Μηχανές Διανυσμάτων Στήριξης SVM

Αναπλ. Καθηγ. Στελιος Ζήμερας
Τμημα Στατιστικης και Αναλογιστικων – Χρηματοοικονομικων Μαθηματικων
Πανεπιστημιο Αιγαιου
Σαμος

2021

Εισαγωγή

Το Support Vector Machine (SVM) είναι ένα εργαλείο πρόβλεψης ταξινόμησης και παλινδρόμησης που χρησιμοποιεί τη θεωρία μάθησης μηχανών για να μεγιστοποιήσει την προβλεπτική ακρίβεια ενώ αποφεύγει αυτόματα την υπερβολική προσαρμογή (overfitting) στα δεδομένα.

Το SVM έχει μια τεχνική που ονομάζεται κόλπο πυρήνα (kernel trick). Πρόκειται για functions που λαμβάνουν ως είσοδο έναν χώρο χαμηλών διαστάσεων και το μετατρέπουν σε ένα υψηλότερο χώρο διαστάσεων. Μετατρέπουν ουσιαστικά το μη διαχωρίσιμο πρόβλημα σε διαχωρίσιμο και ονομάζονται kernels

Εισαγωγή

Τα **πλεονεκτήματα** των μηχανών διανυσμάτων υποστήριξης είναι τα εξής:

1. Αποτελεσματικά σε χώρους μεγάλης διάστασης
2. Αποτελεσματικά στις περιπτώσεις όπου ο αριθμός των διαστάσεων είναι μεγαλύτερος από τον αριθμό των δειγμάτων
3. Ευέλικτο: μπορούν να καθοριστούν διαφορετικές συναρτήσεις πυρήνα για τη συνάρτηση απόφασης

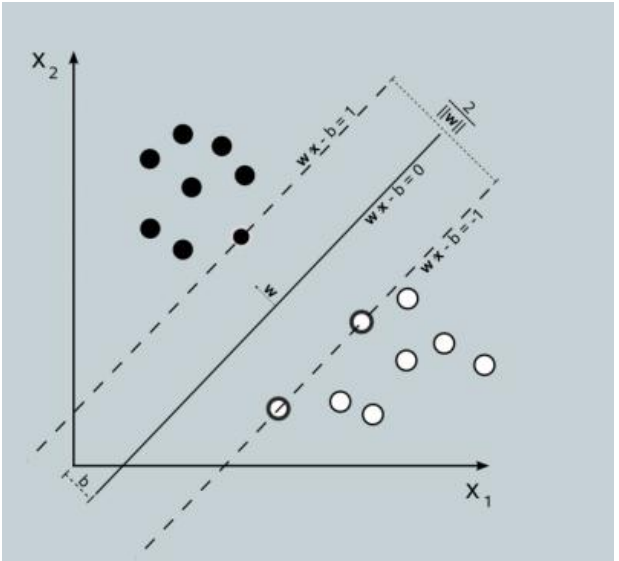
Εισαγωγή

Τα **μειονεκτήματα** των μηχανών διανυσμάτων υποστήριξης είναι τα εξής:

Εάν ο αριθμός των χαρακτηριστικών είναι πολύ μεγαλύτερος από τον αριθμό των δειγμάτων, υπάρχει κίνδυνος για υπερβολικό overfitting και αυτό πρέπει να αποφευχθεί μέσω της επιλογής σωστής συνάρτησης πυρήνα

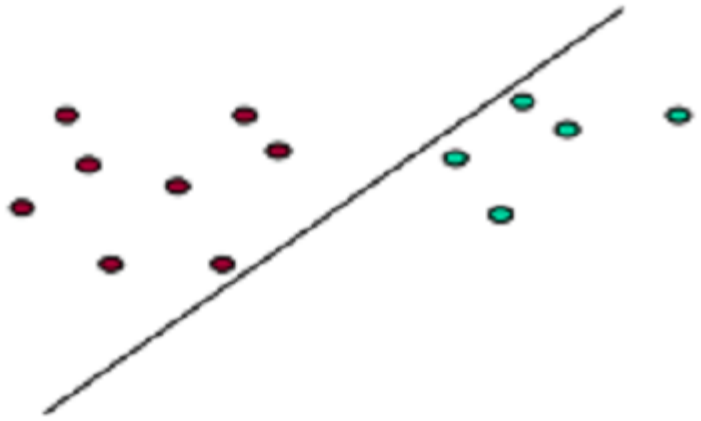
Εισαγωγή

- Προβάλλουν τα σημεία του συνόλου εκπαίδευσης σε έναν χώρο περισσοτέρων διαστάσεων και βρίσκουν το υπερεπίπεδο το οποίο διαχωρίζει βέλτιστα τα σημεία των δύο τάξεων
- Τα άγνωστα σημεία ταξινομούνται σύμφωνα με την πλευρά του υπερεπίπεδου στην οποία βρίσκονται
- Τα διανύσματα τα οποία ορίζουν το υπερεπίπεδο που χωρίζει τις δύο τάξεις ονομάζονται διανύσματα υποστήριξης (support vectors)



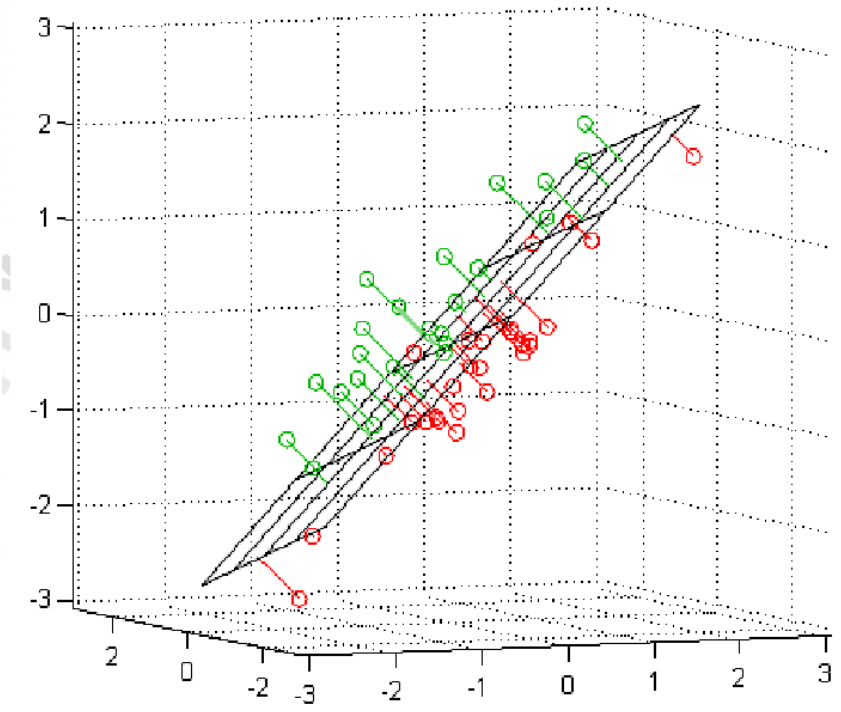
Εισαγωγή

- Στα SVM χρησιμοποιείται συχνά η έννοια του υπερεπιπέδου. Για να γίνει πιο κατανοητός ο όρος υπερεπίπεδα ας υποθέσουμε ότι έχουμε δύο ομάδες σημείων, τα πράσσινα και τα κόκκινα. Αυτό που προσπαθούμε να επιτύχουμε είναι να διαχωρίσουμε τα κόκκινα από τα πράσινα σημεία. Όπως είναι εμφανές και από το παρακάτω σχήμα υπάρχει ευθεία, η οποία διαχωρίζει τις δύο ομάδες.



Εισαγωγή

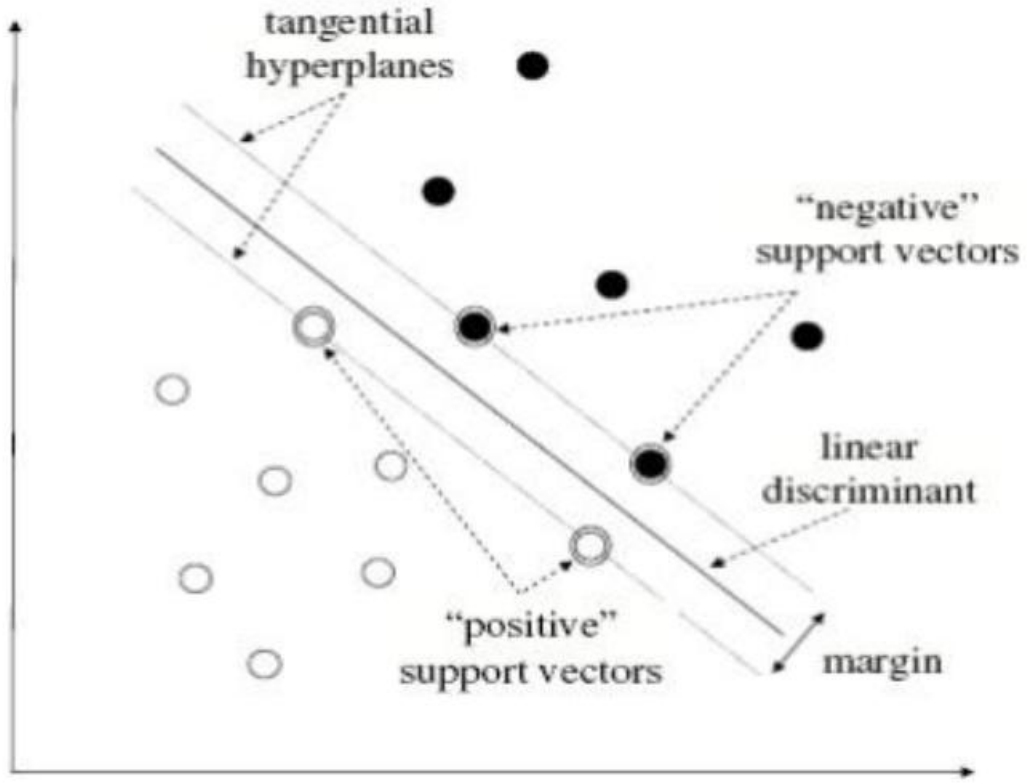
- Έτσι λοιπόν σε δύο διαστάσεις τα σημεία μπορούν να χωρισθούν από μία ευθεία, δηλαδή ένα υπερεπίπεδο μίας διάστασης. Σε τρεις διαστάσεις όπως φαίνεται και στο παρακάτω σχήμα μπορούν να χωρισθούν από ένα επίπεδο, δηλαδή ένα υπερεπίπεδο δύο διαστάσεων.



Εισαγωγή

- Ο κύριος σκοπός αυτής της μεθόδου είναι να βρούμε το βέλτιστο υπερεπίπεδο το οποίο διαχωρίζει καλύτερα τα σημεία μας
- Το βέλτιστο υπερεπίπεδο ονομάζεται υπερεπίπεδο μέγιστου εύρους (maximum margin hyperplane)
- Δημιουργούμε δύο παράλληλα υπερεπίπεδα τέτοια ώστε να μην υπάρχουν ανάμεσά τους δεδομένα του συνόλου εκπαίδευσης

Μέθοδος



Τα σημεία που βρίσκονται πάνω σε αυτά τα δύο υπερεπίπεδα ονομάζονται support vectors και γι' αυτόν τον λόγο και το μοντέλο ονομάζεται Support Vector Machine. Η απόσταση που είναι και η μέγιστη μεταξύ των δύο παράλληλων υπερεπιπέδων ονομάζεται εύρος

Μέθοδος

- Η εξίσωση η οποία συμβολίζει το υπερεπίπεδο έχει την μορφή:

$$\boldsymbol{w} \cdot \boldsymbol{d} + b = 0$$

Με w και b να είναι οι παράμετροι του μοντέλου

D={d_1 , d_2 , …, d_n } το σύνολο των δεδομένων εκπαίδευσης

C={c_1 , c_2 } σύνολο των κατηγοριών

$c_i \in$ {-1,+1} με 1 γνησια δηλωση και -1 εσφαλμενη σηλωση

Για όσων δεδομένων τα διανύσματα τους βρίσκονται πάνω στο υπερεπίπεδο θα επαληθεύουν την εξίσωση

$$\boldsymbol{w} \cdot \boldsymbol{d} + b = 0$$

ενώ τα διανύσματα των υπόλοιπων δεδομένων θα επαληθεύουν την εξίσωση $\boldsymbol{w} \cdot \boldsymbol{d} + b = m$

Μέθοδος

$\boldsymbol{w} \cdot \boldsymbol{d} + b > 0$ τα δεδομένα βρίσκονται πάνω από το υπερεπίπεδο-όριο

$\boldsymbol{w} \cdot \boldsymbol{d} + b < 0$ τα δεδομένα βρίσκονται κατω από το υπερεπίπεδο-όριο

τα παράλληλα υπερεπίπεδα εκφράζονται

$$\boldsymbol{w} \cdot \boldsymbol{d} + b = 1$$
$$\boldsymbol{w} \cdot \boldsymbol{d} + b = -1$$

d_1 βρίσκεται πάνω από το υπερεπίπεδο-όριο

d_2 βρίσκεται κατω από το υπερεπίπεδο-όριο

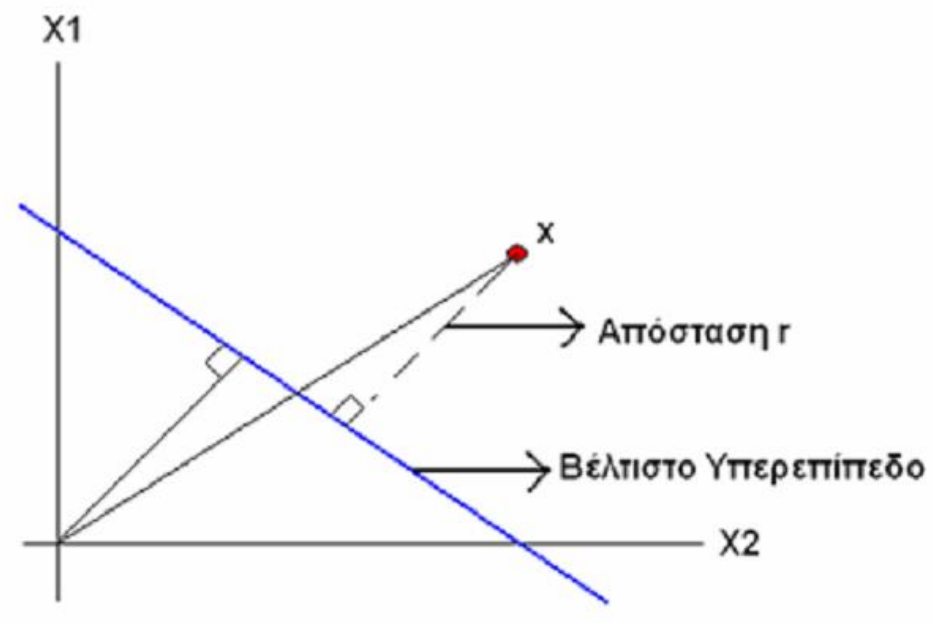
εύρος $(margin)$

$$\left.\begin{matrix} \boldsymbol{w} \cdot \boldsymbol{d_1} + b = 1 \\ \boldsymbol{w} \cdot \boldsymbol{d_2} + b = -1 \end{matrix}\right\} \Leftrightarrow \boldsymbol{w}(\boldsymbol{d_1} - \boldsymbol{d_2}) = 2$$

$$margin = \frac{2}{\|\boldsymbol{w}\|}$$

Μέθοδος

Η απόσταση από την αρχή των αξόνων (x=0) της βέλτιστης διαχωριστικής επιφάνειας δίνεται από τον τύπο $\frac{b_0}{\|w_0\|}$. Εάν $b_0 > 0$, η αρχή των αξόνων βρίσκεται από την θετική πλευρά του βέλτιστου υπερεπιπέδου, εάν $b_0 < 0$ στην αρνητική πλευρά και τέλος εάν $b_0 = 0$ το βέλτιστο υπερεπίπεδο περνά από την αρχή των αξόνων.



Μέθοδος

- Βάσει της προσεγγίσεως των Support Vector Machines, ο«βέλτιστος διαχωριστής» είναι αυτός για τον οποίο, για τα κοντινότερα αντικείμενα προς ταξινόμηση ισχύει

$$f(x) = w \cdot x + b \quad \Rightarrow \quad f(x) = \pm 1$$

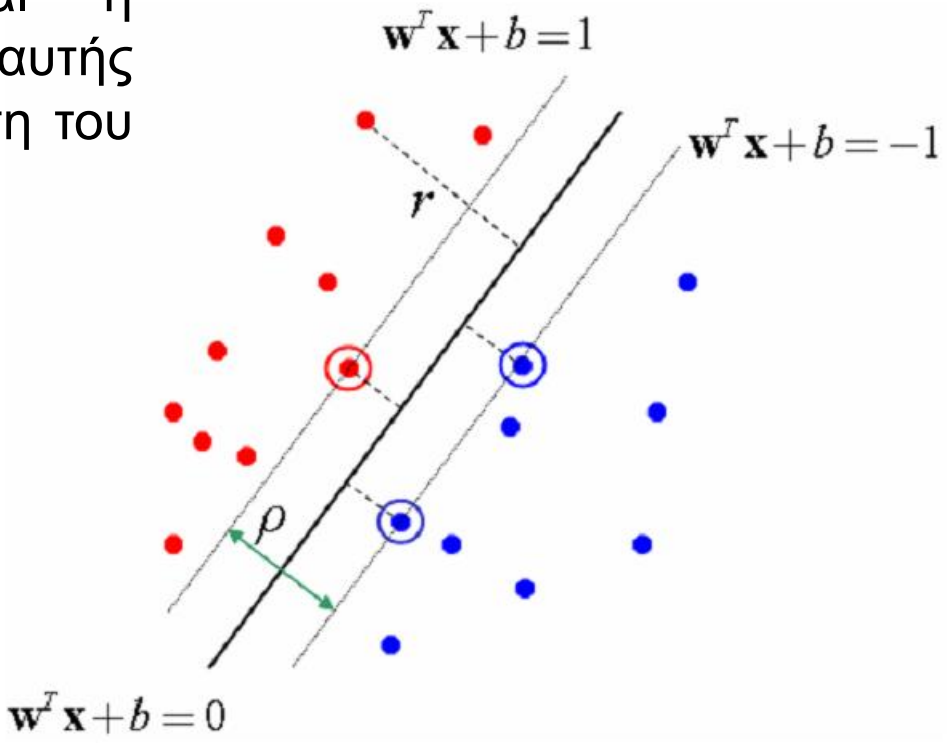
- Η απόσταση d ενός αντικειμένου από τη διαχωριστική υπερεπιφάνεια

$$d = \frac{w \cdot x + b}{\|w\|}$$

Μέθοδος

Στόχος της μεθοδολογίας είναι η μεγιστοποίηση της απόστασης αυτής ή εναλλακτικά η ελαχιστοποίηση του όρου

$$\frac{1}{2} w^T w$$



$$\min \quad \frac{1}{2} w^T w$$

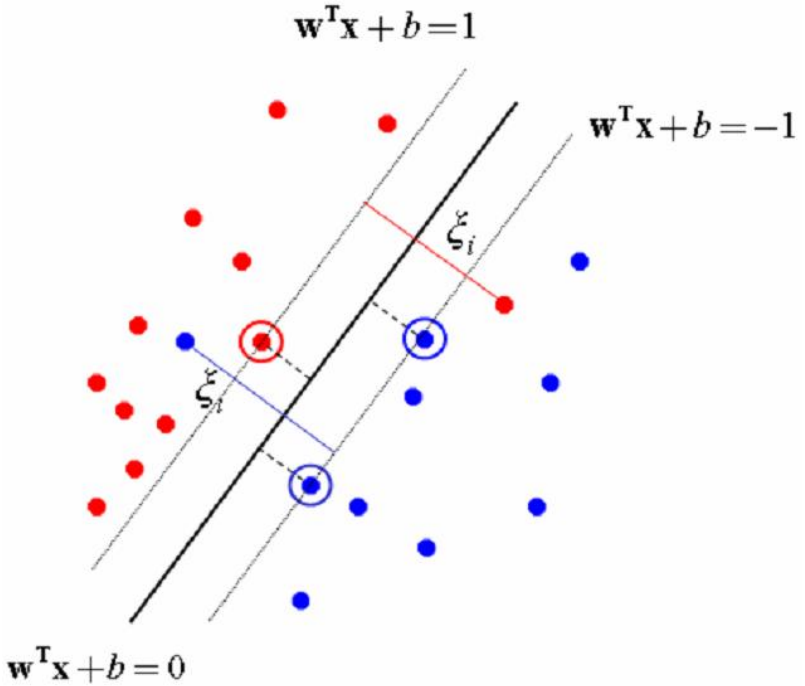
υ.π.

$$w^T x_i + b \geq 1, \quad \text{αν } y_i = 1$$

$$w^T x_i + b \leq -1, \quad \text{αν } y_i = -1$$

Μέθοδος

Στην περίπτωση όμως που ο γραμμικός διαχωρισμός των αντικειμένων δεν είναι εφικτός, εισάγονται στο πρόβλημα κάποιες μεταβλητές απόκλισης-σφάλματος ξ_i



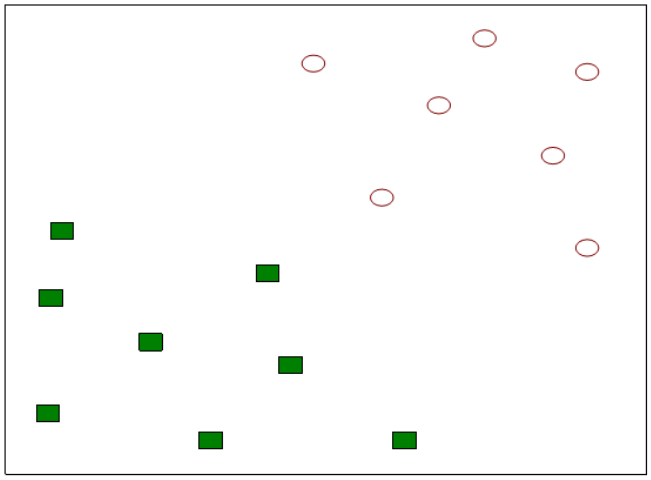
$$\min \quad \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i$$

$$\text{υ.π.} \quad y_i [w^T x_i + b] \geq 1 - \xi_i$$

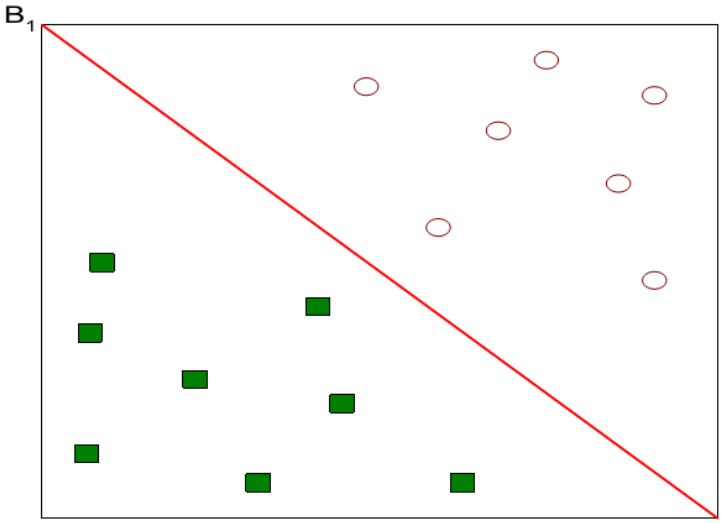
$$\xi_i \geq 0, \ i = 1, \ldots, m$$

όπου, C μια σταθερά (trade-off parameter) η οποία αναπαριστά στην αντικειμενική συνάρτηση τη σχέση μεταξύ της απόστασης των δύο κατηγοριών και των σφαλμάτων ταξινόμησης.

Μέθοδος

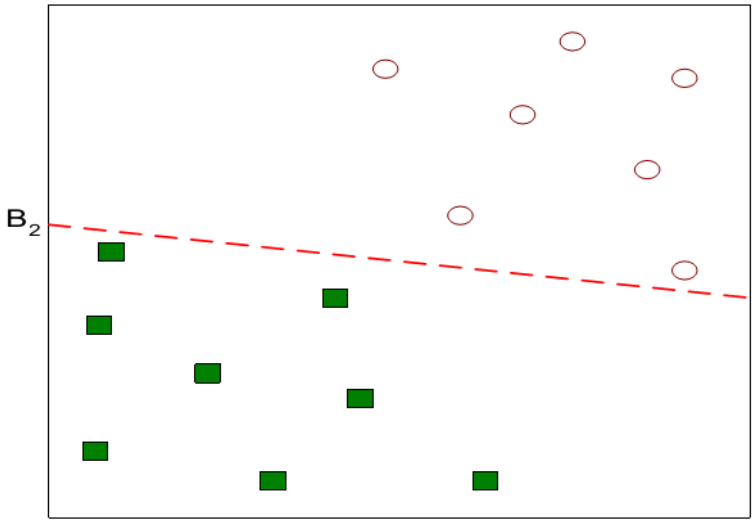


Εύρεση ενός γραμμικού ορίου απόφασης (hyperplane) που διαχωρίζει τα δεδομένα

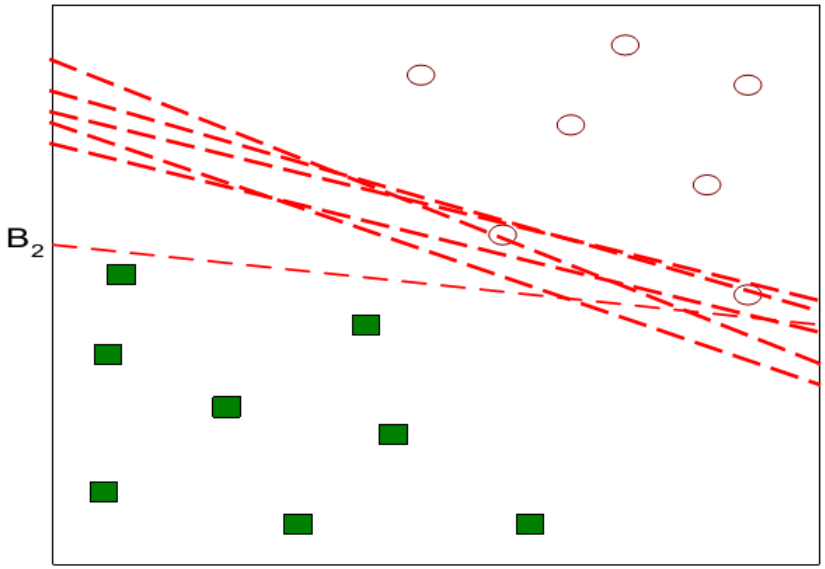


Μια πιθανή λύση

Μέθοδος

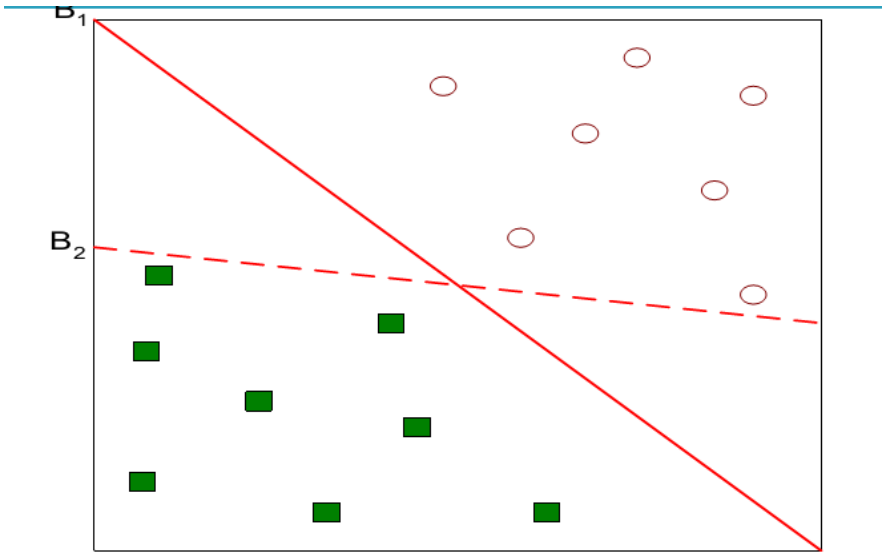


Μια εναλλακτική λύση

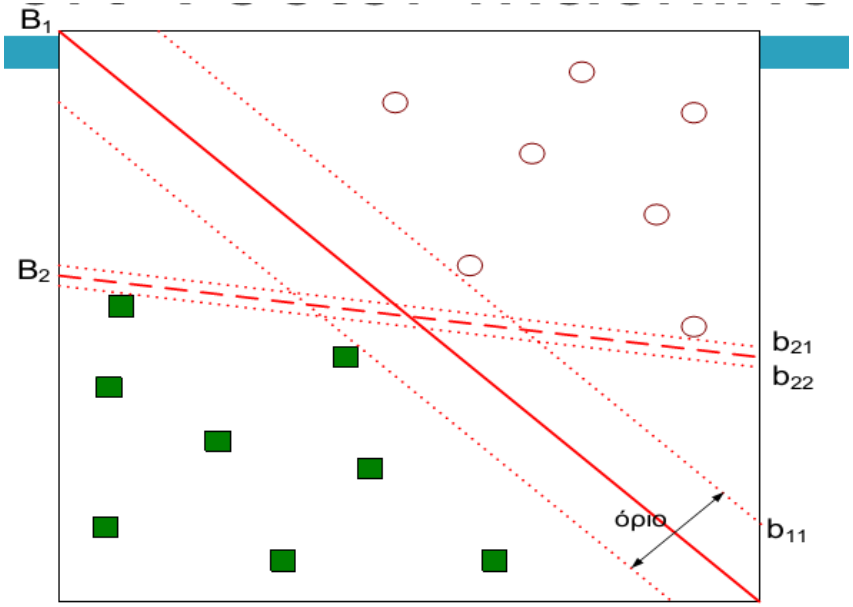


πιθανές λύσεις

Μέθοδος

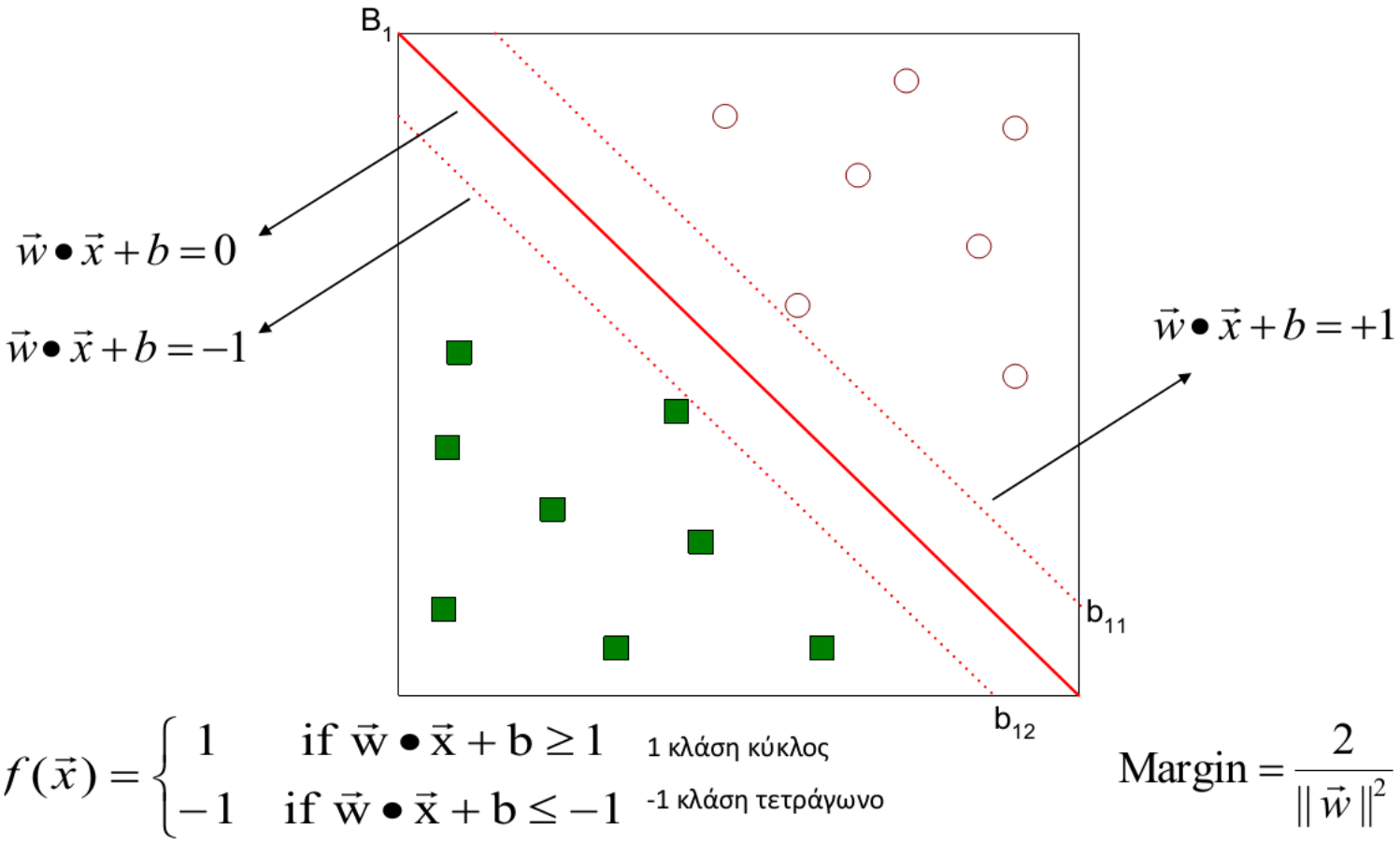


Ποια είναι καλύτερη λύση; Τη B1 ή B2?



Εύρεση του ορίου που μεγιστοποιεί το όριο (margin) που διαχωρίζει τα δεδομένα
Άρα B1 καλύτερη του B2

Μέθοδος



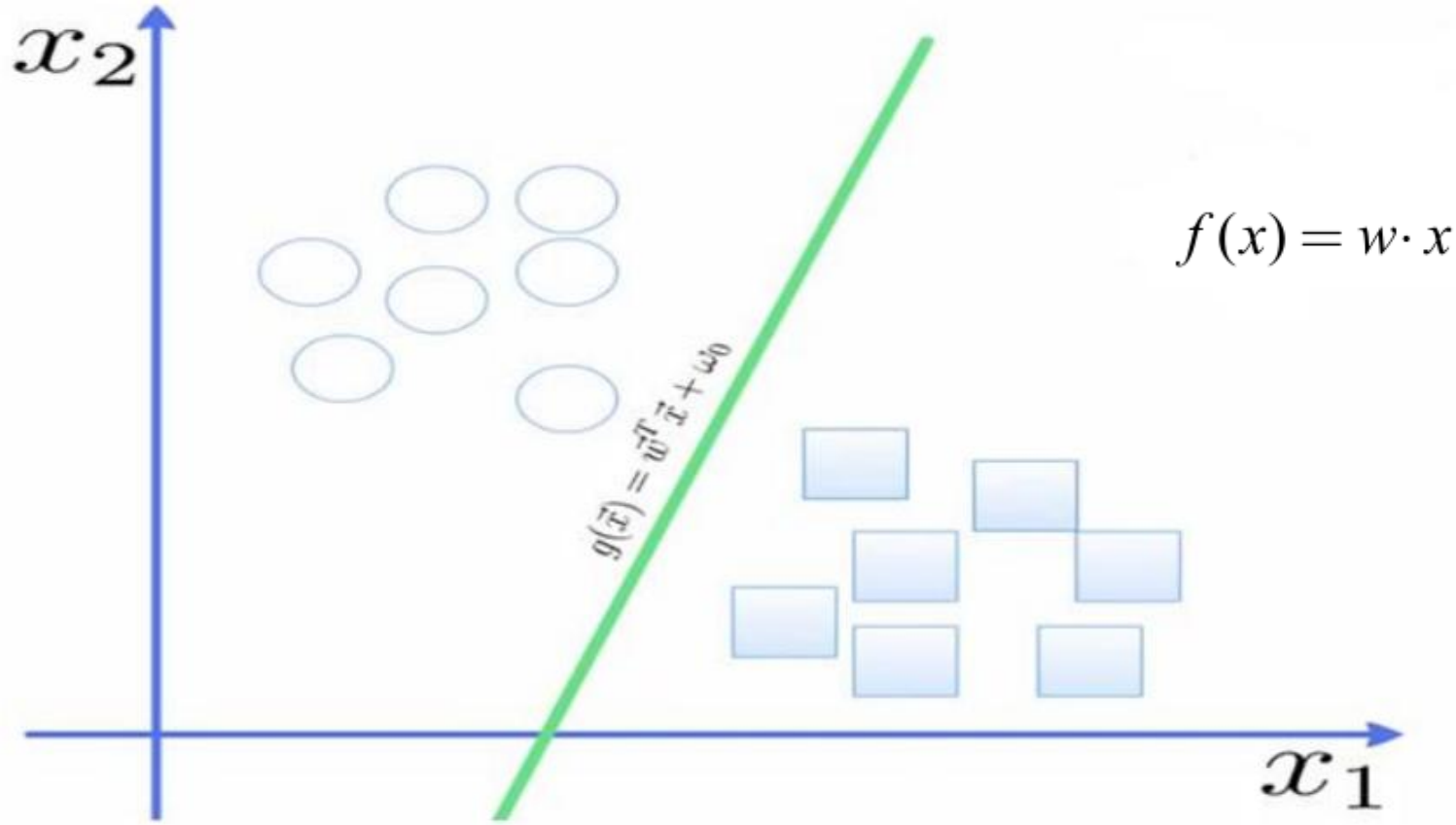
$$f(\vec{x}) = \begin{cases} 1 & \text{if } \vec{w} \bullet \vec{x} + b \geq 1 \\ -1 & \text{if } \vec{w} \bullet \vec{x} + b \leq -1 \end{cases}$$

1 κλάση κύκλος

-1 κλάση τετράγωνο

$$\text{Margin} = \frac{2}{\| \vec{w} \|^2}$$

Μέθοδος



$$f(x) = w \cdot x + b$$

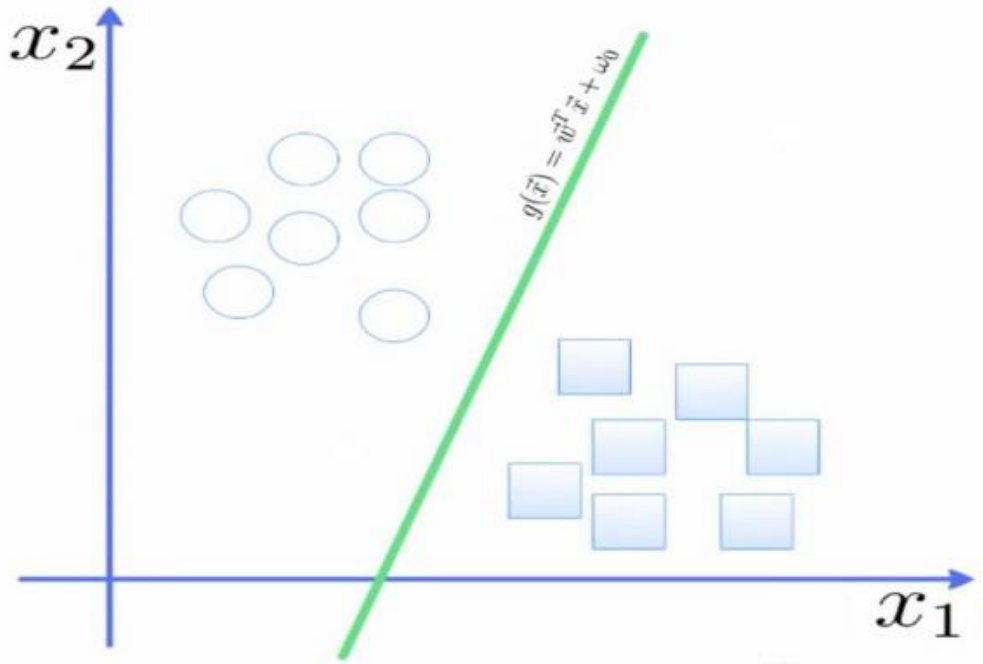
το εσωτερικό γινόμενο διανυσμάτων περιγράφεται:

x=[$x_1, x_2, \ldots, x_N$] και z=[$z_1, z_2, \ldots, z_N$]

$$\sum_{v=1}^{N} x_v . z_v = x^T . z$$

Μέθοδος

$$g(\vec{x}) \geqslant 1, \quad \forall \vec{x} \in \text{class } 1$$
$$g(\vec{x}) \leqslant -1, \quad \forall \vec{x} \in \text{class } 2$$



$$f(x) = \pm 1$$

Μέθοδος

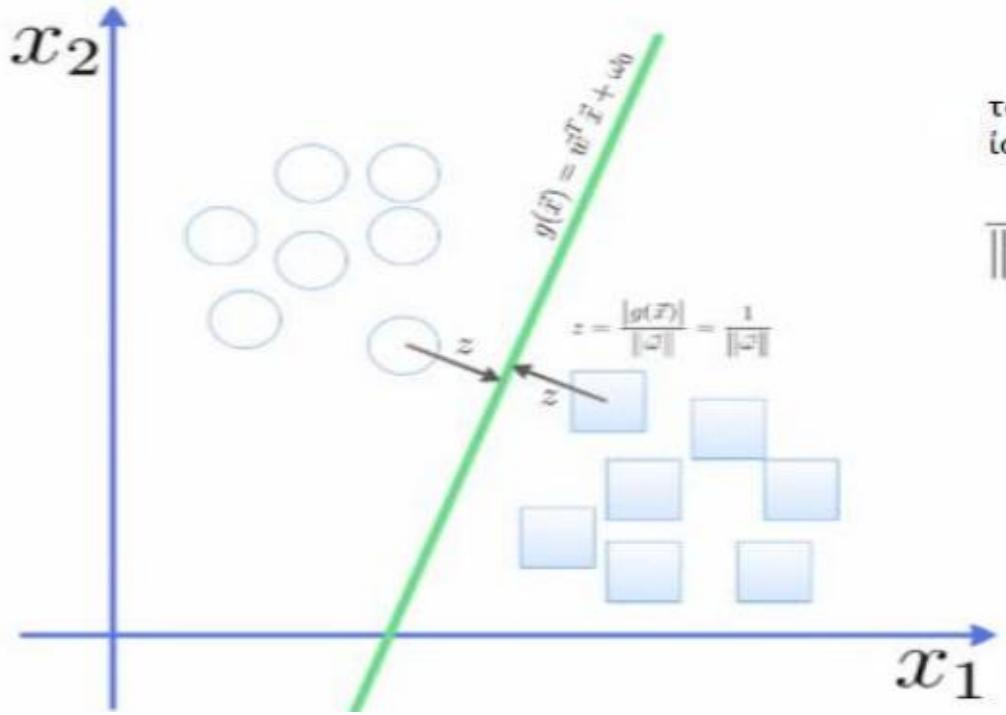
- Θέλουμε να μεγιστοποιήσουμε: $$\text{Margin} = \frac{2}{\| \vec{w} \|^2}$$
  - Το οποίο είναι ισοδύναμο με το να ελαχιστοποιήσουμε: $$L(w) = \frac{\| \vec{w} \|^2}{2}$$
  - Με βάση τους παρακάτω περιορισμούς (constraints):

$$f(\vec{x}_i) = \begin{cases} 1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_\mathrm{i} + \mathrm{b} \geq 1 \\ -1 & \text{if } \vec{\mathrm{w}} \bullet \vec{\mathrm{x}}_\mathrm{i} + \mathrm{b} \leq -1 \end{cases}$$

Μέθοδος

$$d = \frac{w \cdot x + b}{\|w\|}$$

$$g(\vec{x}) \geqslant 1, \quad \forall \vec{x} \in \text{class } 1$$
$$g(\vec{x}) \leqslant -1, \quad \forall \vec{x} \in \text{class } 2$$



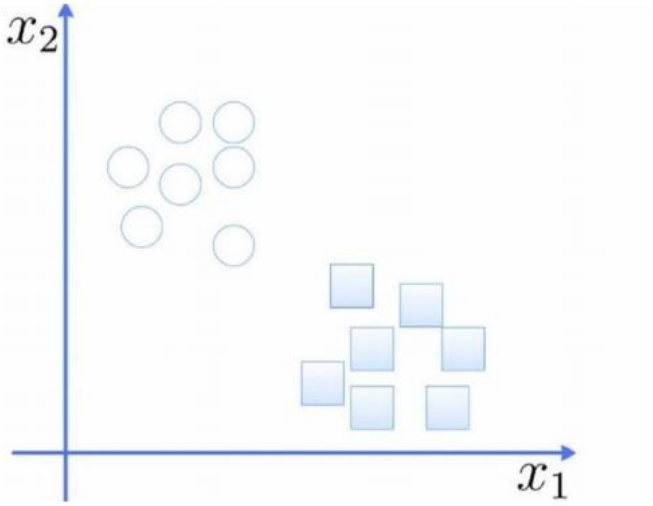
το συνολικό όριο είναι ίσο με:

$$\frac{1}{\|\vec{\omega}\|} + \frac{1}{\|\vec{\omega}\|} = \frac{2}{\|\vec{\omega}\|}$$

ελαχιστοποιώντας αυτόν τον όρο, μεγιστοποιούμε το διαχωρισμό!

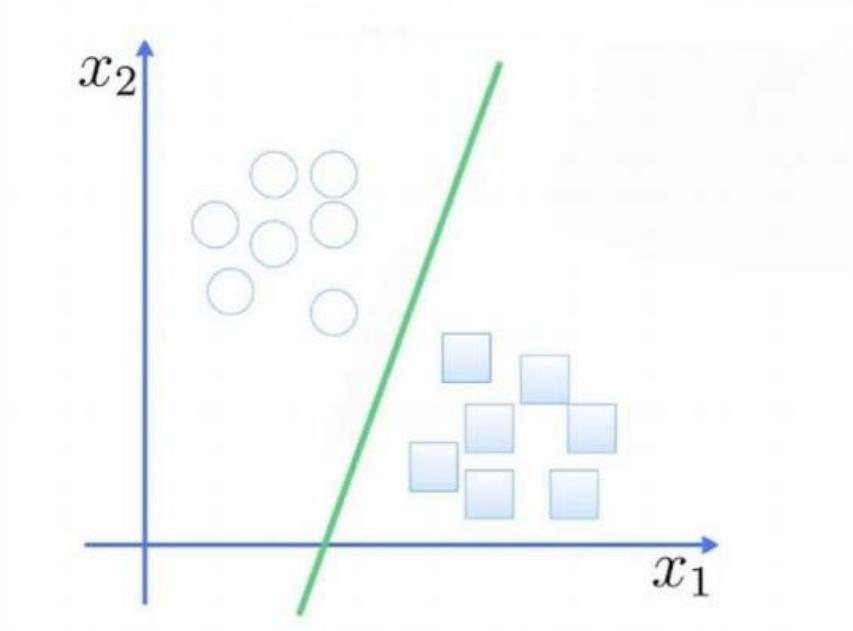
Παράδειγμα

Έστω ότι έχουμε ένα καρτεσιανό επίπεδο με δύο άξονες X_1 και X_2 και πρέπει να ταξινομήσουμε όλα τα στοιχεία που βρίσκονται στο καρτεσιανό μας επίπεδο όπως φαίνονται και παρακάτω στο σχήμα.



Παράδειγμα

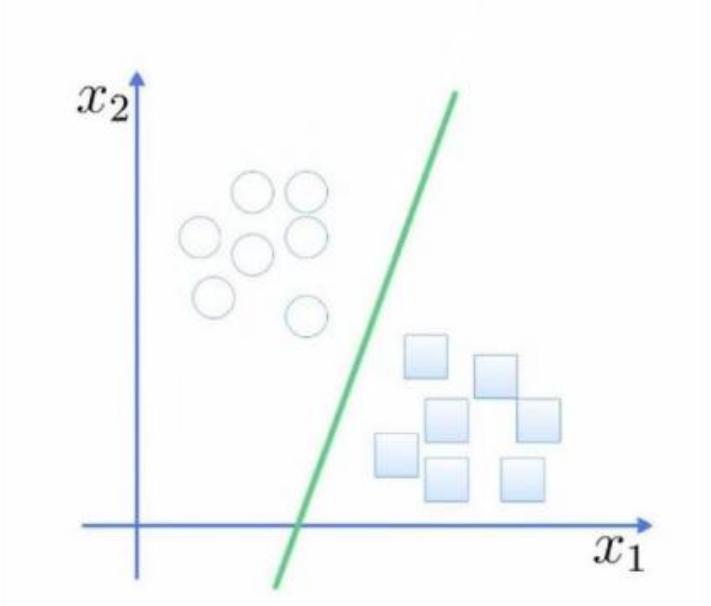
Στη συνέχεια για να ταξινομήσουμε τα στοιχεία που έχουμε πάνω στο επίπεδο μας βάζουμε ως στόχο να ταξινομηθούν όλα τα στοιχεία σε δύο κατηγορίες. Αυτό το καταφέρνουμε σχεδιάζοντας μια ευθεία γραμμή ενδιάμεσα στα κυκλικά και τα τετράγωνα στοιχεία, όπως φαίνεται και το σχήμα παρακάτω.

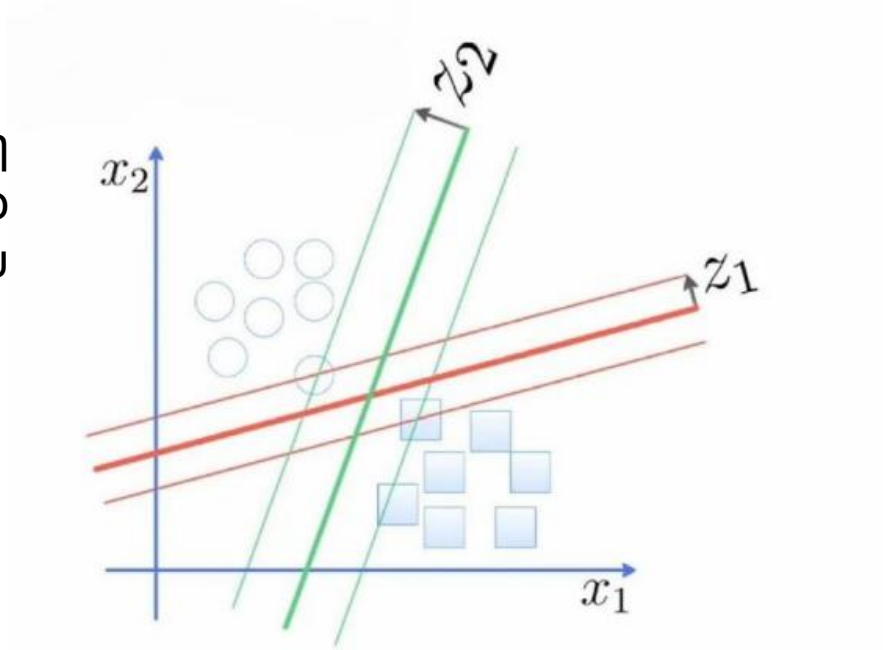


Παράδειγμα

Βέβαια υπάρχουν διάφορες κλίσεις που μπορείς να σχεδιάσεις την ευθεία γραμμή πάνω στο καρτεσιανό επίπεδο, όπως φαίνονται και στο σχήμα παρακάτω.

θα πρέπει να διαλέξουμε την καλύτερη επιλογή η οποία θα έχει το μέγιστο περιθώριο από τις δύο κατηγορίες. Όπου στην δική μας περίπτωση είναι Z2 (Z2 >Z1).





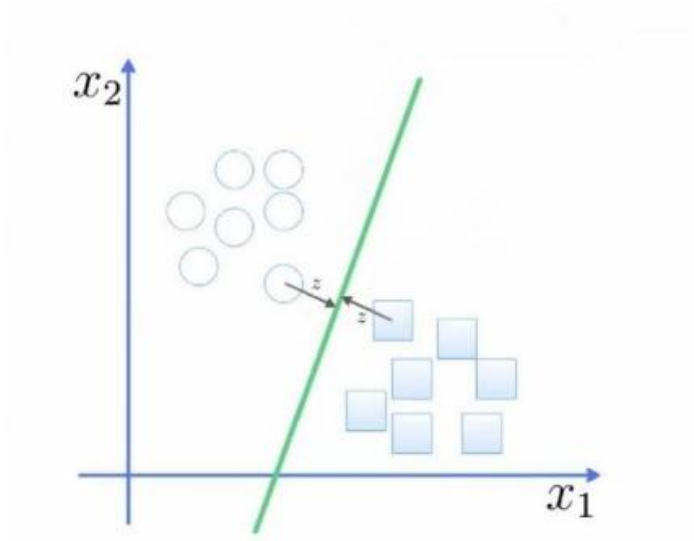
$$g(\vec{x}) \geq 1, \forall \vec{x} \in class1\ (\kappa\acute{\upsilon}\kappa\lambda o\iota)$$

$$g(\vec{x}) \leq -1, \forall \vec{x} \in class2\ (\tau\varepsilon\tau\rho\acute{\alpha}\gamma\omega\nu\alpha)$$

Παράδειγμα

Έστω ότι η απόσταση (Z) μεταξύ της καλύτερης επιλογής γραμμής και του πιο κοντινού στοιχείου από κάθε κατηγορία είναι τουλάχιστον 1. Οπότε με βάση τη γεωμετρία έχουμε ότι

$$Z = \frac{|g(\vec{x})|}{||\vec{\omega}||} = \frac{1}{||\vec{\omega}||}$$



Οπότε με βάση τα παραπάνω έχουμε ότι το συνολικό περιθώριο υπολογίζεται από τον παρακάτω τύπο:

$$\frac{1}{||\vec{\omega}||} + \frac{1}{||\vec{\omega}||} = \frac{2}{||\vec{\omega}||}$$

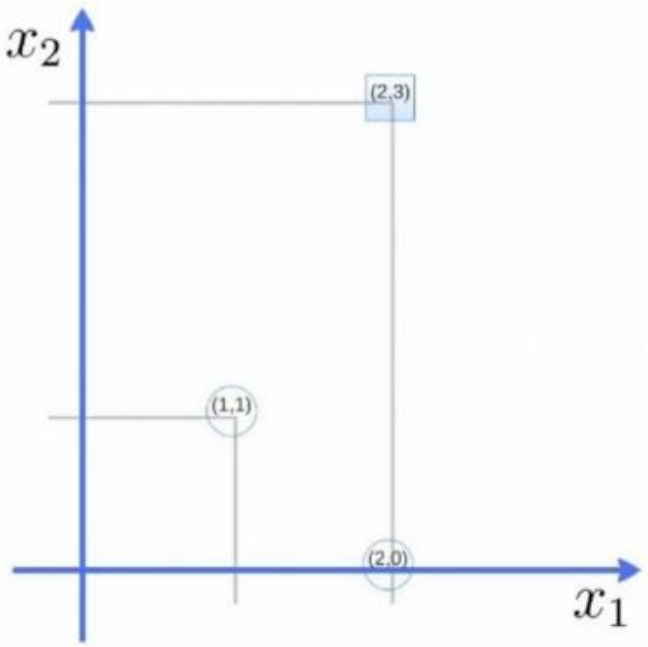
Παράδειγμα

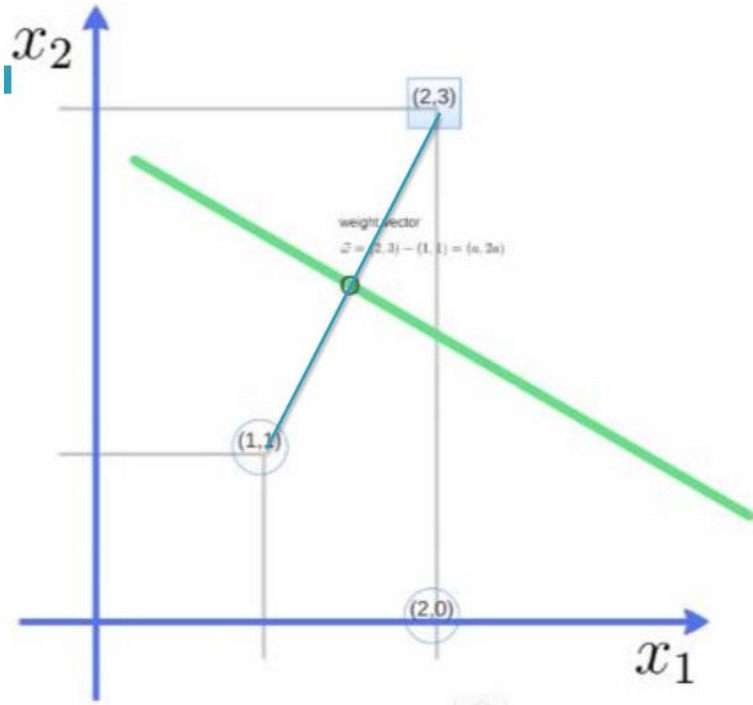
Και εάν ελαχιστοποιήσουμε των παρονομαστή στο τέλος θα έχουμε σαν αποτέλεσμα να μεγιστοποιηθεί η διαχωριστικότητα. Για την ελαχιστοποίηση του $\vec{\omega}$ είναι μια εργασία μη γραμμικής βελτιστοποίησης η οποία μπορεί να λυθεί από τις Karush-Kuhn-Tucker (KKT) συνθήκες, χρησιμοποιώντας πολλαπλασιαστές Langrange λ_ι

$$\vec{\omega} = \sum_{\iota=0}^{N} \lambda_\iota y_i \vec{x}_i$$

$$\sum_{i=0}^{N} \lambda_i \, y_i = 0$$

Παράδειγμα





Από το παραπάνω καρτεσιανό επίπεδο έχουμε ότι το

$$\vec{\omega} = (2,3) - (1,1) = (\alpha, 2\alpha)$$

και για το στοιχείο στη θέση (1,1) έχουμε ότι

g(1,1) = -1 (διότι το (1,1) είναι κύκλος)

όπως και για το στοιχείο στη θέση (2,3) έχουμε ότι

g(2,3) = 1 (διότι το (2,3) είναι τετράγωνο)

Παράδειγμα

Από όλα τα παραπάνω έχουμε ότι

$$\vec{\omega} = (\alpha, 2\alpha),$$

$\alpha+2\alpha+\omega_0= -1$, (από το σημείο (1,1)) και

$2\alpha+6\alpha+\omega_0= 1$, (από το σημείο (2,3)

$$2\alpha + 6\alpha + \omega_0 = 1 \Rightarrow 8\alpha + \omega_0 = 1 \Rightarrow \omega_0 = 1 - 8\alpha$$

σημείο (1,1)

$$3\alpha + \omega_0 = -1 \Rightarrow 3\alpha + 1 - 8\alpha = -1 \Rightarrow 8\alpha - 3\alpha = 1 + 1 \Rightarrow 5\alpha = 2 \Rightarrow \alpha = {}^{2}/_{5}$$

Παράδειγμα

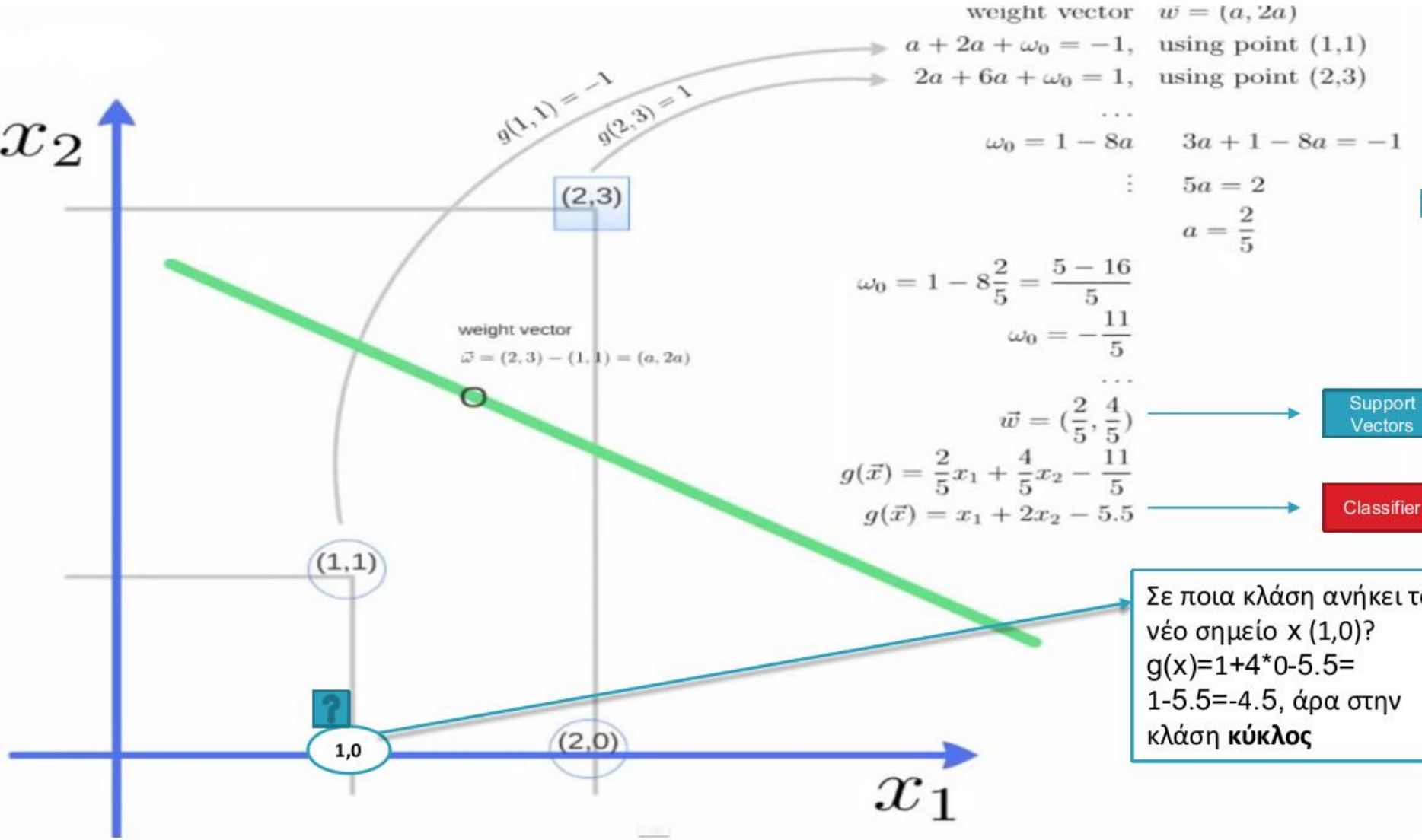
$$\omega_0 = 1 - 8\alpha \Rightarrow \omega_0 = 1 - 8 * \frac{2}{5} = \frac{5 - 16}{5} \Rightarrow \omega_0 = -\frac{11}{5}$$

$$\vec{\omega} = (\frac{2}{5}, \frac{4}{5})$$

$$g(\vec{x}) = \frac{2}{5} * x_1 + \frac{4}{5} * x_2 - \frac{11}{5} \Rightarrow g(\vec{x}) = x_1 + 2x_2 - 5.5$$

Όπου το $\vec{\omega} = (\frac{2}{5}, \frac{4}{5})$ είναι αυτό που λέμε Support Vectors και η τελική μας εξίσωση $g(\vec{x}) = x_1 + 2x_2 - 5.5$ είναι αυτό που λέμε Support Vector Machine.

Παράδειγμα



Σε ποια κλάση ανήκει το νέο σημείο x (1,0)?
g(x)=1+4*0-5.5=
1-5.5=-4.5, άρα στην κλάση **κύκλος**