επανάληψη πραγματοποιείται μετά την επεξεργασία όλων των διανυσμάτων του X.
- Ο αλγόριθμος perceptron συγκλίνει μόνον όταν οι κλάσεις είναι γραμμικώς διαχωρίσιμες. Διαφορετικά, αποτυγχάνει να συγκλίνει.
- Αν οι κλάσεις είναι γραμμικώς διαχωρίσιμες, ο αλγόριθμος συγκλίνει σε πεπερασμένο αριθμό βημάτων αν t

$$\lim_{t\to\infty}\sum_{k=0}^{t}\rho_k\to\infty,\ \lim_{t\to\infty}\sum_{k=0}^{t}\rho_k{}^2<+\infty\ \pi.\chi.\ \rho_t=c/t$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Ο αλγόριθμος perceptron

Παράδειγμα 1:



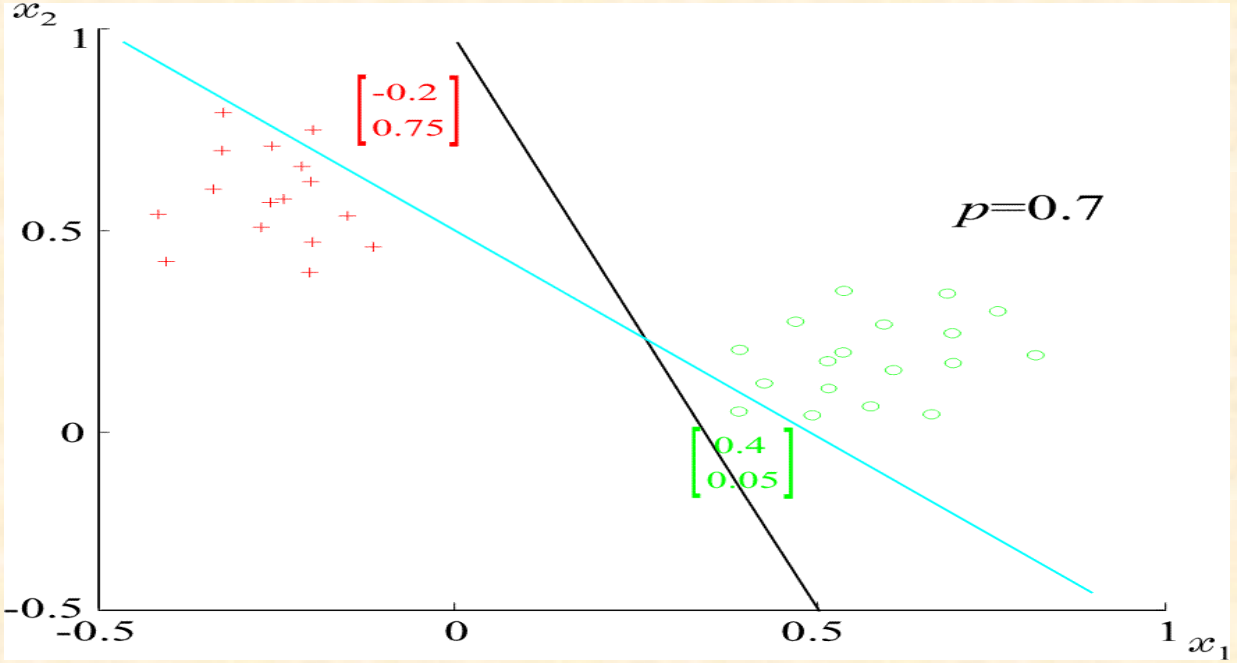
$$w(t+1) = w(t) - \rho_t \delta_x x$$
$$= w(t) + \rho_t x \quad (\delta_x = -1)$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Ο αλγόριθμος perceptron

Παράδειγμα 2: Σε κάποια επανάληψη *t* ο αλγόριθμος perceptron δίνει $w_1 = 1,\ w_2 = 1,\ w_0 = -0.5$ που αντιστοιχεί στην ευθεία $x_1 + x_2 - 0.5 = 0$

Το υπερεπίπεδο της επόμενης επανάληψης είναι



$$\underline{w}(t+1) = \begin{bmatrix} 1 \\ 1 \\ -0.5 \end{bmatrix} - 0.7(-1)\begin{bmatrix} 0.4 \\ 0.05 \\ 1 \end{bmatrix} - 0.7(+1)\begin{bmatrix} -0.2 \\ 0.75 \\ 1 \end{bmatrix} = \begin{bmatrix} 1.42 \\ 0.51 \\ -0.5 \end{bmatrix}$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Παραλλαγές του αλγόριθμου perceptron:

-Ο αλγόριθμος τσέπης (pocket algorithm).

- Κρατά την καλύτερη λύση που βρέθηκε κατά την εκτέλεση του αλγόριθμου perceptron για συγκεκριμένο αριθμό επαναλήψεων.
- Κατάλληλος για προβλήματα μη γραμμικώς διαχωρίσιμων κλάσεων.

-Ο on-line αλγόριθμος perceptron:

- Η ενημέρωση των παραμέτρων λαμβάνει χώρα μετά την μεμονωμένη επεξεργασία κάθε διανύσματος του *X*.

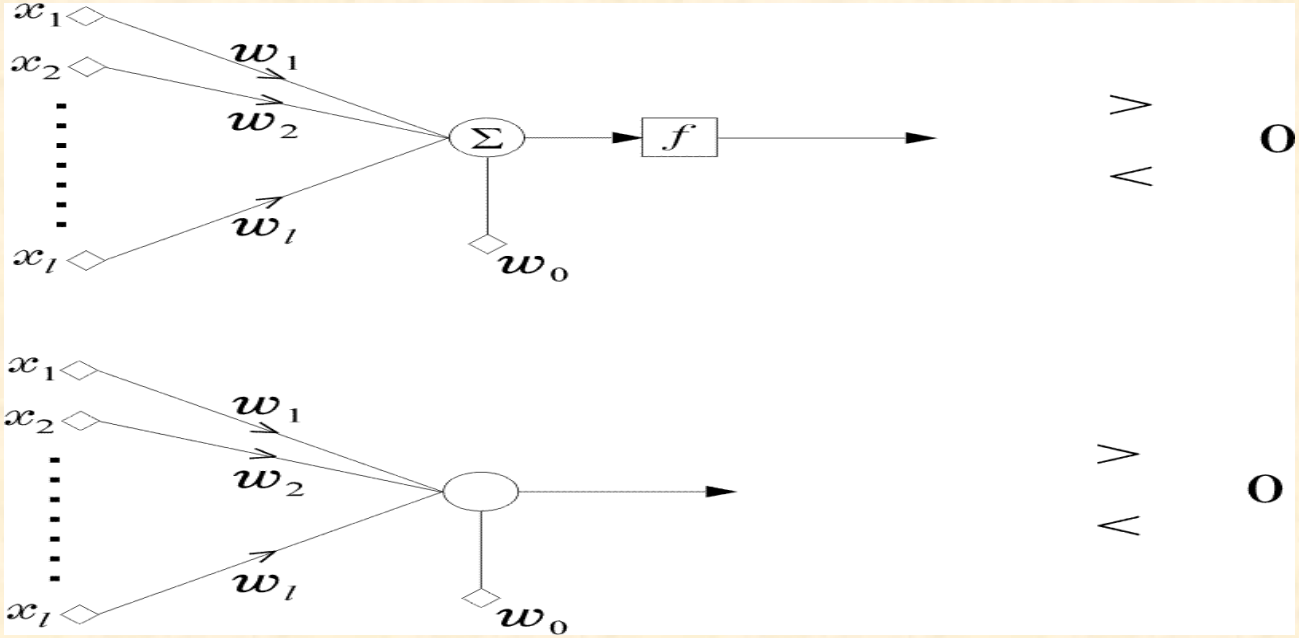
$$w(t+1) = w(t) + \rho x_{(t)}, \quad w^T(t)x_{(t)} \le 0, \; x_{(t)} \in \omega_1$$

$$w(t+1) = w(t) - \rho x_{(t)}, \quad w^T(t)x_{(t)} \ge 0, \; x_{(t)} \in \omega_2$$

$$w(t+1) = w(t) \quad \text{otherwise}$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Η δομή perceptron



$w_i's$ synapsesor synapticweights
w_0 threshold

- Η δομή αυτή καλείται perceptron ή νευρώνας (neuron)
- Είναι μια μηχανή που μπορεί να "μάθει" (δηλ. να προσδιορίσει τις τιμές των εμπλεκόμενων παραμέτρων) από τα διανύσματα εκπαίδευσης μέσω του αλγόριθμου perceptron.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Μέθοδοι ελαχίστων τετραγώνων (Least square methods)

– Αν οι κλάσεις είναι γραμμικώς διαχωρίσιμες, το perceptron θα δώσει σαν έξοδο ± 1

– Αν οι κλάσεις ΔΕΝ είναι γραμμικώς διαχωρίσιμες, θα υπολογίσουμε τα βάρη (παραμέτρους) $w_1, w_2, \ldots, w_0$

έτσι ώστε η διαφορά ανάμεσα

- Στην πραγματική απόκριση του ταξινομητή, $w^T x$, και
- Την αντίστοιχη επιθυμητή απόκριση, δηλ. $\begin{matrix} +1 \text{ if } x \in \omega_1 \\ -1 \text{ if } x \in \omega_2 \end{matrix}$

να είναι όσο το δυνατόν ΜΙΚΡΟΤΕΡΗ για όλα τα διανύσματα του *X*.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Μέθοδοι ελαχίστων τετραγώνων – μικρά ως προς το MSE

– ΜΙΚΡΟΤΕΡΑ, ως προς το κριτήριο του μέσου τετραγωνικού σφάλματος (mean square error - MSE), σημαίνει επιλογή του w ώστε η συνάρτηση κόστους

- $J(w) \equiv E[(d - w^T x)^2]$ να ελαχιστοποιηθεί
  $\hat{w} = \arg\min_w J(w)$

  όπου *d* οι αντίστοιχες επιθυμητές αποκρίσεις

Ελαχιστοποιώντας την *J*(***w***) ως προς το ***w*** έχουμε:

$$\frac{\partial J(w)}{\partial w} = \frac{\partial}{\partial w} E[(d - w^T x)^2] = 0$$
$$= 2E[x(d - x^T w)] \Rightarrow$$

$$E[xx^T]w = E[xd] \Rightarrow \quad \boxed{\hat{w} = R_x^{-1} E[xd]}$$

R_x: Πίν. αυτοσυσχέτισης - **autocorrelation matrix**

$E[xd]$: Διάν. Ετεροσυσχέτ.-**cross-correlation vector**

$$E[xd] = \begin{bmatrix} E[x_1 d] \\ \dots \\ E[x_l d] \end{bmatrix}$$

$$R_x \equiv E[xx^T] = \begin{bmatrix} E[x_1 x_1] & E[x_1 x_2] \dots & E[x_1 x_l] \\ \dots\dots\dots & \dots\dots\dots\dots & \dots\dots\dots \\ E[x_l x_1] & E[x_l x_2] \dots & E[x_l x_l] \end{bmatrix}$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Μέθοδοι ελαχίστων τετραγώνων – μικρά ως προς το MSE: Περ. από 2 κλάσεις

Στην περίπτωση που το υπό μελέτη πρόβλημα περιλαμβάνει M κλάσεις, λύνουμε M προβλήματα σαν το παραπάνω. Συγκεκριμένα:

-Για την j-στή κλάση

- Θέσε $d_i=1$, αν $\boldsymbol{x}_i \in \omega_j$ και $d_i=0$, διαφορετικά.
- Λύσε το πρόβλημα δύο κλάσεων που προκύπτει και έστω $\boldsymbol{w}^j$ το αντίστοιχο διάνυσμα παραμέτρων.

Μετά τον προσδιορισμό των $\boldsymbol{w}^j$'s:

-Για δεδομένο $\boldsymbol{x}_i$:

- **Υπολόγισε** τις ποσότητες $g^j(\boldsymbol{x}_i)=\boldsymbol{w}^{jT}\boldsymbol{x}_i$, $j=1,\ldots,M$.
- **Καταχώρησε** το $\boldsymbol{x}_i$ στην κλάση ω_q για την οποία $g^q(\boldsymbol{x}_i) = \max_{j=1,\ldots,M} g^j(\boldsymbol{x}_i)$

Σημείωση: Το κριτήριο MSE ανήκει σε μια γενικότερη κλάση συναρτήσεων κόστους με την ακόλουθη σημαντική ιδιότητα:

Η τιμή $g^j(\boldsymbol{x}_i)$ είναι μια εκτίμηση, ως προς το κριτήριο MSE, της εκ των υστέρων πιθανότητας, $P(\omega_j|\boldsymbol{x})$, υπό την προϋπόθεση ότι οι επιθυμητές αποκρίσεις που χρησιμοποιούνται κατά την εκπαίδευση είναι $d_i=1$, αν $\boldsymbol{x}_i \in \omega_j$ και $d_i=0$, διαφορετικά.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Μέθοδοι ελαχίστων τετραγώνων – μικρά ως προς το SSE

– ΜΙΚΡΟΤΕΡΟ, ως προς το κριτήριο του αθροίσματος των τετραγώνων των σφαλμάτων, σημαίνει επιλογή του w που ελαχιστοποιεί τη συνάρτηση κόστους

- $J(w) = \sum_{i=1}^{N} (d_i - w^T x_i)^2$

όπου d_i=+1, αν $\boldsymbol{x}_i \in \omega_1$ **και** d_i=-1, αν $\boldsymbol{x}_i \in \omega_1$.

Minimizing $J(\boldsymbol{w})$ with respect to $\boldsymbol{w}$ results in:

$$\frac{\partial J(w)}{\partial w} = \frac{\partial}{\partial w} \sum_{i=1}^{N} (d_i - w^T x_i)^2 = 0 \Rightarrow$$

$$\left(\sum_{i=1}^{N} x_i x_i^T\right) w = \sum_{i=1}^{N} x_i d_i$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ
Μέθοδοι ελαχίστων τετραγώνων – μικρά ως προς το SSE

Ένας εναλλακτικός τρόπος διατύπωσης:

Ορίζουμε τα ακόλουθα

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \dots \\ x_N^T \end{bmatrix}$$ *N*x*l* πίνακας $$\underline{d} = \begin{bmatrix} d_1 \\ \dots \\ d_N \end{bmatrix}$$ Αντίστοιχες επιθυμητές αποκρίσεις

$$X^T = [x_1, x_2, \dots, x_N]$$ *l*x*N* πίνακας

$$X^T X = \sum_{i=1}^{N} x_i x_i^T$$

$$X^T d = \sum_{i=1}^{N} x_i d_i$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ
Μέθοδοι ελαχίστων τετραγώνων – μικρά ως προς το SSE

Ένας εναλλακτικός τρόπος διατύπωσης:

Τότε

$$\left(\sum_{i=1}^{N} x_i^T x_i\right)\hat{w} = \left(\sum_{i=1}^{N} x_i d_i\right) \Leftrightarrow (X^T X)\hat{w} = X^T d \Rightarrow$$

$$\hat{w} = (X^T X)^{-1} X^T d = X^{\neq} d$$

$$X^{\neq} \equiv (X^T X)^{-1} X^T$$

Ψευδοαντίστροφος (pseudoinverse) του *X*

Έστω *N=l*. Τότε ο *X* είναι τετραγωνικός και (γενικά) αντιστρέψιμος. Τότε έχουμε

$$(X^T X)^{-1} X^T = X^{-1} X^{-T} X^T = X^{-1} \Rightarrow \quad X^{\neq} = X^{-1}$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Μέθοδοι ελαχίστων τετραγώνων – μικρά ως προς το SSE

Ένας εναλλακτικός τρόπος διατύπωσης:

Έστω *N>l*. Τότε, γενικά, δεν υπάρχει λύση που να ικανοποιεί ταυτόχρονα όλους τους περιορισμούς

$$Xw = d: \quad \begin{matrix} x_1^T w = d_1 \\ x_2^T w = d_2 \\ .... \\ x_N^T w = d_N \end{matrix} \quad N \text{ equations} > l \text{ unknowns}$$

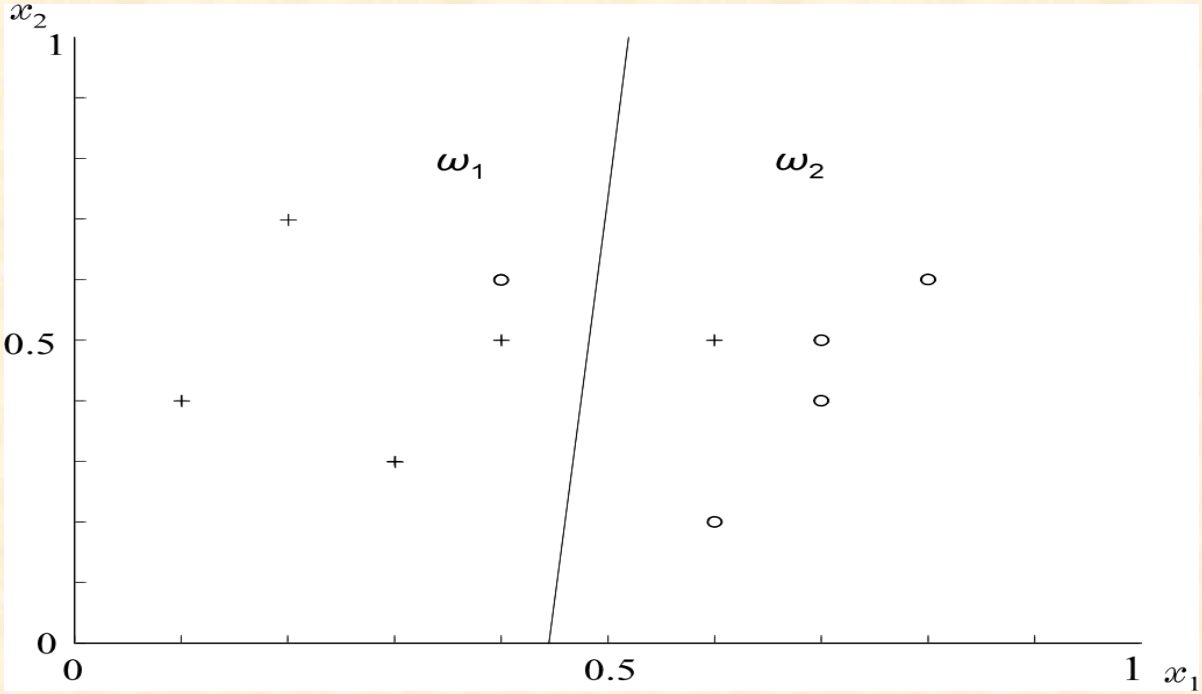
Σ' αυτή την περίπτωση, η "λύση" $w = X^{\neq} d$ ελαχιστοποιεί το άθροισμα των τετραγώνων των σφαλμάτων.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Μέθοδοι ελαχίστων τετραγώνων – μικρά ως προς το SSE

Ένα παράδειγμα:

$$\omega_1: \begin{bmatrix}0.4\\0.5\end{bmatrix}, \begin{bmatrix}0.6\\0.5\end{bmatrix}, \begin{bmatrix}0.1\\0.4\end{bmatrix}, \begin{bmatrix}0.2\\0.7\end{bmatrix}, \begin{bmatrix}0.3\\0.3\end{bmatrix} \quad \omega_2: \begin{bmatrix}0.4\\0.6\end{bmatrix}, \begin{bmatrix}0.6\\0.2\end{bmatrix}, \begin{bmatrix}0.7\\0.4\end{bmatrix}, \begin{bmatrix}0.8\\0.6\end{bmatrix}, \begin{bmatrix}0.7\\0.5\end{bmatrix}$$



$$X = \begin{bmatrix}0.4 & 0.5 & 1\\0.6 & 0.5 & 1\\0.1 & 0.4 & 1\\0.2 & 0.7 & 1\\0.3 & 0.3 & 1\\0.4 & 0.6 & 1\\0.6 & 0.2 & 1\\0.7 & 0.4 & 1\\0.8 & 0.6 & 1\\0.7 & 0.5 & 1\end{bmatrix} \begin{bmatrix}1\\1\\1\\1\\1\\-1\\-1\\-1\\-1\\-1\end{bmatrix} = \underline{d}$$

$$X^T X = \begin{bmatrix}2.8 & 2.24 & 4.8\\2.24 & 2.41 & 4.7\\4.8 & 4.7 & 10\end{bmatrix}, \quad X^T \underline{d} = \begin{bmatrix}-1.6\\0.1\\0.0\end{bmatrix}$$

$$\underline{w} = (X^T X)^{-1} X^T \underline{d} = \begin{bmatrix}-3.13\\0.24\\1.34\end{bmatrix}$$

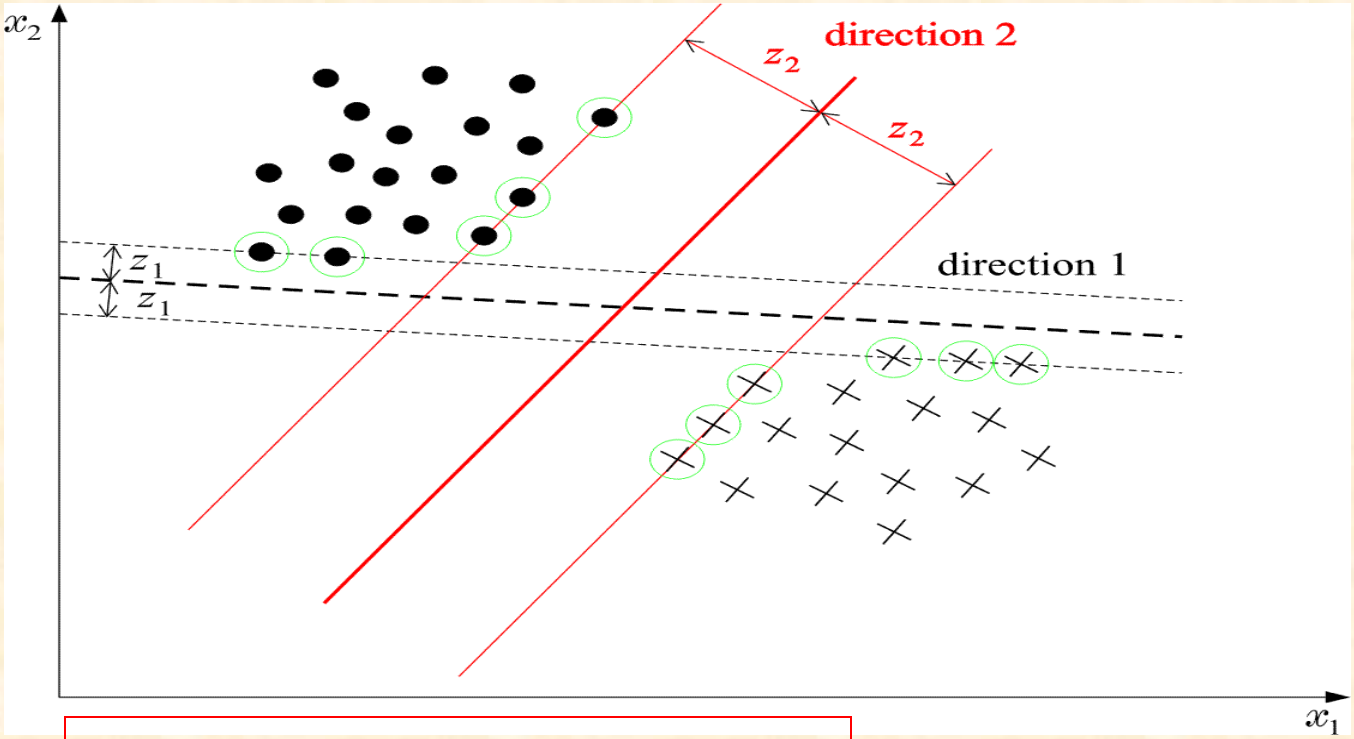
ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – γραμ. διαχ. κλάσεις

Ο στόχος: Δοθέντων δύο γραμμικώς διαχωρίσιμων κλάσεων, προσδιόρισε τον ταξινομητή

$$g(x) = w^T x + w_0 = 0$$

που αφήνει το μέγιστο περιθώριο από τις δύο κλάσεις.



Support vector machines:
Μηχανές διανυσματικής στήριξης

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – γραμ. διαχ. κλάσεις

–Περιθώριο (Margin): Κάθε υπερεπίπεδο χαρακτηρίζεται από

- Τον προσανατολισμό του στο χώρο, δηλ. το διάνυσμα ***w***
- Τη θέση του χώρο, i.e., w_0

- Για ΚΑΘΕ προσανατολισμό, ***w***, επέλεξε το υπερεπίπεδο που αφήνει την ΙΔΙΑ απόσταση από τα εγγύτερα σημεία από κάθε κλάση. Το περιθώριο είναι το διπλάσιο αυτής της απόστασης.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – γραμ. διαχ. κλάσεις

–Η απόσταση ενός σημείου $\hat{x}$ από ένα υπερεπίπεδο είναι

$$z_{\hat{x}} = \frac{g(\hat{x})}{\|w\|}$$

–Διαβάθμισε τα , w, w_0, ώστε για τα εγγύτερα σημεία από κάθε κλάση να είναι:

$$|g(\underline{x})| = 1 \ \{g(\underline{x}) = +1 \text{ for } \omega_1 \text{ and } g(\underline{x}) = -1 \text{ for } \omega_2\}$$

–Τότε το περιθώριο είναι

$$\frac{1}{\|\underline{w}\|} + \frac{1}{\|\underline{w}\|} = \frac{2}{\|w\|}$$

–Επίσης, ισχύουν και οι ακόλουθοι περιορισμοί

$$w^T x + w_0 \geq 1 \ \ \forall x \in \omega_1$$
$$w^T x + w_0 \leq -1 \ \ \forall x \in \omega_2$$

(δηλ. όλα τα στοιχεία της κλάσης +1 (-1) βρίσκονται στη θετική (αρνητική) πλευρά του υπερεπιπέδου)

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – γραμ. διαχ. κλάσεις

- SVM (γραμμικός) ταξινομητής

$$g(\underline{x}) = \underline{w}^T \underline{x} + w_0$$

- (Πρόβλημα SVM) Ελαχιστοποίησε την

$$J(\underline{w}) = \frac{1}{2}\|\underline{w}\|^2$$

- Υπό τους περιορισμούς

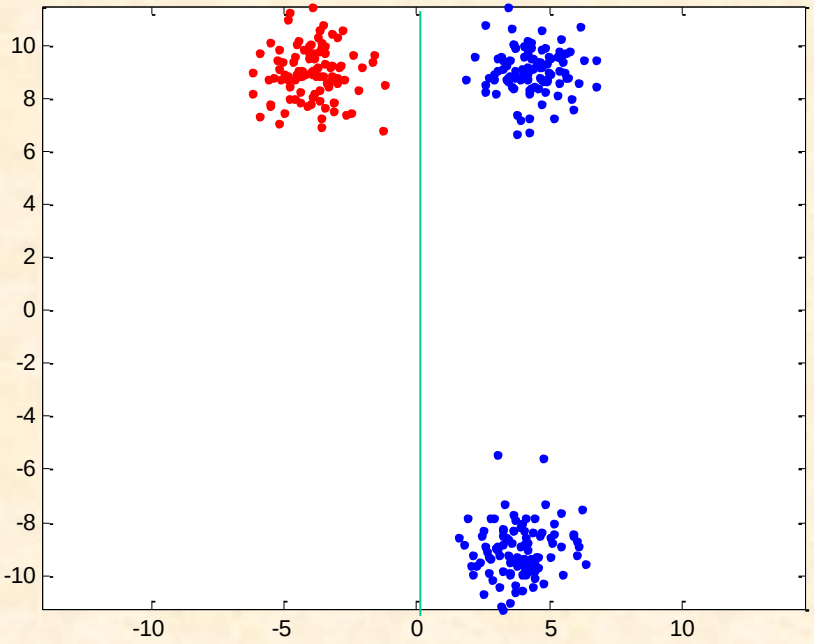
$$d_i(\underline{w}^T \underline{x}_i + w_0) \geq 1, \; i = 1,2,\ldots,N \qquad \begin{array}{l} d_i = 1, \text{ for } \underline{x}_i \in \omega_i, \\ d_i = -1, \text{ for } \underline{x}_i \in \omega_2 \end{array}$$

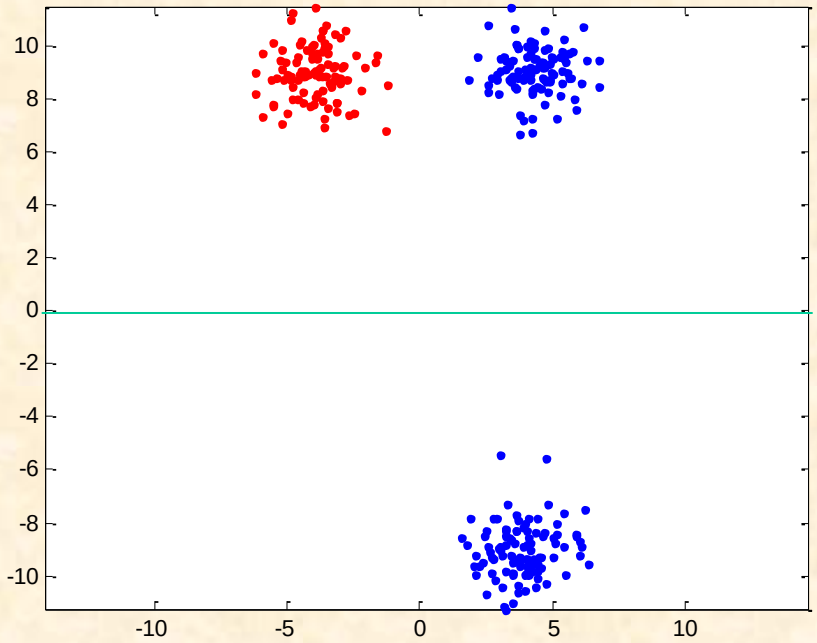
- Τα παραπάνω δικαιολογούνται από το γεγονός ότι ελαχιστοποιώντας το $\|\underline{w}\|^2$

μεγιστοποιείται το περιθώριο $\frac{2}{\|w\|}$.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – γραμ. διαχ. κλάσεις





Οι επιπλέον περιορισμοί ισότητας αποθαρρύνουν λύσεις σαν αυτή του 2^{ου} σεναρίου πιο πάνω.

ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΥΝΑΡΤΗΣΗΣ ΥΠΟ ΓΡΑΜ. ΑΝΙΣΟΤΙΚΟΥΣ ΠΕΡΙΟΡΙΣΜΟΥΣ

Έστω το πρόβλημα

Min $J(\vartheta)$

Υπό τους περιορισμούς $f_i(\vartheta)\geq 0$, $i=1,\ldots m$.

Ορίζουμε τη συνάρτηση Lagrange

$$L(\vartheta,\lambda)=J(\vartheta) - \boldsymbol{\Sigma}_i \lambda_i f_i(\vartheta)$$

Συνθήκες ΚΚΤ (Karush-Kuhn-Tacker): Για τη θέση του **ελαχίστου** ισχύουν τα ακόλουθα

$$\frac{\partial}{\partial \vartheta} \mathrm{L}(\vartheta, \lambda) = 0$$

$$\lambda_i \geq 0, i = 1,2,\ldots, m$$

$$\lambda_i f_i(\vartheta) = 0, i = 1,2,\ldots, m$$

ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ ΣΥΝΑΡΤΗΣΗΣ ΥΠΟ ΓΡΑΜ. ΑΝΙΣΟΤΙΚΟΥΣ ΠΕΡΙΟΡΙΣΜΟΥΣ

Έστω

$$L^*(\vartheta) = max_\lambda\, L(\vartheta,\lambda) = max_\lambda\, (J(\vartheta) - \boldsymbol{\Sigma}_i \lambda_i f_i(\vartheta))$$

Αφού $\lambda_i \geq 0$, $f_i(\vartheta) \geq 0$ έχουμε

$$L^*(\vartheta) = J(\vartheta)$$

Άρα

$$\min_\vartheta J(\vartheta) = \min_\vartheta L^*(\vartheta) = \min_\vartheta \max_{\lambda\geq 0} L(\vartheta,\lambda)$$

Θεώρημα:

- Έστω (α) η $J(\vartheta)$ είναι κυρτή και (β) οι $f_i(\vartheta)$ είναι γραμμικές.
- Έστω ϑ^* μία θέση ελαχίστου για το πρόβλημα ελαχιστοποίησης και λ^* το αντίστοιχο διάνυσμα των πολ/στών Lagrange.

➢ Τότε το (ϑ^*,λ^*) είναι ένα σαγματικό σημείο (saddle point) της συνάρτησης Lagrange, για το οποίο ισχύει

$$L(\vartheta^*,\lambda^*) = \min_\vartheta \max_{\lambda\geq 0} L(\vartheta,\lambda) = \max_{\lambda\geq 0} \min_\vartheta L(\vartheta,\lambda)$$

Συνεπώς, σύμφωνα με το θεώρημα, για τον προσδιορισμό του ελαχίστου **δεν παίζει ρόλο η σειρά με την οποία βελτιστοποιούμε** την $L(\vartheta,\lambda)$ ως προς τα ϑ και λ.

Στο **SVM πρόβλημα** θα **βελτιστοποιήσουμε** την $L(.)$ πρώτα ως προς ϑ και μετά ως προς λ.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – γραμ. διαχ. κλάσεις

–Το πρόβλημα SVM είναι μια διαδικασία (κυρτής) τετραγωνικής βελτιστοποίησης (quadratic optimization task), με γραμμικούς περιορισμούς. Οι συνθήκες Karush-Kuhn-Tucker ορίζουν ότι για τη θέση του ελαχίστου ισχύουν τα ακόλουθα:

- (1) $\frac{\partial}{\partial w} \mathrm{L}(w, w_0, \lambda) = 0$
- (2) $\frac{\partial}{\partial w_0} L(w, w_0, \lambda) = 0$
- (3) $\lambda_i \geq 0, i = 1,2,\ldots, N$
- (4) $\lambda_i \left[d_i (w^T x_i + w_0) - 1 \right] = 0, i = 1,2,\ldots, N$
- Όπου $L(\bullet,\bullet,\bullet)$ είναι η συνάρτηση Lagrange

$$L(w, w_0, \lambda) \equiv \frac{1}{2} w^T w - \sum_{i=1}^{N} \lambda_i [d_i (w^T x_i + w_0)]$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – γραμ. διαχ. κλάσεις

– Η λύση: από τα παραπάνω προκύπτει ότι είναι

$$w = \sum_{i=1}^{N} \lambda_i d_i x_i \qquad \sum_{i=1}^{N} \lambda_i d_i = 0$$

Αντικαθιστώντας τα παραπάνω στην $L(.)$, τα λ_i's εκτιμώνται ως οι λύσεις του δυικού SVM προβλήματος:

(Δυικό SVM)

Μεγιστοποίησε την $\left(\sum_{i=1}^{N} \lambda_i - \frac{1}{2} \sum_{ij} \lambda_i \lambda_j d_i d_j x_i^T x_j\right)$ ως προς το $\boldsymbol{\lambda}=[\lambda_1,\ldots,\lambda_N]^T$

Έτσι ώστε

$$\sum_{i=1}^{N} \lambda_i d_i = 0$$

$$\underline{\lambda} \geq \underline{0}$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – γραμ. διαχ. κλάσεις

Παρατηρήσεις:

- Οι πολ/στές Lagrange λ_i είναι είτε θετικοί είτε μηδέν. Έτσι

$$w = \sum_{i=1}^{N_s} \lambda_i d_i x_i$$

όπου N_s είναι ο αριθμός των διανυσμάτων με θετικά λ_i's.

- Θετικά λ_is έχουν τα διανύσματα που ικανοποιούν τη συνθήκη $w^T x_i + w_0 = \pm 1$ λόγω των περιορισμών $\lambda_i[d_i(\underline{w}^T \underline{x}_i + w_0) - 1] = 0, \quad i = 1,2,\ldots,N$

Τα διανύσματα αυτά ονομάζονται ΔΙΑΝΥΣΜΑΤΑ ΣΤΗΡΙΞΗΣ (SUPPORT VECTORS) και είναι τα εγγύτερα προς το υπερεπίπεδο του ταξινομητή διανύσματα από κάθε κλάση. Αυτά είναι που καθορίζουν το ***w***.

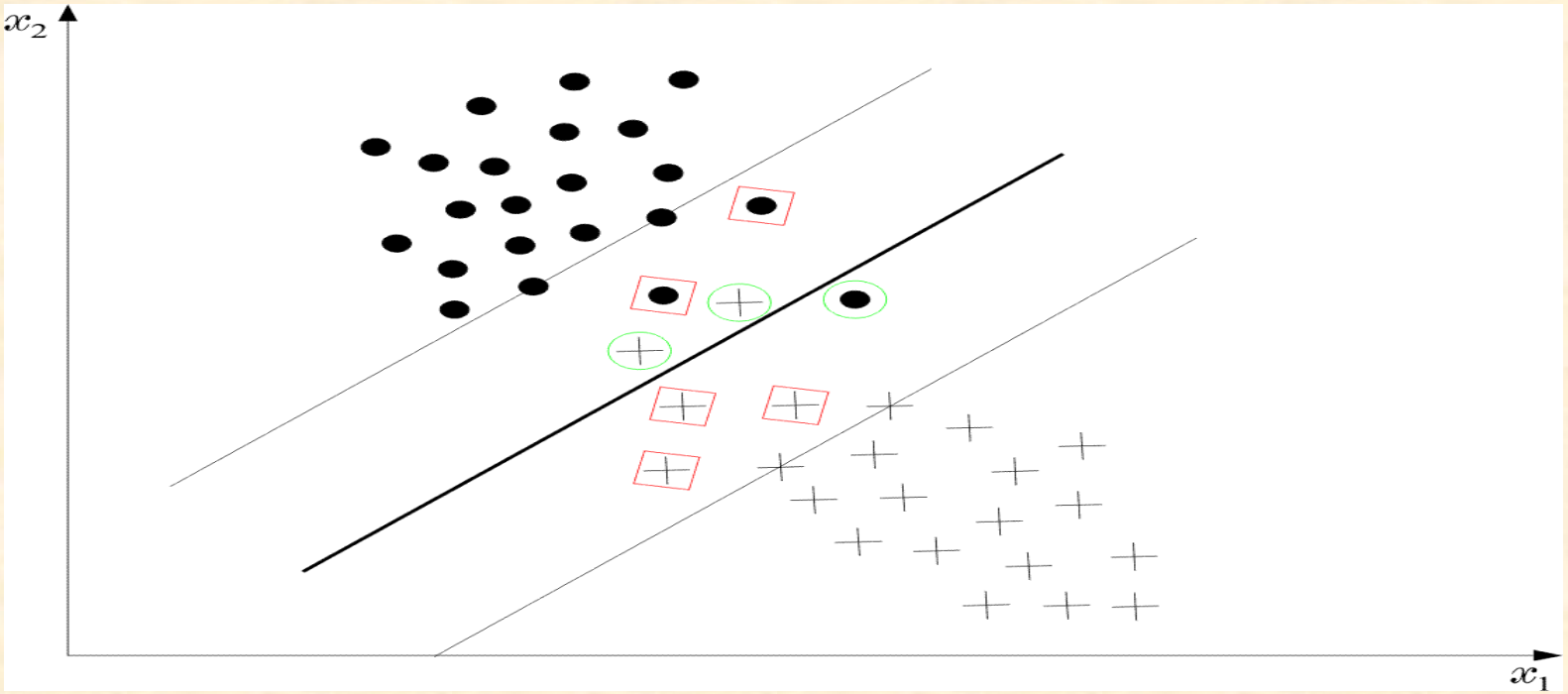
- Μετά τον προσδιορισμό του ***w***, το w_0 προσδιορίζεται από τις συνθήκες (4).

- Το βέλτιστο υπερεπίπεδο ως προς το κριτήριο της μεγιστοποίησης του περιθωρίου είναι ΜΟΝΑΔΙΚΟ.

- Παρότι η λύση είναι μοναδική, οι πολ. Lagrange δεν είναι απαραίτητα μοναδικοί.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – μη γραμ. διαχ. κλάσεις



ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – μη γραμ. διαχ. κλάσεις

- Στην περίπτωση αυτή, δεν υπάρχει υπερεπίπεδο έτσι ώστε

$$w^T x + w_0 (><) 1, \ \forall x$$

- Υπενθυμίζεται ότι το περιθώριο ορίζεται ως το διπλάσιο της απόστασης μεταξύ των ακόλουθων δύο υπερεπιπέδων

$$w^T x + w_0 = 1 \ \text{and} \ w^T x + w_0 = -1$$

–Για τα διανύσματα εκπαίδευσης έχουμε ένα από τα ακόλουθα τρία δυνατά σενάρια

1) Διανύσματα εκτός της ζώνης των δύο επιπέδων που είναι ορθά ταξινομημένα

$$d_i (w^T x + w_0) > 1$$

2) Διανύσματα εντός της ζώνης των δύο επιπέδων που είναι ορθά ταξινομημένα

$$0 \le d_i (w^T x + w_0) < 1$$

3) Διανύσματα τα οποία είναι μη ορθά ταξινομημένα

$$d_i (w^T x + w_0) < 0$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – μη γραμ. διαχ. κλάσεις

– Οι παραπάνω περιπτώσεις μπορούν να παρασταθούν με συμπαγή τρόπο ως εξής

$$d_i(\underline{w}^T \underline{x} + w_0) \geq 1 - \xi_i$$

Όπου για

1ο σενάριο $\rightarrow \xi_i = 0$

2ο σενάριο $\rightarrow 0 < \xi_i \leq 1$

3ο σενάριο $\rightarrow 1 < \xi_i$

Οι μεταβλητές ξ_i είναι γνωστές ως μεταβλητές χαλαρότητας (slack variables)

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – μη γραμ. διαχ. κλάσεις

–Ο στόχος της βελτιστοποίησης είναι τώρα διττός

- **Μεγιστοποίηση του** περιθωρίου
- **Ελαχιστοποίηση** του αριθμού των διανυσμάτων με $\xi_i > 0$,
- Ένας τρόπος να επιτύχουμε τον παραπάνω στόχο είναι μέσω της ακόλουθης συνάρτησης κόστους

$$J(\underline{w}, w_0, \underline{\xi}) = \frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{N} I(\xi_i)$$

όπου *C* είναι μια σταθερά και

$$I(\xi_i) = \begin{cases} 1 & \xi_i > 0 \\ 0 & \xi_i = 0 \end{cases}$$

- Η συνάρτηση *I*(.) ΔΕΝ είναι διαφορίσιμη. Στην πράξη, μπορούμε να χρησιμοποιήσουμε μια προσέγγιση

$$J(\underline{w}, w_0, \underline{\xi}) = \frac{1}{2}\|\underline{w}\|^2 + C\sum_{i=1}^{N} \xi_i$$

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – μη γραμ. διαχ. κλάσεις

- Η συνάρτηση Lagrange για την περίπτωση αυτή γίνεται

$$L(w, w_0, \xi, \lambda, \mu) \equiv \frac{1}{2} w^T w + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \mu_i \xi_i - \sum_{i=1}^{N} \lambda_i [d_i (w^T x_i + w_0) - 1 + \xi_i]$$

(τα λ_i και μ_i είναι πολ/στές Lagrange)

- Οι αντίστοιχες ΚΚΤ συνθήκες είναι:

$$(1)\ \ w = \sum_{i=1}^{N} \lambda_i d_i x_i$$

$$(2)\ \ \sum_{i=1}^{N} \lambda_i d_i = 0$$

$$(3)\ \ C - \mu_i - \lambda_i = 0, i = 1,2,\ldots, N$$

$$(4)\ \ \lambda_i [y_i (w^T x_i + w_0) - 1 + \xi_i] = 0, \quad i = 1,2,\ldots, N$$

$$(5)\ \ \mu_i \xi_i = 0, \quad i = 1,2,\ldots, N$$

$$(6)\ \ \mu_i, \lambda_i \geq 0, \quad i = 1,2,\ldots, N$$

➢Αντικαθιστώντας τα παραπάνω στην L(.), οι πολ/στές Lagrange προσδιορίζονται ως οι λύσεις του ακόλουθου δυϊκού SVM προβλήματος

Μεγιστοποίησε $_{\lambda\geq 0}(\sum_{i=1}^{N}\lambda_i - \frac{1}{2}\sum_{ij}\lambda_i\lambda_j d_i d_j x_i^T x_j)$ ως προς το $\boldsymbol{\lambda}=[\lambda_1,\ldots,\lambda_N]^T$

υπό τις προϋποθέσεις

$$0 \leq \lambda_i \leq C,\ i = 1,2,\ldots,N$$
$$\sum_{i=1}^{N}\lambda_i d_i = 0$$

➢**<u>Σχόλιο</u>**:

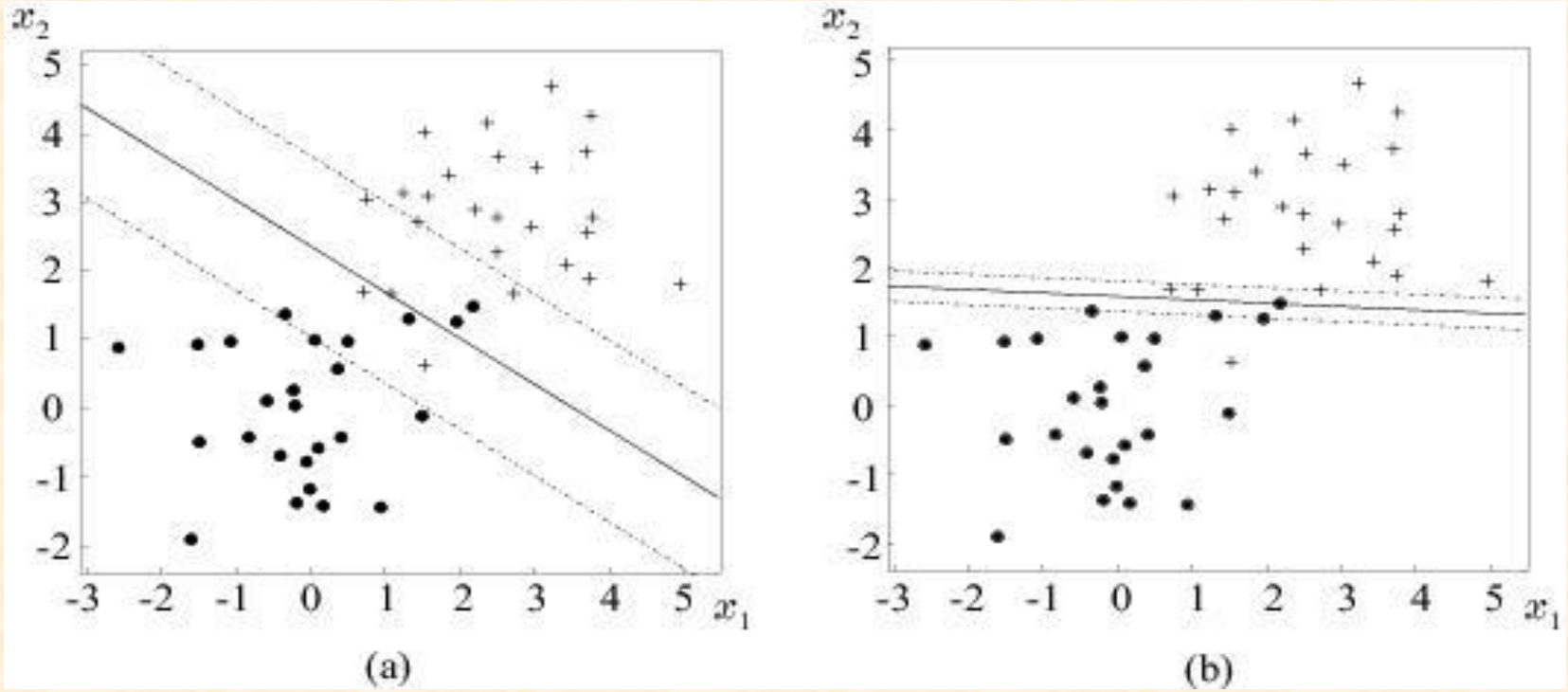
Η μόνη διαφορά με την περίπτωση των διαχωρίσιμων κλάσεων είναι η ύπαρξη του *C* στους περιορισμούς.

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – μη γραμ. διαχ. κλάσεις

Η λύση του SVM προβλήματος ακολουθεί βήματα ανάλογα με αυτά της προηγούμενης περίπτωσης.

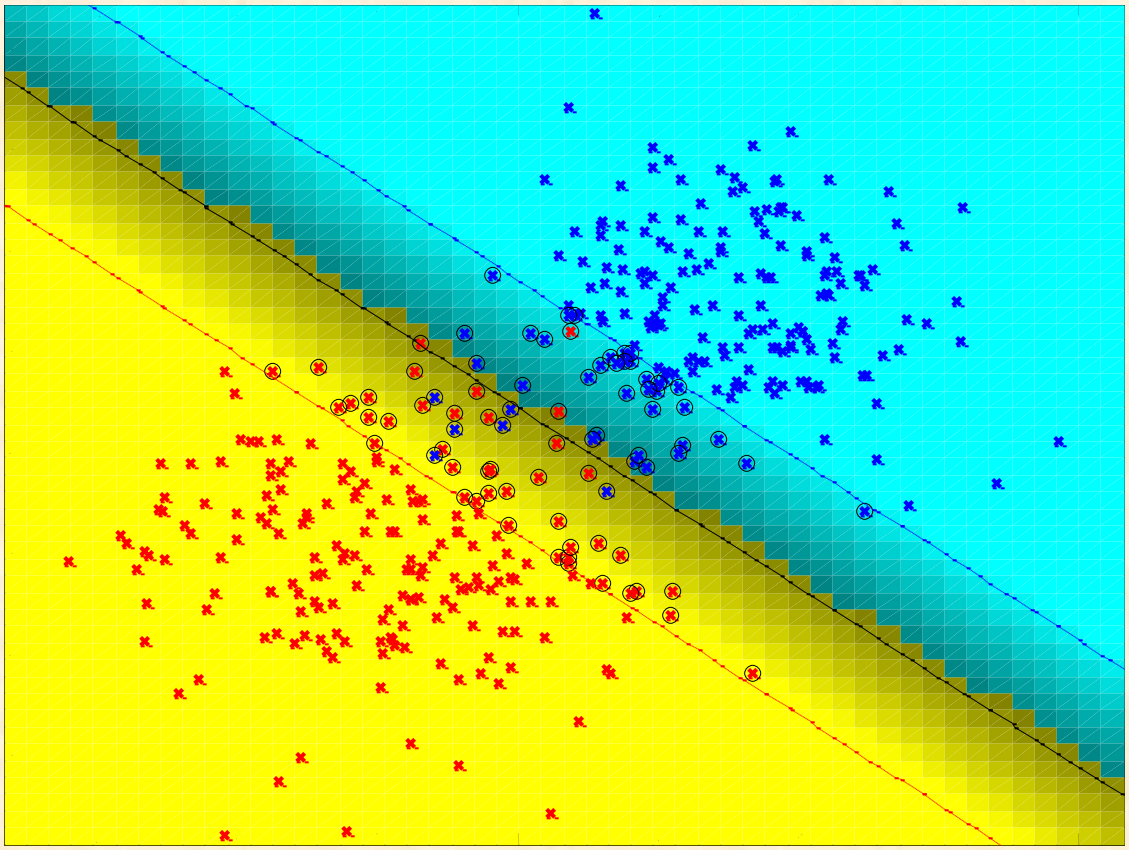
Ωστόσο, στην παρούσα περίπτωση, η παράμετρος *C* επηρεάζει την επιλογή της τελικής λύσης.



Στο παραπάνω παράδειγμα η *C* έχει "μικρότερη τιμή" για το σχήμα (a) και "μεγαλύτερη τιμή" για το σχήμα (b).

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – μη γραμ. διαχ. κλάσεις



C =0.1

Pe_tr =0.0225

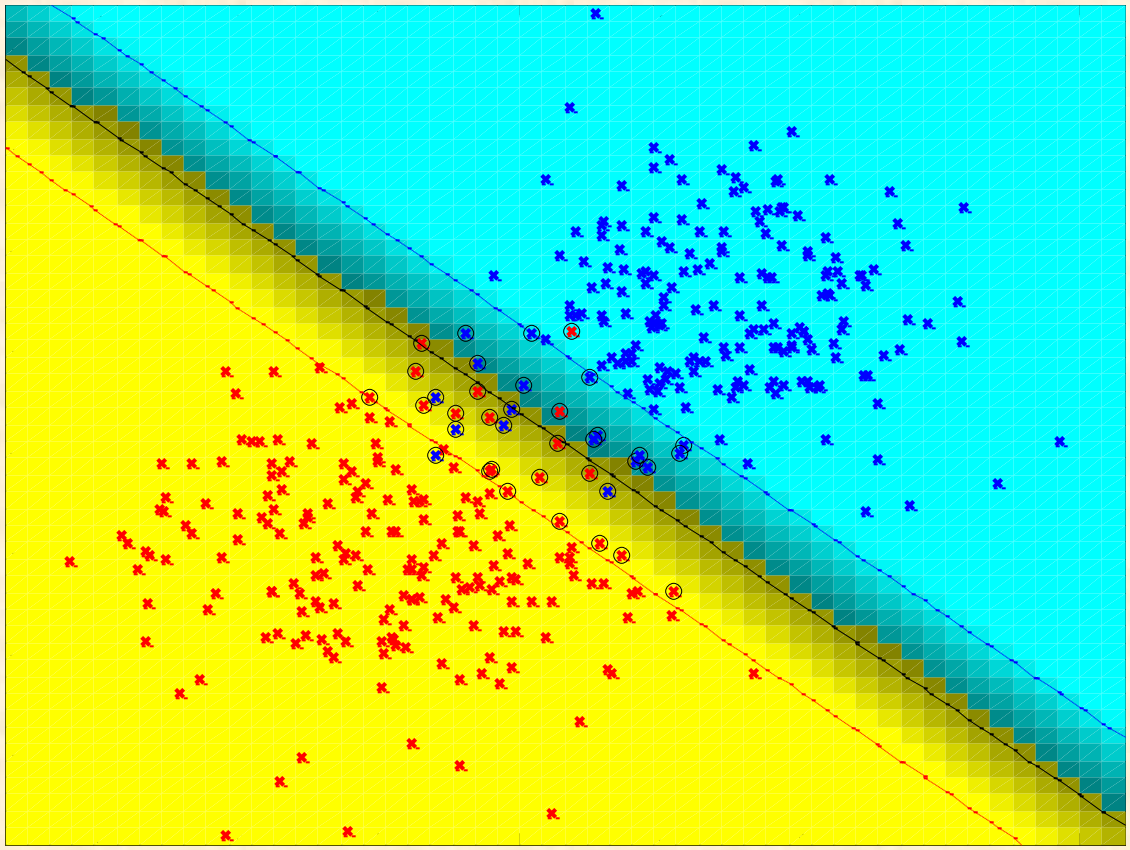
Pe_te =0.0325

sup_vec =82

marg = 0.9412

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – μη γραμ. διαχ. κλάσεις



C =1

Pe_tr =0.0225

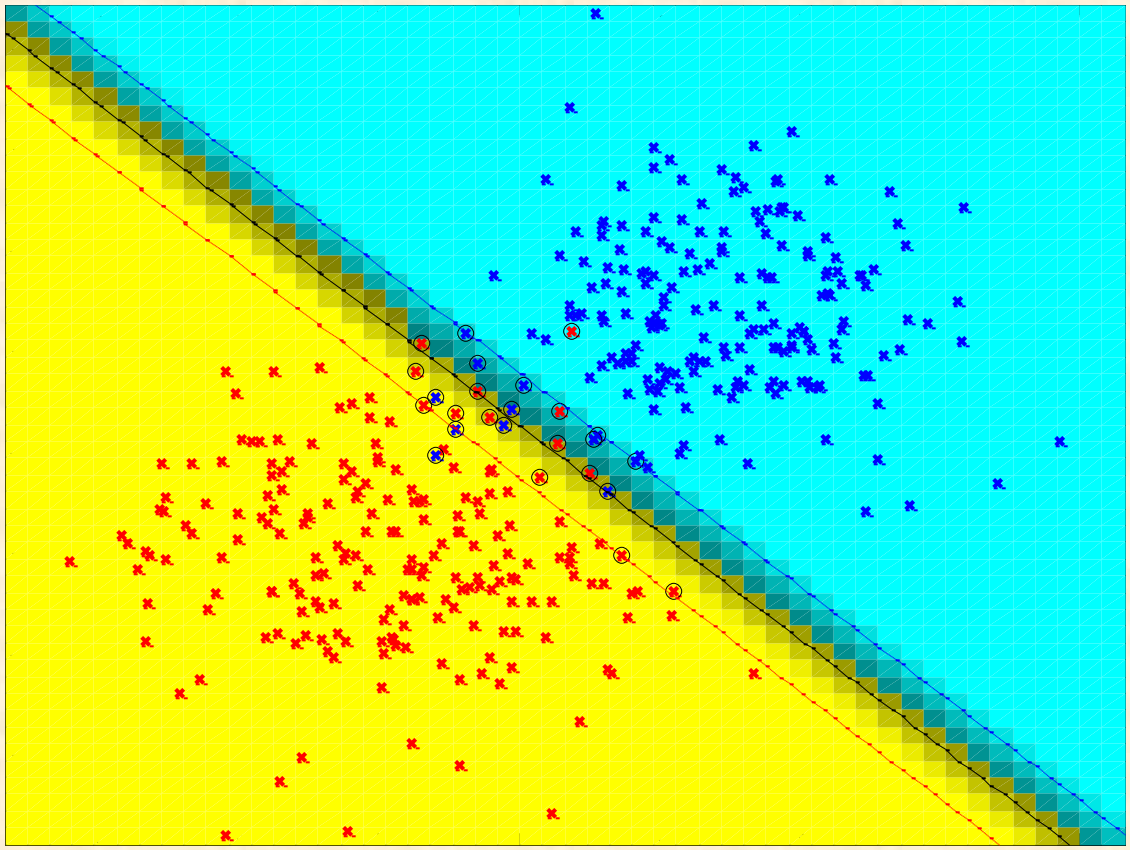
Pe_te =0.0325

sup_vec =37

marg = 0.6317

ΓΡΑΜΜΙΚΟΙ ΠΑΡΑΜΕΤΡΙΚΟΙ ΤΑΞΙΝΟΜΗΤΕΣ ΕΠΙΦΑΝΕΙΩΝ ΑΠΟΦΑΣΗΣ

Support vector machines – η γραμμική περίπτωση – μη γραμ. διαχ. κλάσεις



C =20

Pe_tr =0.0225

Pe_te =0.0350

sup_vec =25

marg = 0.3573