Μηχανική Μάθηση

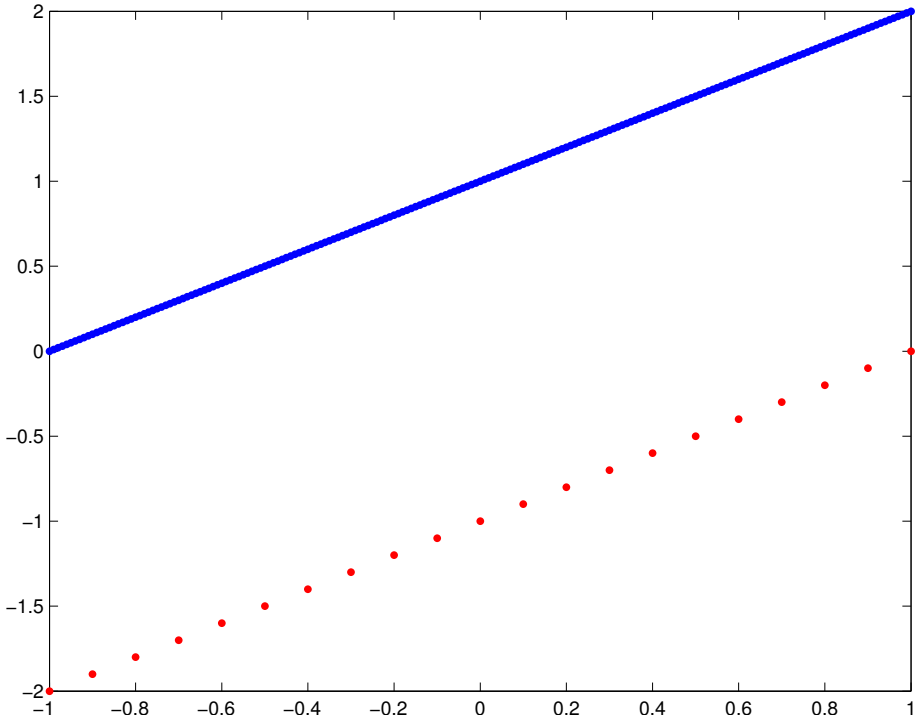
Μιχάλης Τίτσιας

Διάλεξη 6ή

Support Vector Machines

Περιεχόμενα

Support Vector Machines

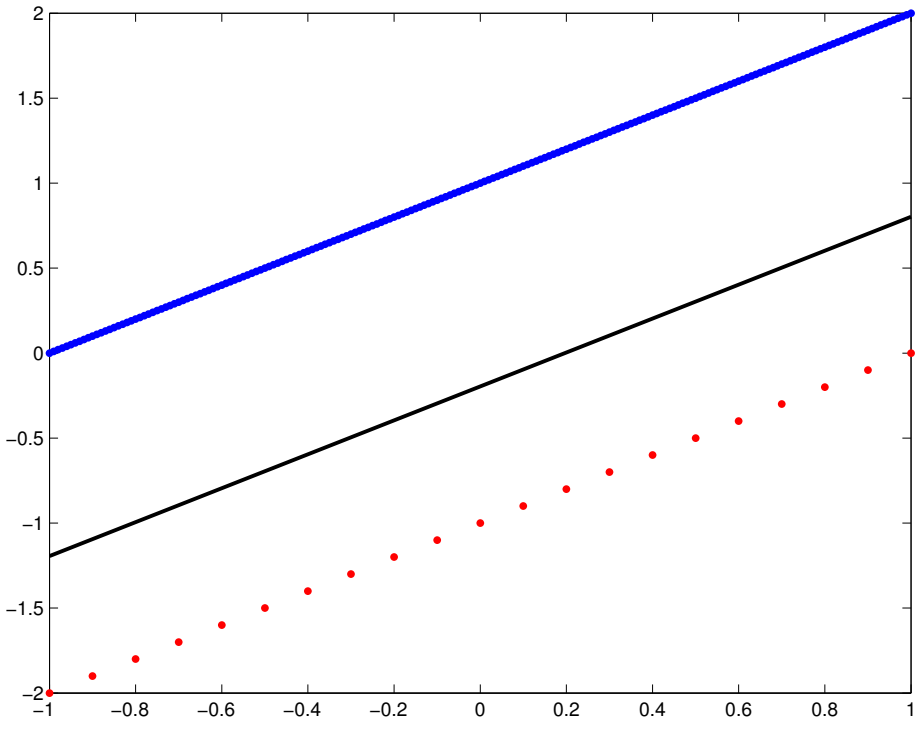


Έστω τα δεδομένα του σχήματος τα οποία είναι διατεταγμένα σε δύο ευθείες

Τα δεδομένα της μπλε κατηγορίας είναι πολύ περισσότερα από αυτής της κόκκινης

- Τι συνέπειες θα έχει αυτό στη λύση που δίνει η λογιστική παλινδρόμηση;

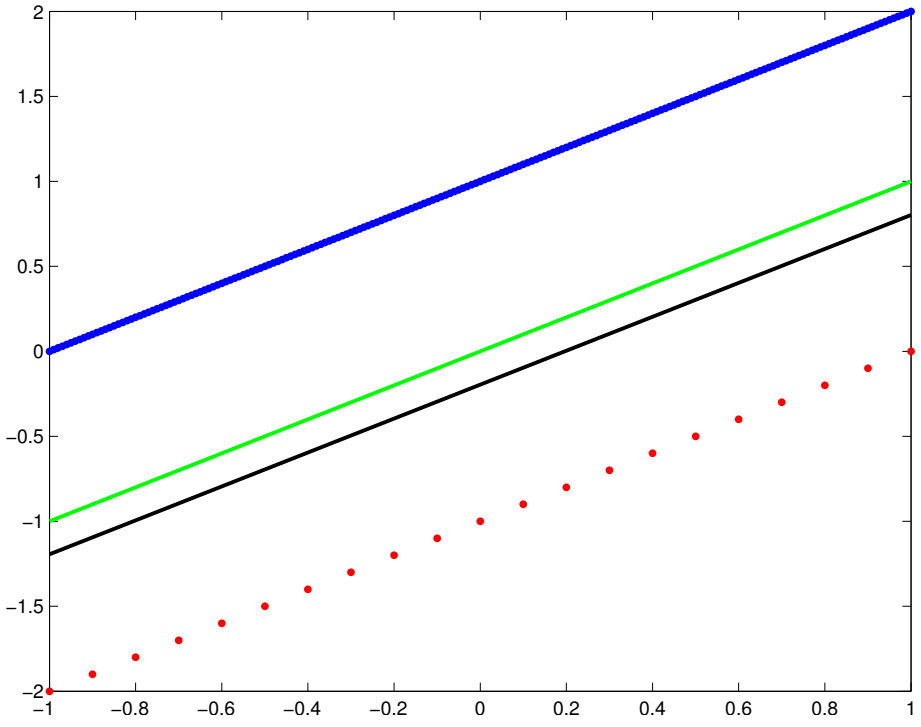
Support Vector Machines



Το σύνορο απόφασης (μαύρη γραμμή) που επιτυγχάνει η λογιστική παλινδρόμηση βρίσκεται πιο κοντά στα δεδομένα της κόκκινης κατηγορίας

- λόγω της συνάρτησης κόστους που ευννοεί την μετακίνηση του συνόρου απόφασης προς τα λίγα δεδομένα!

Support Vector Machines



Ωστόσο αν τοποθετούσαμε το σύνορο απόφασης ακριβώς στο διάμεσο των δεδομένων των δύο κατηγοριών (πράσινη γραμμή) αυτό διαισθητικά τουλάχιστον φαίνεται ως καλύτερη λύση

- δυστυχώς η λύση αυτή δεν μπορεί να επιτευχθεί μέσω του αλγορίθμου εκπαίδευσης της λογιστικής παλινδρόμησης
- πώς θα μπορούσαμε να επιτυχούμε μια τέτοια λύση;

Support Vector Machines

Πώς θα μπορούσαμε να επιτυχούμε μια τέτοια λύση;

- Θα κουβεντιάσουμε πρώτα πως μπορούμε να επιτύχουμε μια τέτοια λύση ξεκινώντας από την συνάρτηση κόστους της λογιστικής παλινδρομήσης και εισάγωντας σιγά σιγά τις κατάλληλες τροποποιήσεις
- ώστε να οδηγηθούμε στα **Support Vector Machines (SVMs)** που επιτυγχάνουν μια τέτοια λύση

Για αύτο το μάθημα, για τα δεδομένα εξόδου t_n θα θεωρήσουμε ότι οι δυαδικές τιμές που παίρνουν δεν είναι οι $\{0, 1\}$ αλλά οι $\{-1, 1\}$, δηλ. τα δεδομένα εξόδου θα είναι τέτοια ώστε

$$t_n \in \{-1, 1\}$$

Αυτό γίνεται καθάρα για λόγους ευκολίας αφού είναι πιο απλό και σύνηθες τα SVMs να παρουσιάζονται χρησιμοποιώντας τις $\{-1, 1\}$ δυαδικές τιμές για τις δύο κατηγορίες

Support Vector Machines

Η συνάρτηση κόστους της λογιστικής παλινδρόμησης (υποθέτωντας ότι τις εξόδους τις αναπαριστούμε με $\{-1, 1\}$ και εφαρμόζοντας ελαχιστοποίηση αντί για μεγιστοποίηση) είναι

$$E(\mathbf{w}) = -\sum_{n=1}^{N}(t_n = 1)\log\left(\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}_n}}\right)+(t_n = -1)\log\left(1-\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}_n}}\right)+\frac{1}{2}\lambda||\mathbf{w}||^2$$

$$E(\mathbf{w}) = -\sum_{n=1}^{N}(t_n = 1)\log\left(\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}_n}}\right)+(t_n = -1)\log\left(\frac{1+e^{-\mathbf{w}^T\mathbf{x}_n}-1}{1+e^{-\mathbf{w}^T\mathbf{x}_n}}\right)+\frac{1}{2}\lambda||\mathbf{w}||^2$$

$$E(\mathbf{w}) = -\sum_{n=1}^{N}(t_n = 1)\log\left(\frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}_n}}\right)+(t_n = -1)\log\left(\frac{1}{1+e^{\mathbf{w}^T\mathbf{x}_n}}\right)+\frac{1}{2}\lambda||\mathbf{w}||^2$$

$$E(\mathbf{w}) = -\sum_{n=1}^{N}\log\left(\frac{1}{1+e^{-t_n\mathbf{w}^T\mathbf{x}_n}}\right)+\frac{1}{2}\lambda||\mathbf{w}||^2$$

$$E(\mathbf{w}) = \sum_{n=1}^{N}\log\left(1+e^{-t_n\mathbf{w}^T\mathbf{x}_n}\right)+\frac{1}{2}\lambda||\mathbf{w}||^2$$

όπου $(t_n = 1)$ και $(t_n = -1)$ συμβολίζουν λογικές συναρτήσεις που παίρνουν τιμές 0 ή 1 ανάλογα με το αν η συνθήκη εντός της παρένθεσης ικανοποιείται

$$E(\mathbf{w}) = \sum_{n=1}^{N} \log\left(1 + e^{-t_n \mathbf{w}^T \mathbf{x}_n}\right) + \frac{1}{2}\lambda ||\mathbf{w}||^2$$

Η σημαντική ποσότητα σε αυτή τη συνάρτηση κόστους (την οποία θέλουμε να ελαχιστοποιήσουμε) είναι το κόστος ανά δεδομένο

$$E_n = \log\left(1 + e^{-t_n \mathbf{w}^T \mathbf{x}_n}\right) \geq 0$$

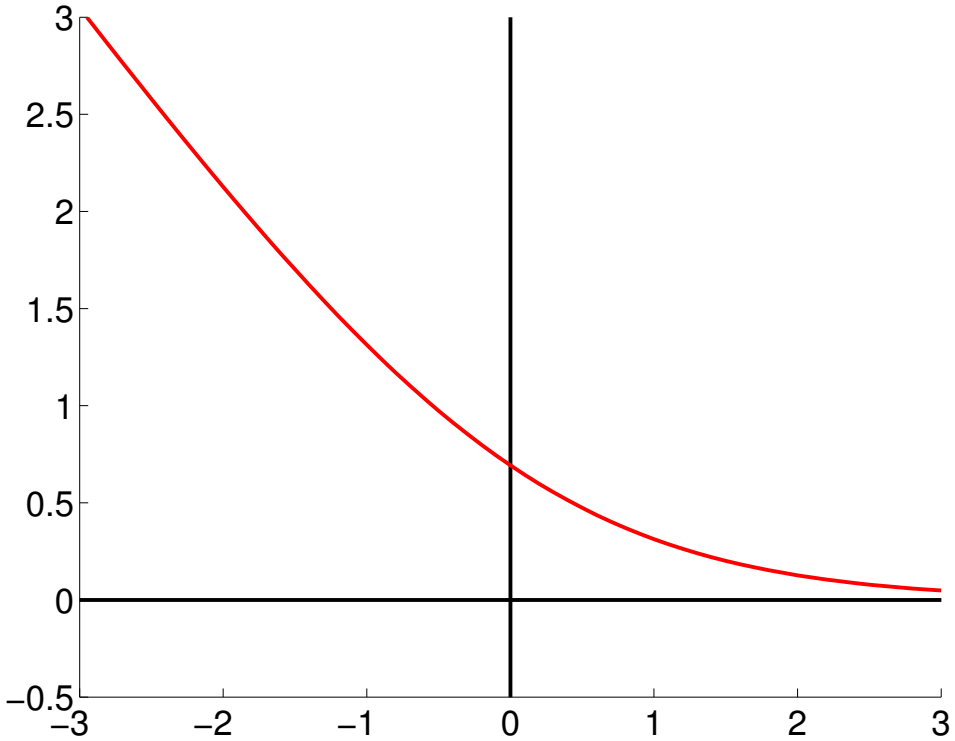
το οποίο για να παίρνει τιμές κοντά στο ελάχιστο (που είναι το 0) θα πρέπει

$$e^{-t_n \mathbf{w}^T \mathbf{x}_n}$$

να παίρνει τιμές κοντά στο μηδέν. Ακολούθως αυτό για να συμβεί θα πρέπει

$$\text{Αν} \quad t_n = 1 \Rightarrow \quad \mathbf{w}^T \mathbf{x}_n \gg 0$$

$$\text{Αν} \quad t_n = -1 \Rightarrow \quad \mathbf{w}^T \mathbf{x}_n \ll 0$$

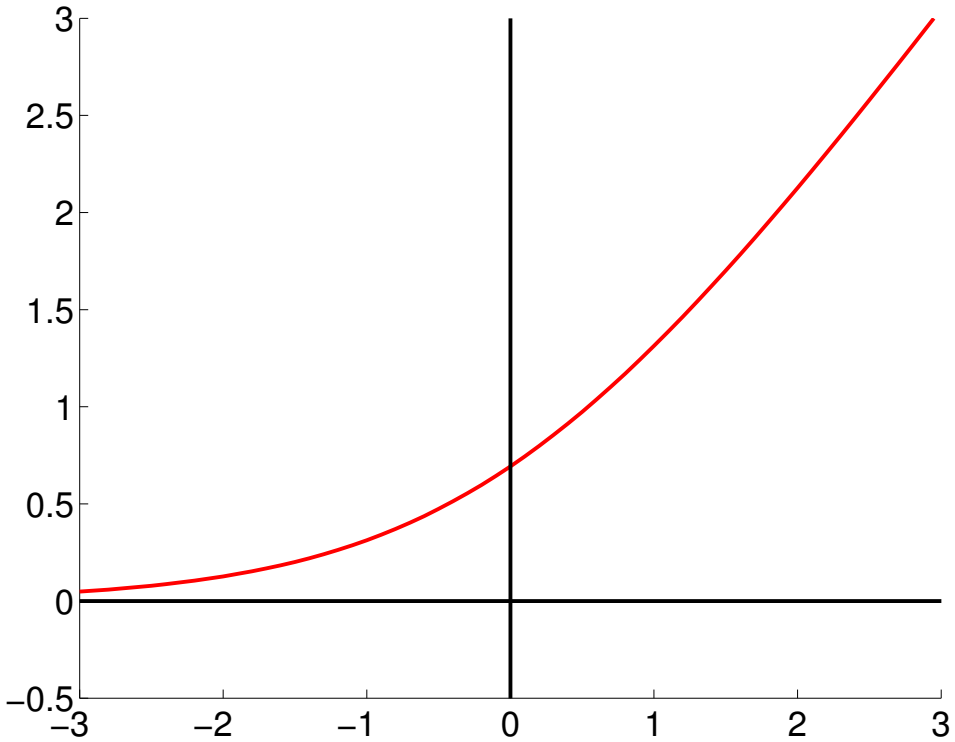


Σχήμα: Οριζόντιος άξονας αντιστοιχεί στο z και ο κάθετος στο $E(z)$.

Για δεδομένο της θετικής κατηγορίας (δηλ. $t = 1$) το κόστος γίνεται

$$E(z) = \log\left(1 + e^{-z}\right), \quad z = \mathbf{w}^T\mathbf{x}$$

και απεικονίζεται στο σχήμα ως συνάρτηση του z



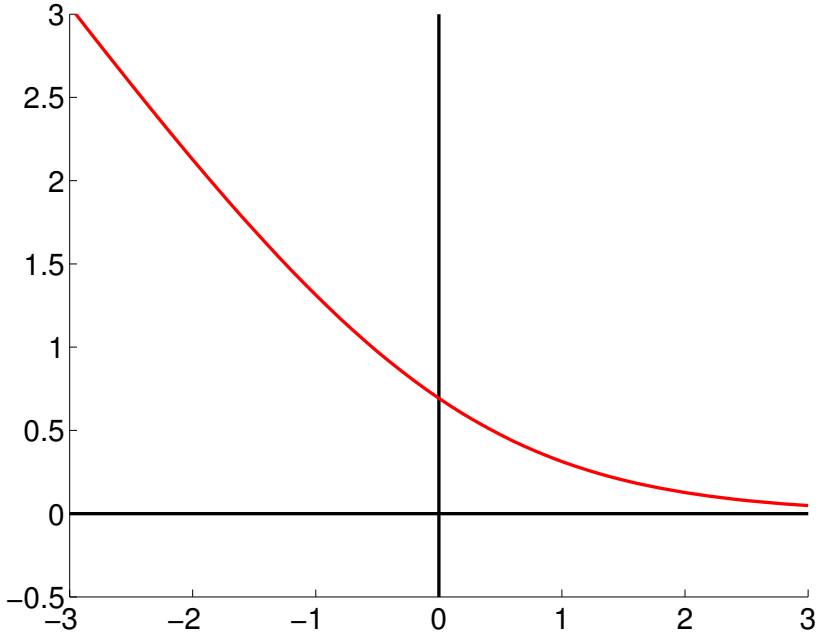
Σχήμα: Οριζόντιος άξονας αντιστοιχεί στο z και ο κάθετος στο $E(z)$.

Για δεδομένο της αρνητικής κατηγορίας (δηλ. $t = -1$) το κόστος γίνεται

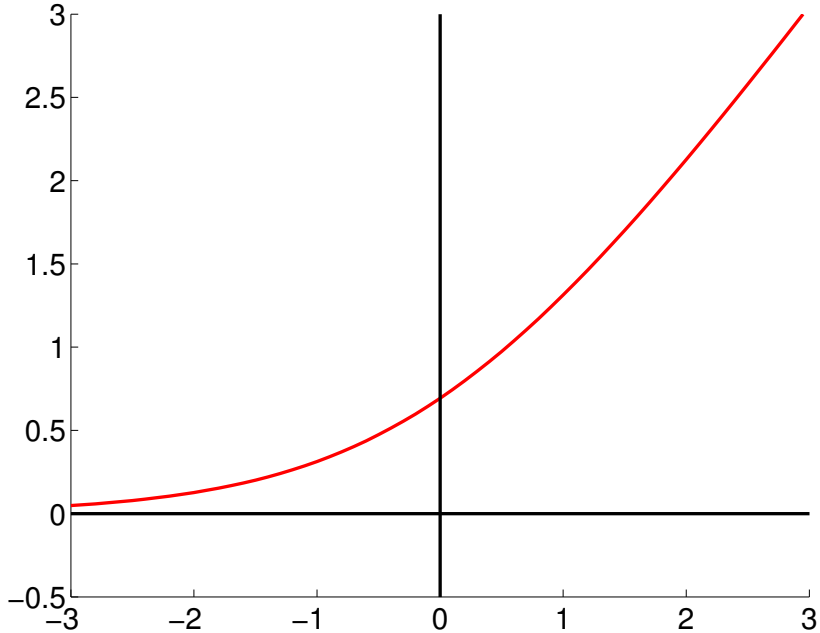
$$E(z) = \log\left(1 + e^{z}\right), \quad z = \mathbf{w}^{T}\mathbf{x}$$

και απεικονίζεται στο σχήμα ως συνάρτηση του z

Support Vector Machines



$t = 1$



$t = -1$

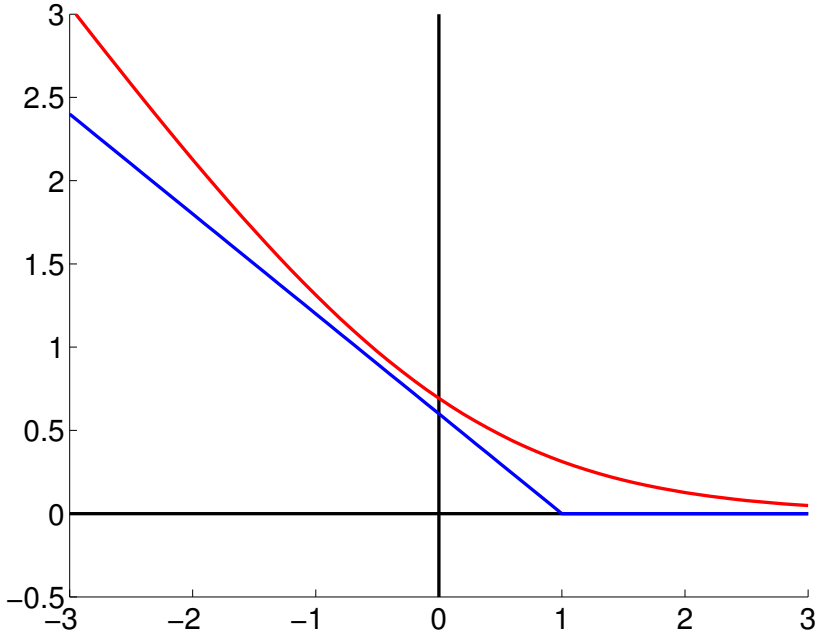
Η συνάρτηση κόστους συνεχίζει να τιμωρεί ακόμα και αν το δεδομένο έχει κατηγοριοποιηθεί σωστά κατά μεγάλο περιθώριο (**margin**)

Μια ιδέα θα ήταν να θέσουμε ένα περιθώριο ασφαλείας και να μην τιμωρούμε πέρα από αυτό. Π.χ.

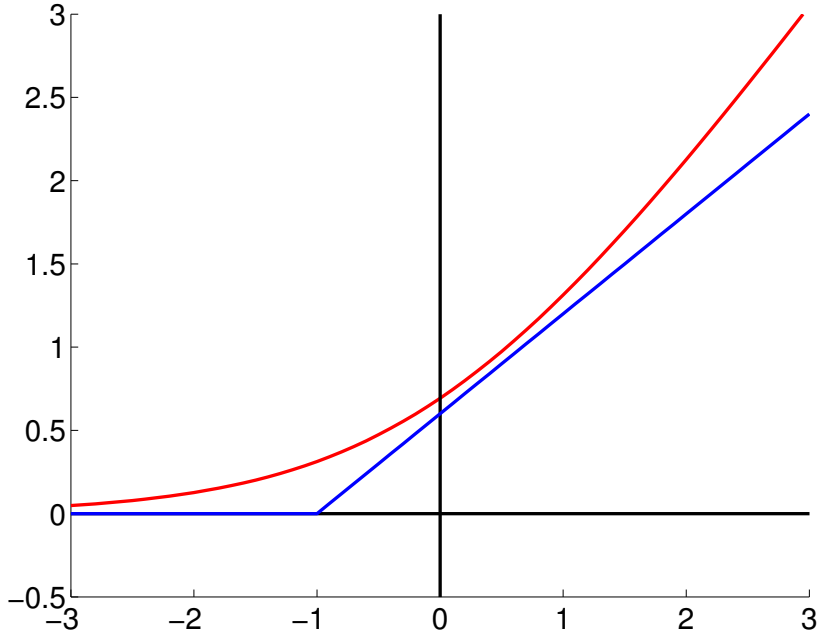
$$\text{Αν} \quad t_n = 1 \text{ και } \quad z = \mathbf{w}^T \mathbf{x}_n \geq 1 \quad \Rightarrow \text{σταμάτα να τιμωρείς!}$$

$$\text{Αν} \quad t_n = -1 \text{ και } \quad z = \mathbf{w}^T \mathbf{x}_n \leq -1 \quad \Rightarrow \text{σταμάτα να τιμωρείς!}$$

Support Vector Machines



$t = 1$



$t = -1$

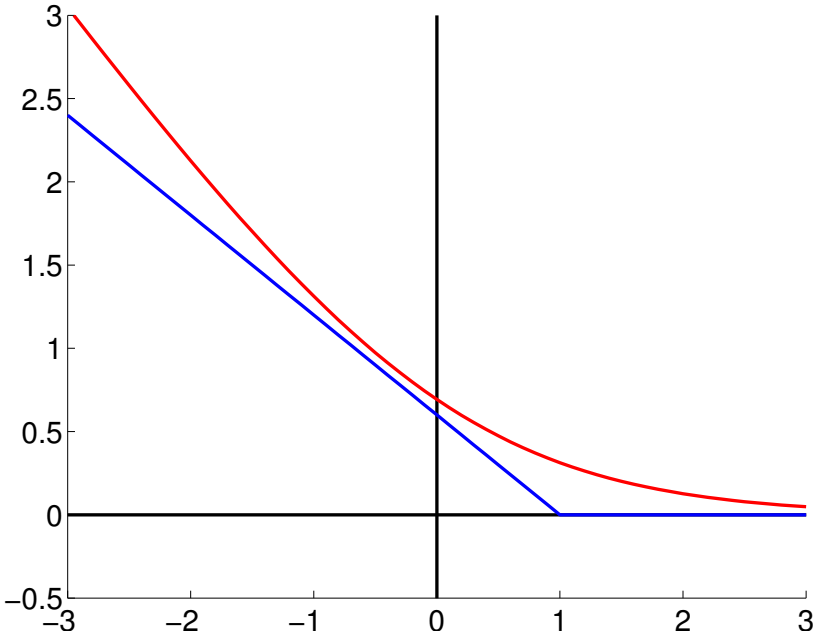
Μια ιδέα θα ήταν να θέσουμε ένα περιθώριο ασφαλείας

$$\text{Αν} \quad t_n = 1 \text{ και } \quad z = \mathbf{w}^T \mathbf{x}_n \geq 1 \quad \Rightarrow \text{σταμάτα να τιμωρείς!}$$

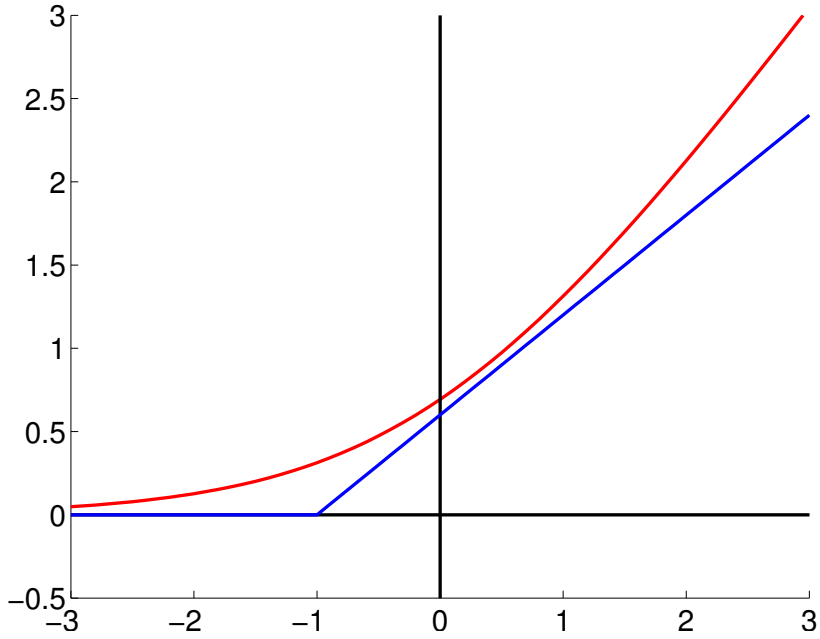
$$\text{Αν} \quad t_n = -1 \text{ και } \quad z = \mathbf{w}^T \mathbf{x}_n \leq -1 \quad \Rightarrow \text{σταμάτα να τιμωρείς!}$$

Αυτό θα μπορούσε να γίνει αλλάζοντας τις αρχικές συναρτήσεις κόστους που στο σχήμα φαίνονται με κόκκινο χρώμα με αυτές που φαίνονται με μπλε χρώμα

Support Vector Machines



$t = 1$



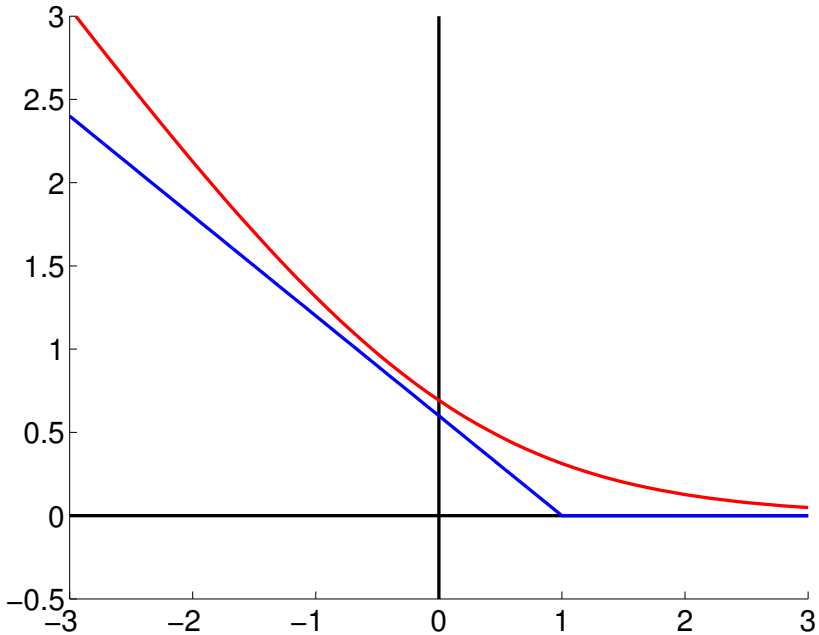
$t = -1$

Αν $t_n = 1$ χρησιμοποίησε κόστος $\max(0, 1 - z), \quad z = \mathbf{w}^T \mathbf{x}_n$

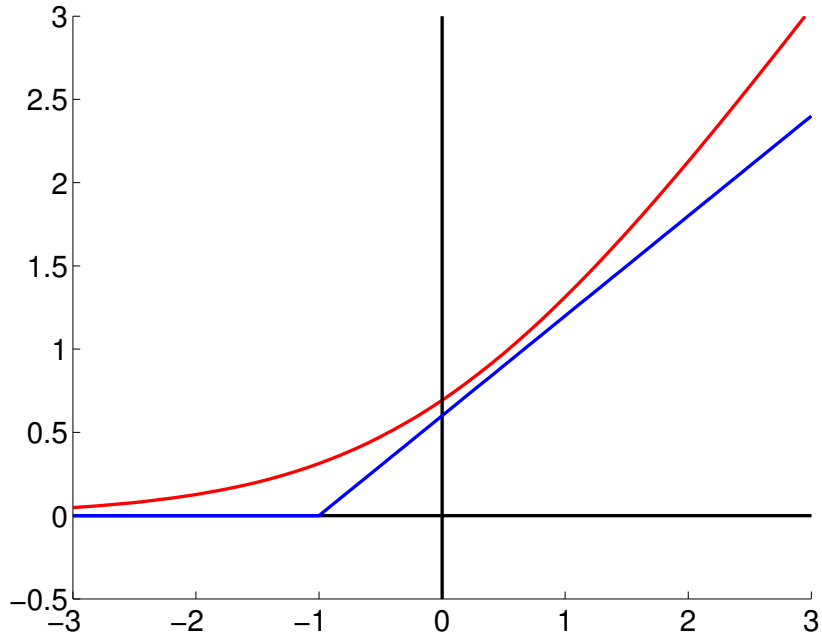
Αν $t_n = -1$ χρησιμοποίησε κόστος $\max(0, 1 + z), \quad z = \mathbf{w}^T \mathbf{x}_n$

- αυτές είναι οι συναρτήσεις κόστους (ανά) δεδομένο του αλγορίθμου των SVMs

Support Vector Machines



$t = 1$



$t = -1$

Με μία εξίσωση η συνάρτηση κόστους ανά δεδομένο για τον αλγόριθμο των support vector machines γράφεται ως

$$(1 - t_n \mathbf{w}^T \mathbf{x}_n)_+ = \max(0, 1 - t_n \mathbf{w}^T \mathbf{x}_n)$$

το οποίο ονομάζεται hinge loss και απεικονίζεται στο σχήμα με τις μπλε γραμμές

Συνοψίζοντας ως τώρα έχουμε

- Για την λογιστική παλινδρόμηση η συνάρτηση κόστους είναι

$$E(\mathbf{w}) = \sum_{n=1}^{N} \log\left(1 + e^{-t_n \mathbf{w}^T \mathbf{x}_n}\right) + \frac{1}{2}\lambda ||\mathbf{w}||^2$$

- Ο αλγόριθμος των support vector machines αλλάζει το κόστος $\log\left(1 + e^{-t_n \mathbf{w}^T \mathbf{x}_n}\right)$ με το hinge loss ώστε η συνολική συνάρτηση κόστους που ελαχιστοποιεί είναι η

$$E(\mathbf{w}) = \sum_{n=1}^{N} (1 - t_n \mathbf{w}^T \mathbf{x}_n)_+ + \frac{1}{2}\lambda ||\mathbf{w}||^2$$

όπου

$$(1 - t_n \mathbf{w}^T \mathbf{x}_n)_+ = \max(0, 1 - t_n \mathbf{w}^T \mathbf{x}_n)$$

Support Vector Machines

$$E(\mathbf{w}) = \sum_{n=1}^{N}(1 - t_n \mathbf{w}^T \mathbf{x}_n)_+ + \frac{1}{2}\lambda||\mathbf{w}||^2$$

Στη βιβλιογραφία των SVMs συνηθίζεται η παραπάνω συνάρτηση κόστους να γράφεται με ένα ισοδύναμο αλλά λίγο διαφορετικό τρόπο. Συγκεκριμένα, βγάζοντας κοινό παράγοντα το λ έχουμε

$$E(\mathbf{w}) = \lambda \left(\frac{1}{\lambda} \sum_{n=1}^{N}(1 - t_n \mathbf{w}^T \mathbf{x}_n)_+ + \frac{1}{2}||\mathbf{w}||^2 \right)$$

και ισοδύναμα μεγιστοποιούμε την ποσότητα

$$E(\mathbf{w}) = C \sum_{n=1}^{N}(1 - t_n \mathbf{w}^T \mathbf{x}_n)_+ + \frac{1}{2}||\mathbf{w}||^2$$

όπου $C = \frac{1}{\lambda}$ είναι η παράμετρος κανονικοποίησης

Support Vector Machines

Επίσης στην τελική συνάρτηση κόστους των SVMs δεν βάζουμε κανένα πέναλτυ στην παράμετρο του bias w_0, δηλ. η τελική συνάρτηση κόστους έχει τη μορφή

$$E(\mathbf{w}) = C\sum_{n=1}^{N}(1 - t_n(\mathbf{w}^T\mathbf{x}_n + w_0))_+ + \frac{1}{2}||\mathbf{w}||^2$$

όπου $\mathbf{x}_n$ θεωρείται ότι δεν έχει προσαυξηθεί με το 1 και $||\mathbf{w}||^2 = w_1^2 + \ldots + w_D^2$ (ενώ πριν ήταν $||\mathbf{w}||^2 = w_0^2 + w_1^2 + \ldots + w_D^2$)

Ως μια επλιπλέον αλλαγή συμβολισμού παρακάτω θα συμβολίζουμε το bias w_0 ως b προκειμένου να μην το συγχέουμε με τα υπόλοιπα $\mathbf{w}$. Δηλ. θα γράφουμε την συνάρτηση κόστους ως

$$E(\mathbf{w}) = C\sum_{n=1}^{N}(1 - t_n(\mathbf{w}^T\mathbf{x}_n + b))_+ + \frac{1}{2}||\mathbf{w}||^2$$

Support Vector Machines

$$E(\mathbf{w}) = C\sum_{n=1}^{N}(1 - t_n(\mathbf{w}^T\mathbf{x}_n + b))_+ + \frac{1}{2}||\mathbf{w}||^2$$

Λόγω ότι η $(1 - t_n(\mathbf{w}^T\mathbf{x}_n + b))_+ = \max(0, 1 - t_n(\mathbf{w}^T\mathbf{x}_n + b))$ δεν είναι παραγωγίσιμη (συγκεκριμένα εξαιτίας του max), δεν μπορούμε να την μεγιστοποιήσουμε χρησιμοποιώντας ένα αλγόριθμο όπως αυτός της ανοδικής κλίσης (δηλ. όπως στην περίπτωση της συνάρτησης κόστους της λογιστικής παλινδρόμησης)

Ωστόσο μπορεί να δειχθεί ότι η ελαχιστοποίηση της παραπάνω συνάρτησης κόστους είναι ισοδύναμη με το ακόλουθο πρόβλημα βελιστοποίησης

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} C\sum_{n=1}^{N}\xi_n + \frac{1}{2}||\mathbf{w}||^2$$

$$\text{subject to: } \xi_n \geq 0, \quad t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n, \ n = 1, \ldots, N$$

Support Vector Machines

$$E(\mathbf{w}) = C\sum_{n=1}^{N}(1 - t_n(\mathbf{w}^T\mathbf{x}_n + b))_+ + \frac{1}{2}||\mathbf{w}||^2$$

Η ελαχιστοποίηση της παραπάνω συνάρτησης κόστους μπορεί ισοδύναμα να διατυπωθεί ως το ακόλουθο πρόβλημα βελιστοποίησης με γραμμικούς περιορισμούς

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} C\sum_{n=1}^{N}\xi_n + \frac{1}{2}||\mathbf{w}||^2$$

$$\text{subject to: } \xi_n \geq 0, \quad t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1 - \xi_n, \ n = 1, \ldots, N$$

Το πρόβλημα αυτό ονομάζεται quadratic program και έχει μοναδική λύση η οποία μπορεί να βρεθεί από διάφορους αλγορίθμους

- δεν θα αναλύσουμε τέτοιους αλγόριθμους και ούτε το πως προκύπτει το ισοδύναμο quadratic program
- Ωστόσο θα προσπαθήσουμε διαισθητικά να καταλάβουμε το τι συμβαίνει στην περίπτωση που τα δεδομένα είναι γραμμικά διαχωρίσιμα και έπειτα στην περίπτωση που δεν είναι γραμμικά διαχωρίσιμα

Support Vector Machines

$$E(\mathbf{w}) = C\sum_{n=1}^{N}(1 - t_n(\mathbf{w}^T\mathbf{x}_n + b))_+ + \frac{1}{2}||\mathbf{w}||^2$$

ή ισοδύναμα

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} C\sum_{n=1}^{N}\xi_n + \frac{1}{2}||\mathbf{w}||^2$$

$$\text{subject to: } \xi_n \geq 0, \quad t_n(\mathbf{w}^T\mathbf{x}_n + w_0) \geq 1 - \xi_n, \ n = 1, \ldots, N$$

Στην περίπτωση που τα δεδομένα είναι γραμμικά διαχωρίσιμα, και επιλέξουμε σημαντικά μεγάλη τιμή για το C (κοντά στο συν άπειρο) τα ξ_n θα γίνουν μηδέν, δηλ. $\xi_n = 0, n = 1, \ldots, N$. Οπότε σε αυτή την περίπτωση το πρόβλημα βελιστοποίησης απλοποείται στη μορφή

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$$

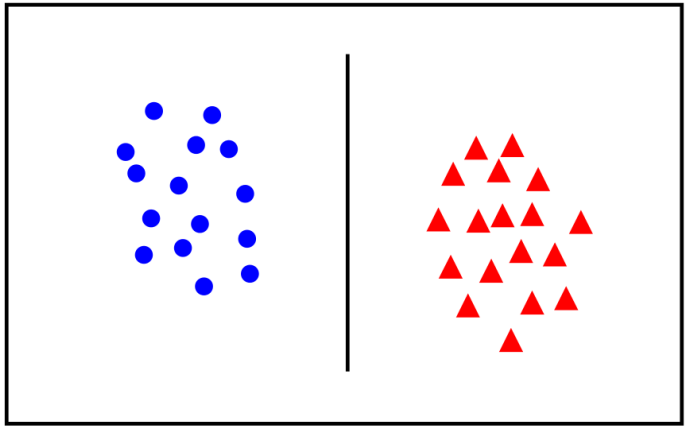
$$\text{subject to: } \quad t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, \ n = 1, \ldots, N$$

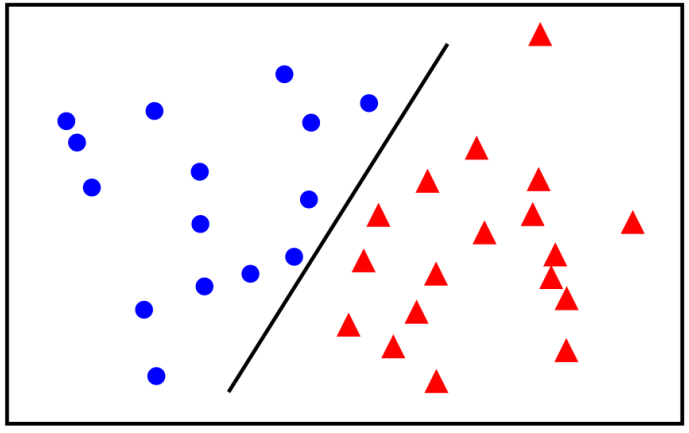
το οποίο έχει πάντα λύση εφόσον τα δεδομένα είναι γραμμικά διαχωρίσιμα

Support Vector Machines

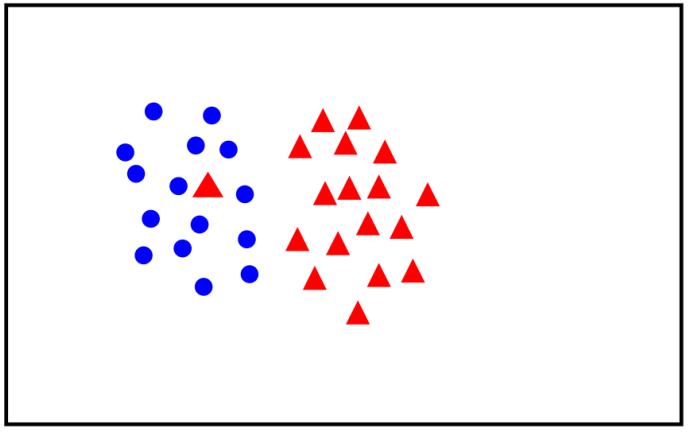
(Υπενθύμιση: για το τι είναι γραμμικά διαχωρίσιμα δεδομένα δύο κατηγοριών)

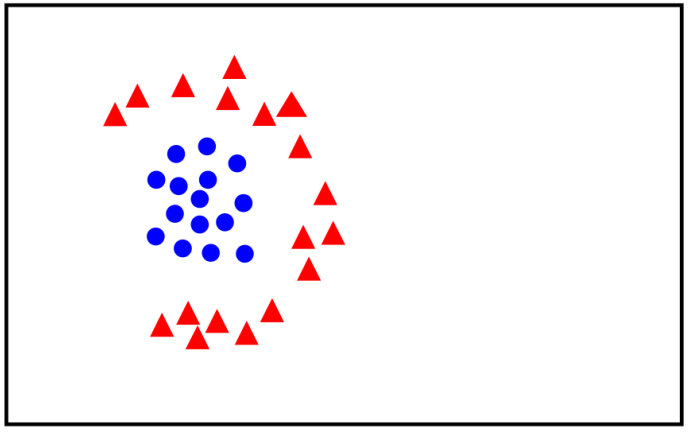
linearly separable





not linearly separable





Support Vector Machines

(Υπενθύμιση: για το τι είναι γραμμικά διαχωρίσιμα δεδομένα δύο κατηγοριών)

Μαθηματικά γραμμικά διαχωρίσιμα δεδομένα (στο σύνολο εκπαίδευσης) σημαίνει ότι υπάρχουν $(\mathbf{w}, b)$ τέτοια ώστε

$$t_n(w^T\mathbf{x} + b) > 0, \quad n = 1, \ldots, N$$

Support Vector Machines

Για γραμμικά διαχώσιμα δεδομένα το ακόλουθο πρόβλημα έχει μοναδική λύση

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$$

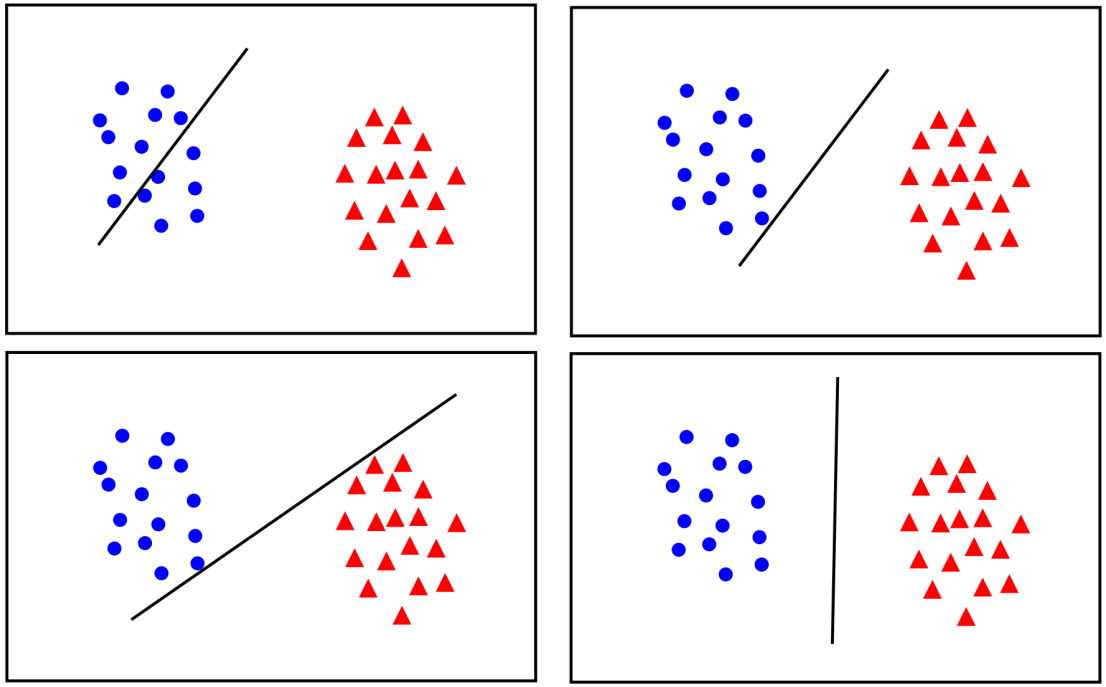
$$\text{subject to:} \quad t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, \ n = 1, \ldots, N$$

και βρίσκει ένα σύνορο απόφασης που μεγιστοποιεί το margin

Προκειμένου να κατανοήσουμε τα SVMs θα πρέπει να αναλύσουμε τα εξής

1. τι είναι το margin;
2. γιατί/πώς ο αλγόριθμος των SVMs το μεγιστοποίει;

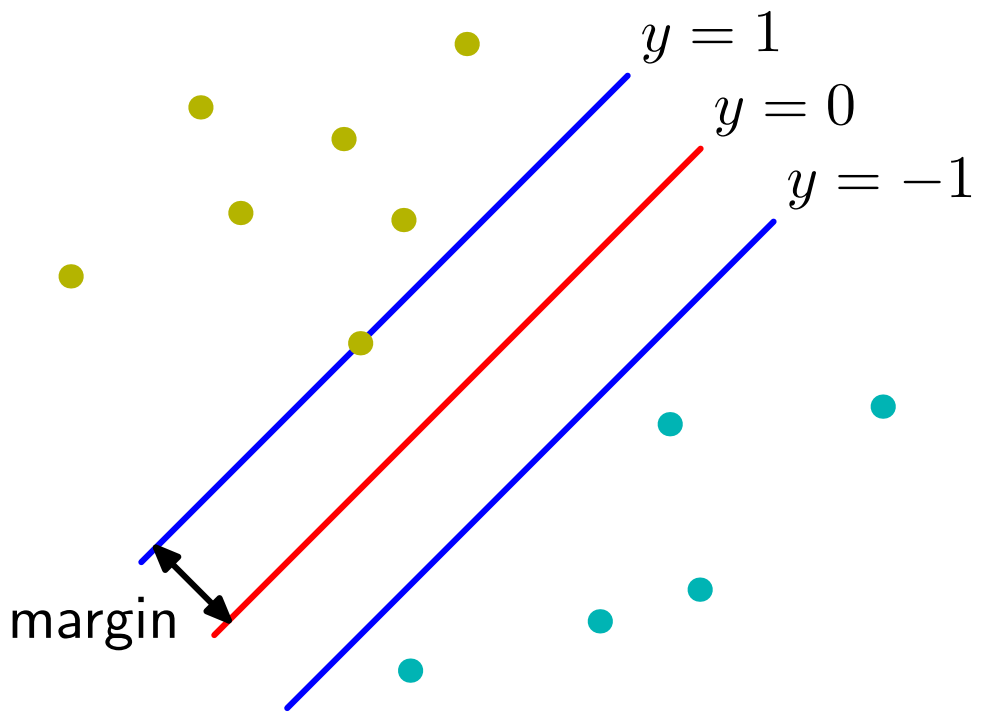
Γραμμικά διαχωρισίμα δεδομένα και η έννοια του margin



Όταν τα δεδομένα είναι γραμμικά διαχωρίσιμα, τότε υπάρχουν άπειρες γραμμές που τα διαχωρίζουν

Τα (SVMs) θα βρούν μια συγκεκριμένη γραμμή που στο σχήμα θα είναι η κάτω δεξιά $\Rightarrow$ είναι η γραμμή που μεγιστοποιεί το margin

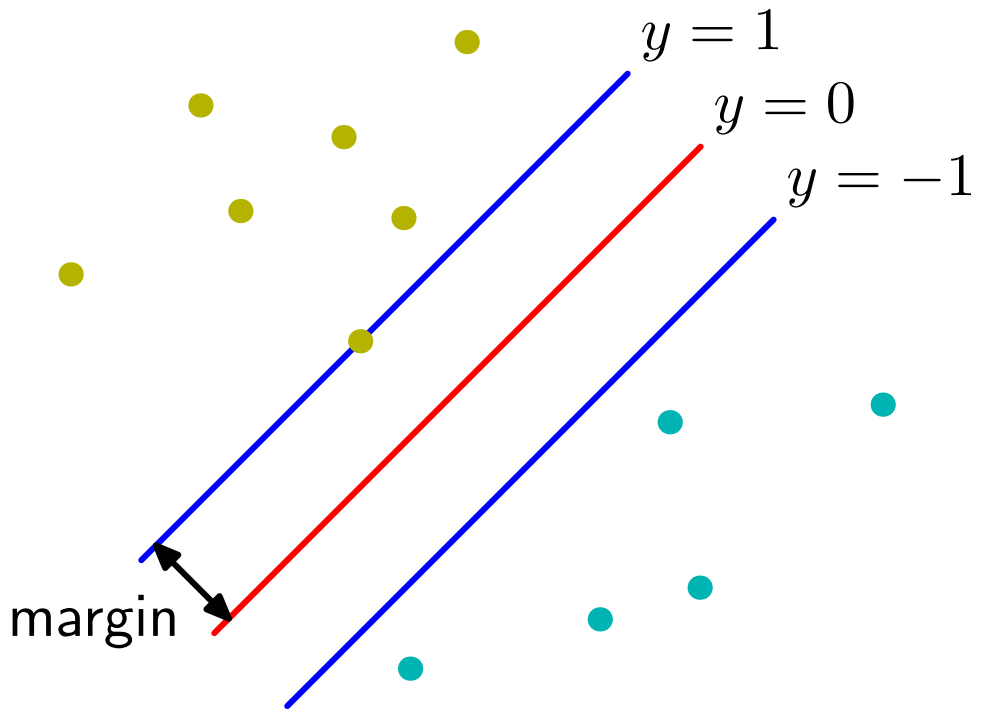
Γραμμικά διαχωρισίμα δεδομένα και η έννοια του margin



Margin: Ορίζεται ως η μικρότερη απόσταση ανάμεσα στο σύνορο απόφασης και κάθε δεδομένο εκπαίδευσης

Η έννοια του margin χαρακτηρίζει οποιοδήποτε σύνορο απόφασης σε κάθε σύστημα κατηγοριοποίησης και δεν είναι κάτι που εμφανίζεται μόνο στα SVMs!

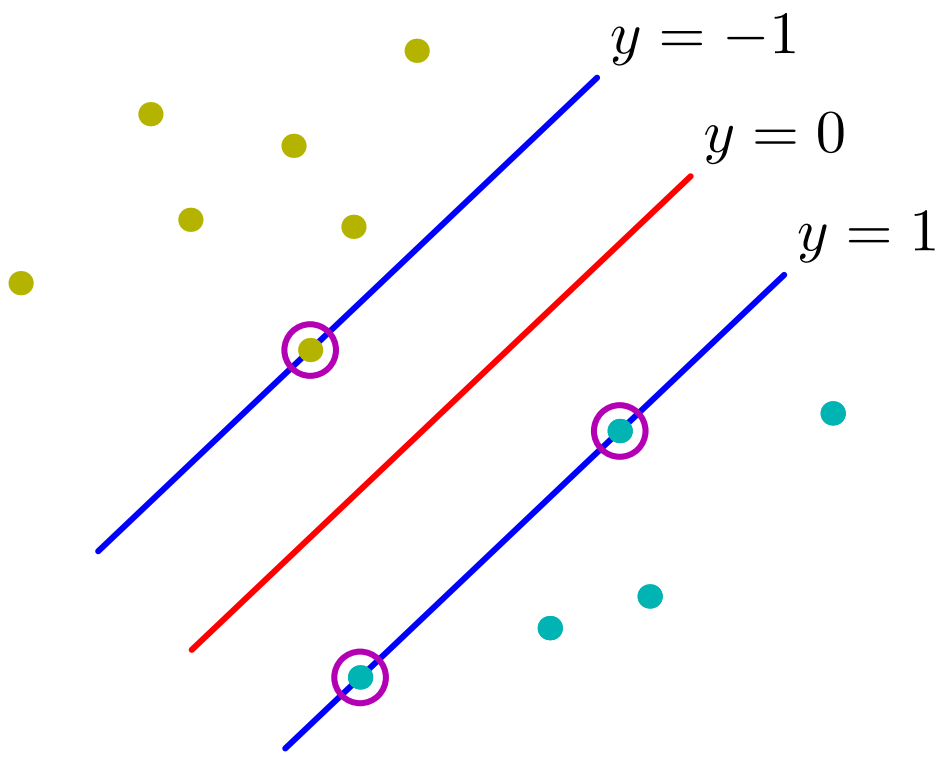
Γραμμικά διαχωρισίμα δεδομένα και η έννοια του margin



Margin: Ορίζεται ως η μικρότερη απόσταση ανάμεσα στο σύνορο απόφασης και κάθε δεδομένο εκπαίδευσης

Τα SVMs επιλέγουν εκείνο το σύνορο απόφασης που **μεγιστοποιεί** το margin

Γραμμικά διαχωρισίμα δεδομένα και η έννοια του margin



Η μορφή του συνόρου απόφασης (δηλ. οι τιμές των παραμέτρων $(\mathbf{w}, b)$) καθορίζεται από ένα υποσύνολο των δεδομένων (που ικανοποιούν το margin)

- $\Rightarrow$ ονομάζονται support vectors

Γραμμικά διαχωρισίμα δεδομένα και η έννοια του margin

Μορφή της λύσης για τις παραμέτρους w και b

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$$

$$\text{subject to:} \quad t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, \ n = 1, \ldots, N$$

Αφού βρούμε την μοναδική λύση του παραπάνω προβλήματος, μπορεί να δειχθεί ότι οι παράμετροι **w** θα δίνονται ως γραμμικός συνδυασμός των support vectors. Έστω το υποσύνολο των δεδομένων που είναι support vectors είναι το $\mathcal{S}$. Όλα αυτά τα δεδομένα ικανοποιούν το margin (δηλ. έχουν όλα την ίδια ελάχιστη απόσταση από το σύνορο απόφασης). Το **w** γράφεται ως

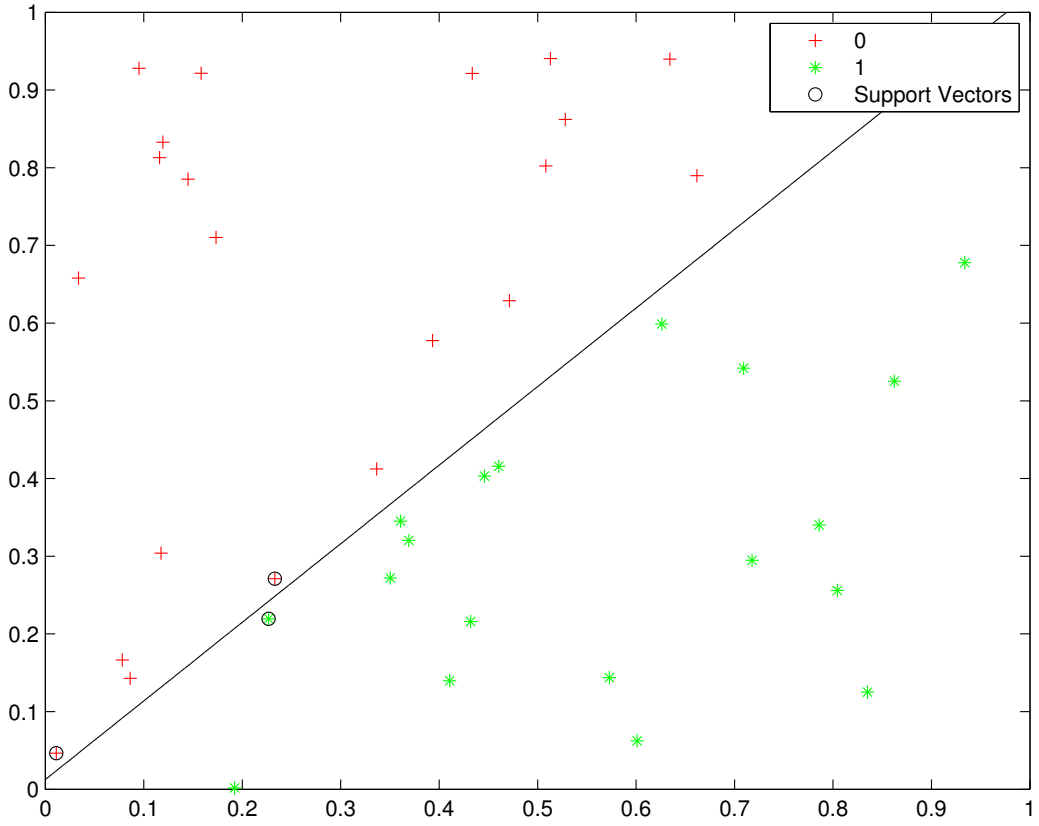
$$\mathbf{w} = \sum_{n \in \mathcal{S}} t_n \alpha_n \mathbf{x}_n$$

όπου η τιμή του κάθε $\alpha_n > 0$ προκύπτει από την λύση του παραπάνω προβλήματος βελτιστοποίησης

Ομοίως και η βέλτιστη λύση για το bias b εξαρτάται μόνο από τα support vectors

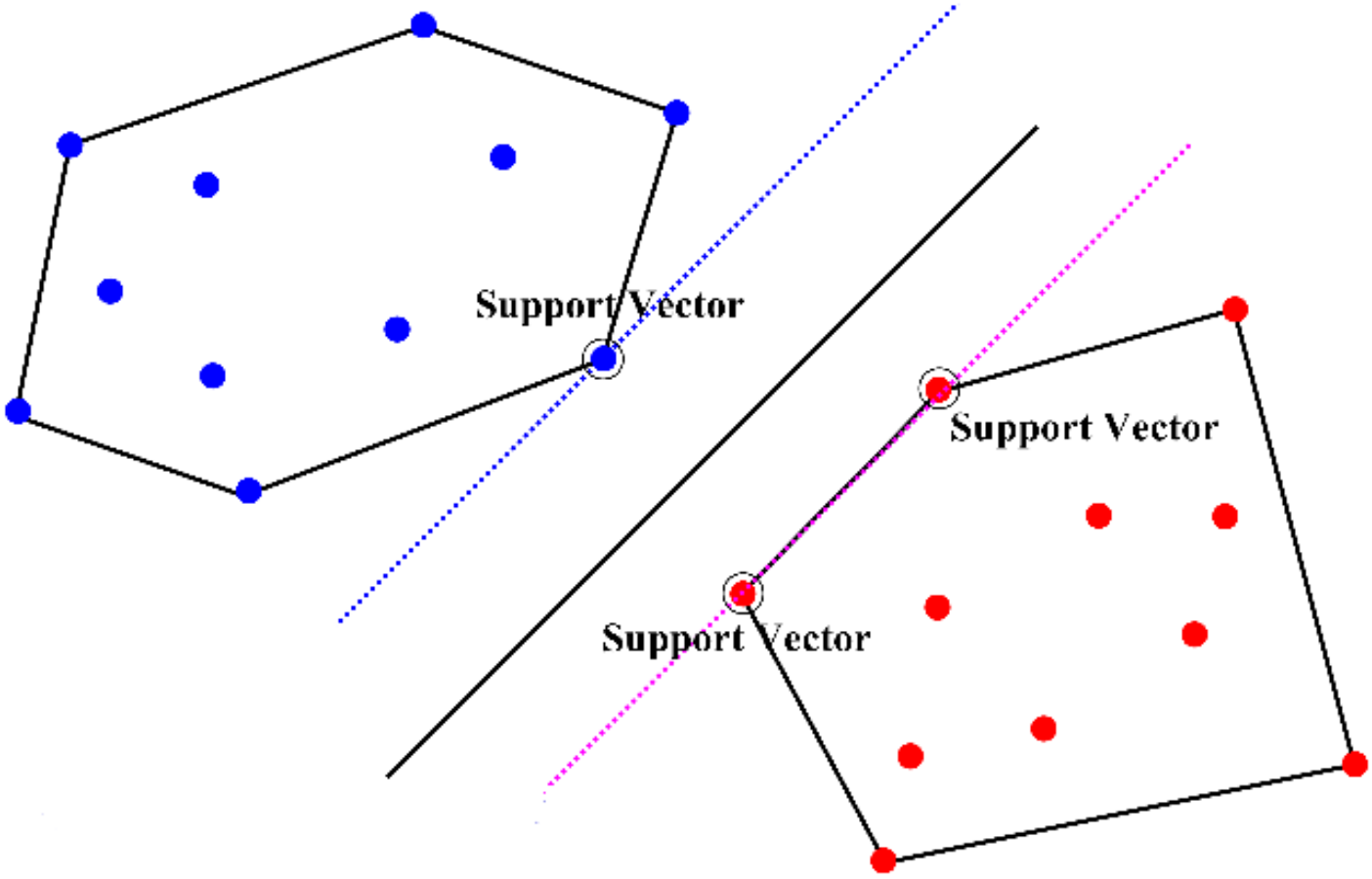
Γραμμικά διαχωρισίμα δεδομένα και η έννοια του margin

Παράδειγμα εφαρμογής στο σύνολο data2Tr. Δες demo_svmSepar.m στο e-class



Το σύνορο απόφασης (δηλ. οι τιμές των παραμέτρων $(\mathbf{w}, b)$) καθορίζεται μόνο από τρία δεδομένα (δηλ. υπάρχουν μόνο τρία support vectors)

Γραμμικά διαχωρισίμα δεδομένα και η έννοια του margin



Τα εσωτερικά δεδομένα δεν πρόκειται ποτέ να γίνουν support vectors. Τα support vectors επιλέγονται από το περίγραμμα (convex hull) των δεδομένων

Γραμμικά διαχωρισίμα δεδομένα και η έννοια του margin

Γιατί ο αλγόριθμος των SVMs μεγιστοποιεί το margin;

Για να το θέσουμε διαφορεικά, γιατί για γραμμικά διαχωρίσιμα δεδομένα η μοναδική λύση του προβλήματος

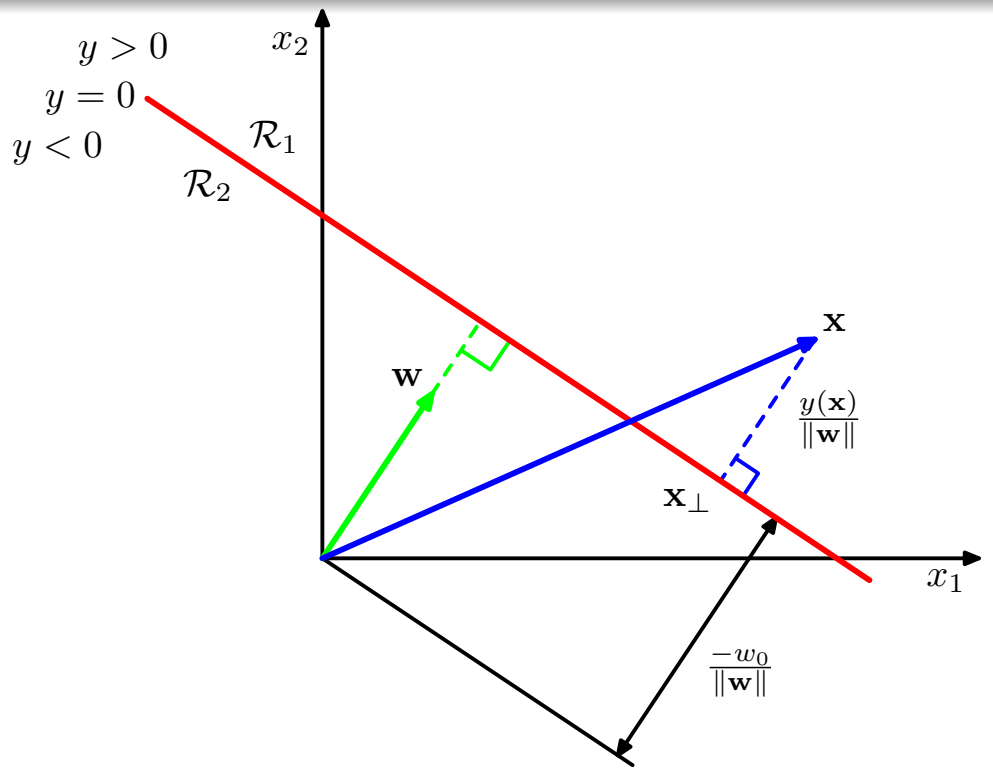
$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$$

$$\text{subject to:} \quad t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, \;\; n = 1, \ldots, N$$

αποτελεί ένα σύνορο απόφαση που μεγιστοποιεί το margin;

Για να απαντήσουμε στο ερώτημα αυτό θα ακολουθήσουμε μια κατασκευαστική απόδειξη. Θα ορίσουμε απευθείας την συνάρτηση που αντιστοιχεί στην μεγιστοποίηση του margin και έπειτα θα διαπιστώσουμε ότι το πρόβλημα βελιστοποίησης που προκύπτει είναι ισοδύναμο με το παραπάνω

Γιατί ο αλγόριθμος των SVMs μεγιστοποιεί το margin;



- Ισχύει: Το κάθε $\mathbf{x}$ απέχει από το σύνορο απόφασης απόσταση

$$||\mathbf{x} - \mathbf{x}_\perp|| = \frac{|y(\mathbf{x})|}{||w||}, \quad \text{όπου } \mathbf{x}_\perp \text{ η προβολή του } \mathbf{x} \text{ στο σύνορο}$$

- $\frac{y(\mathbf{x})}{||\mathbf{w}||}$ (χωρίς απόλυτη τιμή) μας δίνει επιπλέον την πληροφορία για τη πλευρά του συνόρου που περιέχει το $\mathbf{x}$

(για αναλυτική εξήγηση δες στο τέλος Παράρτημα: Γεωμετρία συνόρων απόφασης)

Γιατί ο αλγόριθμος των SVMs μεγιστοποιεί το margin;

Μεγιστοποίηση του margin

- Δοθέντος ότι το σύνορο απόφασης διαχωρίζει γραμμικά τα δεδομένα, η απόσταση του κάθε δεδομένου από το σύνορο (σε απόλυτη τιμή) είναι

$$\frac{|y(\mathbf{x}_n)|}{||\mathbf{w}||} = \frac{t_n y(\mathbf{x}_n)}{||\mathbf{w}||} = \frac{t_n(\mathbf{w}^T\mathbf{x}_n + b)}{||\mathbf{w}||}$$

- Εξ ορισμού το margin είναι η ελάχιστη τέτοια απόσταση, δηλ.

$$\text{margin}(\mathbf{w}, b) = \min_n \left[\frac{t_n(\mathbf{w}^T\mathbf{x}_n + b)}{||\mathbf{w}||}\right] = \frac{1}{||\mathbf{w}||} \min_n \left[t_n(\mathbf{w}^T\mathbf{x}_n + b)\right]$$

- Επομένως η μεγιστοποίηση του margin ισοδυναμεί με την λύση του ακόλουθου προβλήματος βελτιστοποίησης

$$\max_{\mathbf{w},b} \left\{\frac{1}{||\mathbf{w}||} \min_n \left[t_n(\mathbf{w}^T\mathbf{x}_n + b)\right]\right\}$$

Γιατί ο αλγόριθμος των SVMs μεγιστοποιεί το margin;

Μεγιστοποίηση του margin

$$\max_{\mathbf{w},b} \left\{ \frac{1}{||\mathbf{w}||} \min_n \left[t_n(\mathbf{w}^T \mathbf{x}_n + b) \right] \right\}$$

- Η λύση του παραπάνω προβλήματος φαντάζει δύσκολη με βάση την τρέχουσα αναπαράσταση
- Θα θέλαμε να μετασχηματίσουμε το πρόβλημα σε μια ισοδύναμη μορφή που να οδηγεί σε ευκολότερη επίλυση
- Πρώτα από όλα παρατηρούμε ότι για $k > 0$ αν κάνουμε rescaling τις παραμέτρους, $\mathbf{w} \to k\mathbf{w}$ και $b \to kb$, η απόσταση των δεδομένων από το σύνορο απόφασης δεν μεταβάλλεται, δηλ.

$$\frac{t_n(k\mathbf{w}^T \mathbf{x}_n + kb)}{||k\mathbf{w}||} = \frac{kt_n(\mathbf{w}^T \mathbf{x}_n + b)}{k||\mathbf{w}||} = \frac{t_n(\mathbf{w}^T \mathbf{x}_n + b)}{||\mathbf{w}||}$$

Γιατί ο αλγόριθμος των SVMs μεγιστοποιεί το margin;

Μεγιστοποίηση του margin

- Για $k > 0$ αν κάνουμε rescaling τις παραμέτρους $\mathbf{w} \to k\mathbf{w}$ και $b \to kb$, η απόσταση των δεδομένων από το σύνορο απόφασης δεν μεταβάλλεται

$$\frac{t_n(k\mathbf{w}^T\mathbf{x}_n + kb)}{||k\mathbf{w}||} = \frac{t_n(\mathbf{w}^T\mathbf{x}_n + b)}{||\mathbf{w}||}$$

- Οπότε θα μπορούσαμε να επιλέξουμε ένα τέτοιο rescaled $(\mathbf{w}, b)$ ώστε

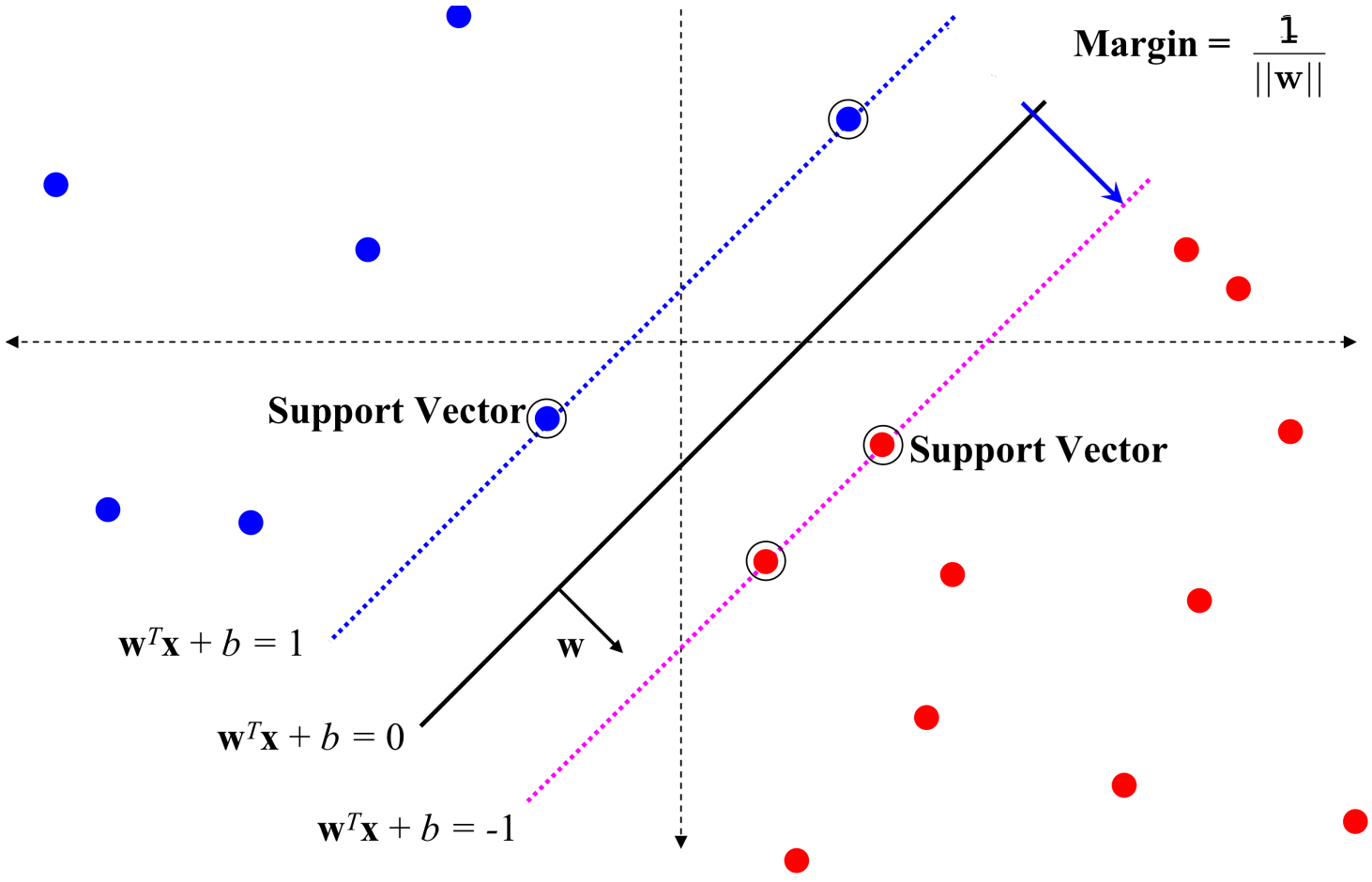
$$t_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$$

για όλα τα δεδομένα που ικανοποιούν το margin. Οπότε όλα τα δεδομένα γενικά θα ικανοποιούν

$$t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, \quad n = 1, \ldots, N$$

και το margin θα είναι ίσο με $\frac{1}{||\mathbf{w}||}$!

Γιατί ο αλγόριθμος των SVMs μεγιστοποιεί το margin;



Γιατί ο αλγόριθμος των SVMs μεγιστοποιεί το margin;

Μεγιστοποίηση του margin

$$\max_{\mathbf{w},b} \left\{ \frac{1}{||\mathbf{w}||} \min_n \left[t_n(\mathbf{w}^T \mathbf{x}_n + b) \right] \right\}$$

- Όλα τα δεδομένα γενικά θα ικανοποιούν

$$t_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1, \quad n = 1, \ldots, N$$

- Αφού $\min_n \left[t_n(\mathbf{w}^T \mathbf{x}_n + b) \right] = 1$, το πρόβλημα βελτιστοποίησης μετασχηματίζεται σε

$$\max_{\mathbf{w}} \left\{ \frac{1}{||\mathbf{w}||} \right\} \Rightarrow \min_{\mathbf{w}} \frac{||\mathbf{w}||^2}{2}$$

υπό τον περιορισμό ότι

$$t_n(\mathbf{w}^T \mathbf{x}_n + b) \geq 1, \quad n = 1, \ldots, N$$

Γιατί ο αλγόριθμος των SVMs μεγιστοποιεί το margin;

Μεγιστοποίηση του margin
Οποτέ οδηγούμαστε στο πρόβλημα βελτιστοποίησης

$$\min_{\mathbf{w}} \frac{||\mathbf{w}||^2}{2}$$

υπό τον περιορισμό ότι

$$t_n(\mathbf{w}^T\mathbf{x}_n + b) \geq 1, \quad n = 1, \ldots, N$$

- που είναι το πρόβλημα βελιστοποίησης των SVMs που είχαμε ορίσει αρχικά για γραμμικά διαχωρίσιμα δεδομένα
- $\Rightarrow$ οπότε καταλήγουμε στο συμπέρασμα ότι ο αλγόριθμος των SVMs πράγματι μεγιστοποιεί το margin!

Γραμμική διαχωρισιμότητα στο χώρο των μετασχηματισμένων δεδομένων

Όπως και στην περίπτωση της λογιστικής παλινδρόμησης θα μπορούσαμε να μετασχηματίσουμε κάθε δεδομένο από $\mathbf{x}$ σε $\boldsymbol{\phi} = \boldsymbol{\phi}(\mathbf{x})$

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=1}^{M} w_j \phi_j + b$$

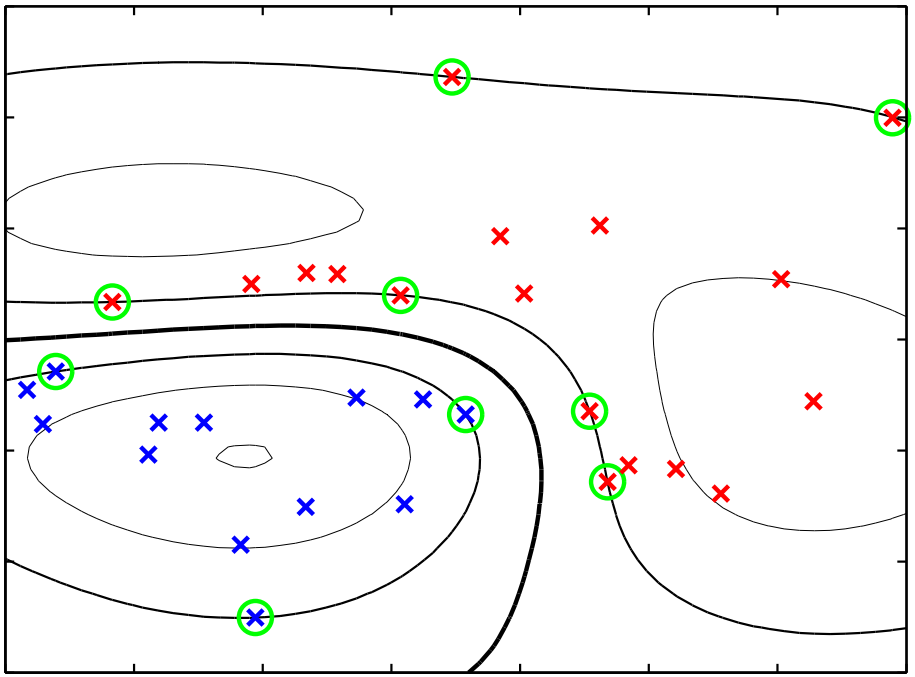
όπου $\boldsymbol{\phi} = (\phi_1(\mathbf{x}), \ldots, \phi_M(\mathbf{x}))$ είναι το feature vector

Ο αλγόριθμος των SVMs χρησιμοποιείται προφανώς χωρίς καμιά αλλαγή

Η έννοια του margin θα αφορά αποστάσεις των μετασχηματισμένων δεδομένων με το σύνορο απόφασης. Επίσης το σύνορο απόφασης θα είναι μη γραμμικό ως προς το $\mathbf{x}$ αλλά γραμμικό ως προς $\boldsymbol{\phi}$

Γραμμική διαχωρισιμότητα στο χώρο των μετασχηματισμένων δεδομένων

Παράδειγμα



Μη γραμμικά διαχωρισίμα δεδομένα: κατηγορίες με επικάλυψη

Στη πράξη ακόμα και αν τα δεδομένα είναι γραμμικά διαχωρίσιμα (σε κάποιο χώρο!) εκπαιδεύουμε τα SVMs επιτρέποντας λάθη στην κατηγοριοποίηση των δεδομένων του συνόλου εκπαίδευσης. Δηλ. χρησιμοποιούμε την αρχική μορφή του προβλήματος βελτιστοποίησης όπου οι slack variables ξ_n δεν είναι μηδέν. Δηλ. λύνουμε το πρόβλημα

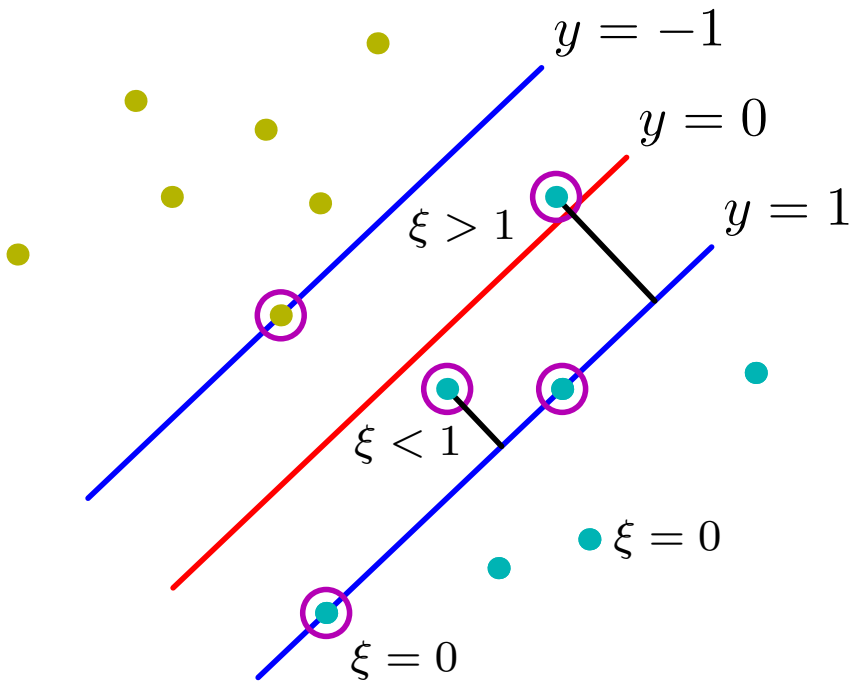
$$\min_{\mathbf{w},b,\boldsymbol{\xi}} C\sum_{n=1}^{N}\xi_n + \frac{1}{2}||\mathbf{w}||^2$$

$$\text{subject to: } \xi_n \geq 0, \quad t_n(\mathbf{w}^T\mathbf{x}_n + w_0) \geq 1 - \xi_n, \ n = 1, \ldots, N$$

Ερμηνεία της κάθε slack variable:

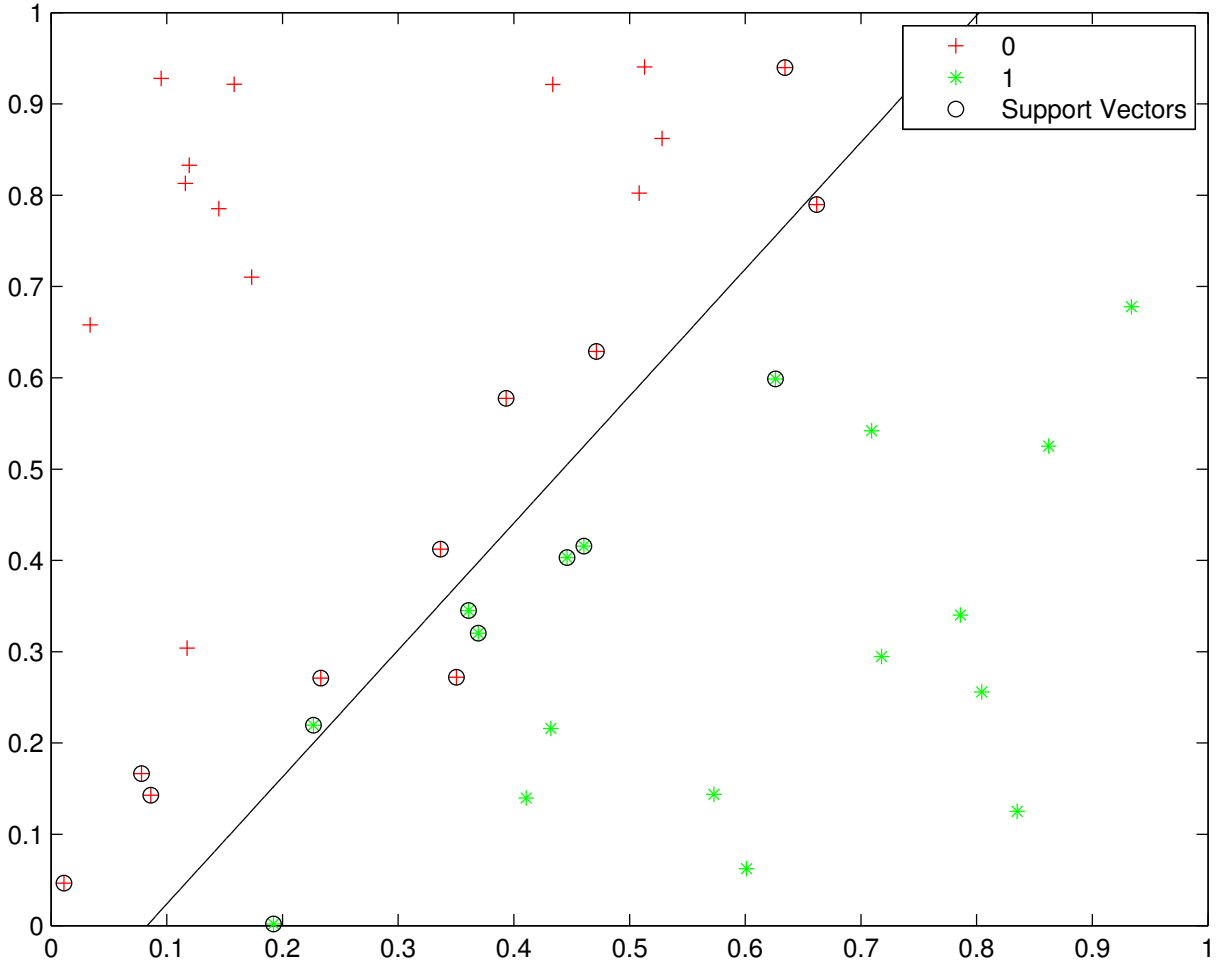
- $\xi_n = 0$: Δεδομένο που είναι σωστά κατηγοριοποιημένο και είτε ικανοποιεί ακριβώς το margin είτε βρίσκεται στη σωστή πλευρά
- $0 < \xi_n \leq 1$: Δεδομένο που είναι σωστά κατηγοριοποιημένο αλλά βρίσκεται μεταξύ του συνόρου απόφασης και του margin
- $\xi_n > 1$: Δεδομένο δεν κατηγοριοποιείται σωστά

Μη γραμμικά διαχωρισίμα δεδομένα: κατηγορίες με επικάλυψη



Μη γραμμικά διαχωρισίμα δεδομένα: κατηγορίες με επικάλυψη

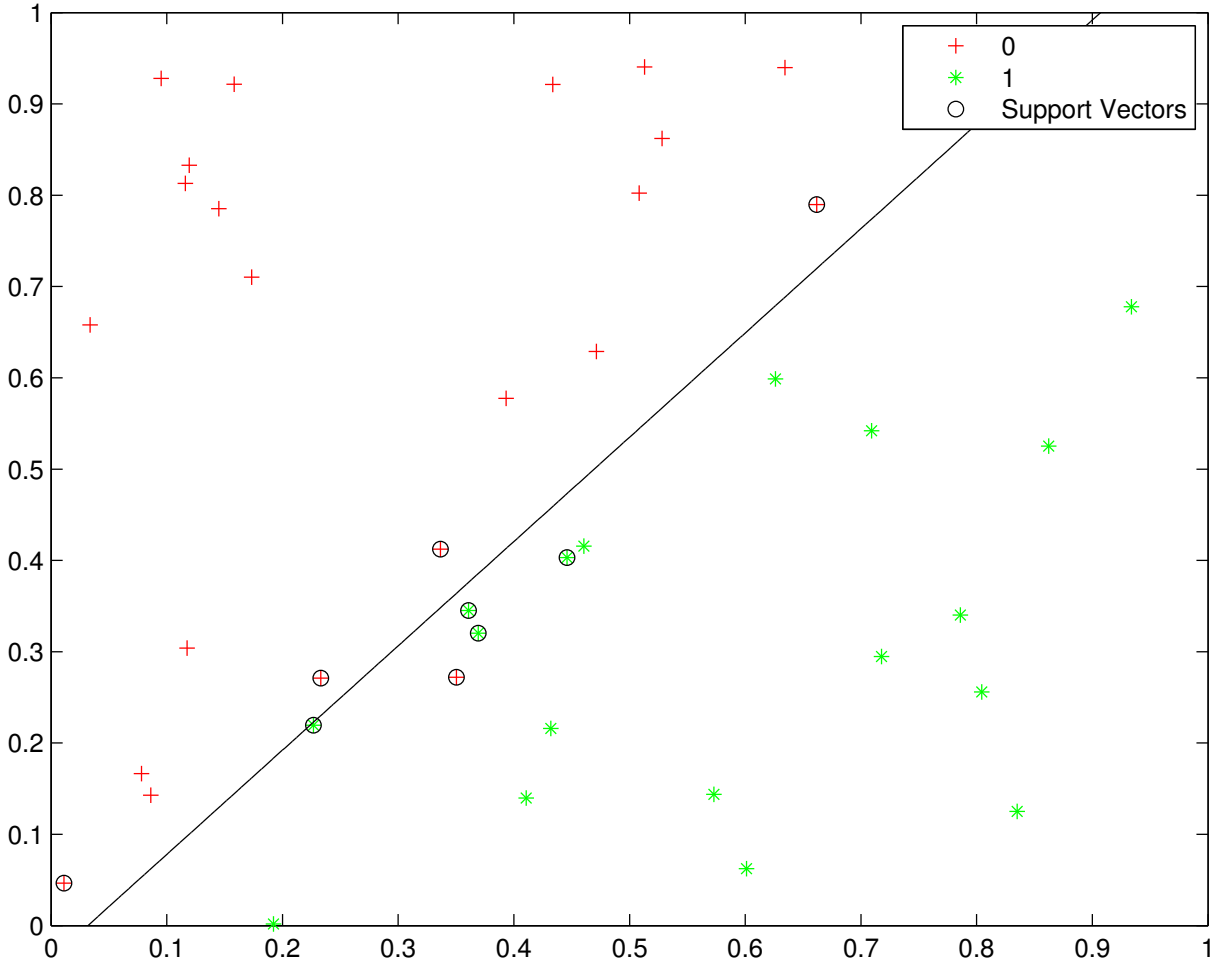
Δες demo_svmNonSepar.m στο e-class



Η παράμετρος κανονικοποίησης ήταν $C = 1$

Μη γραμμικά διαχωρισίμα δεδομένα: κατηγορίες με επικάλυψη

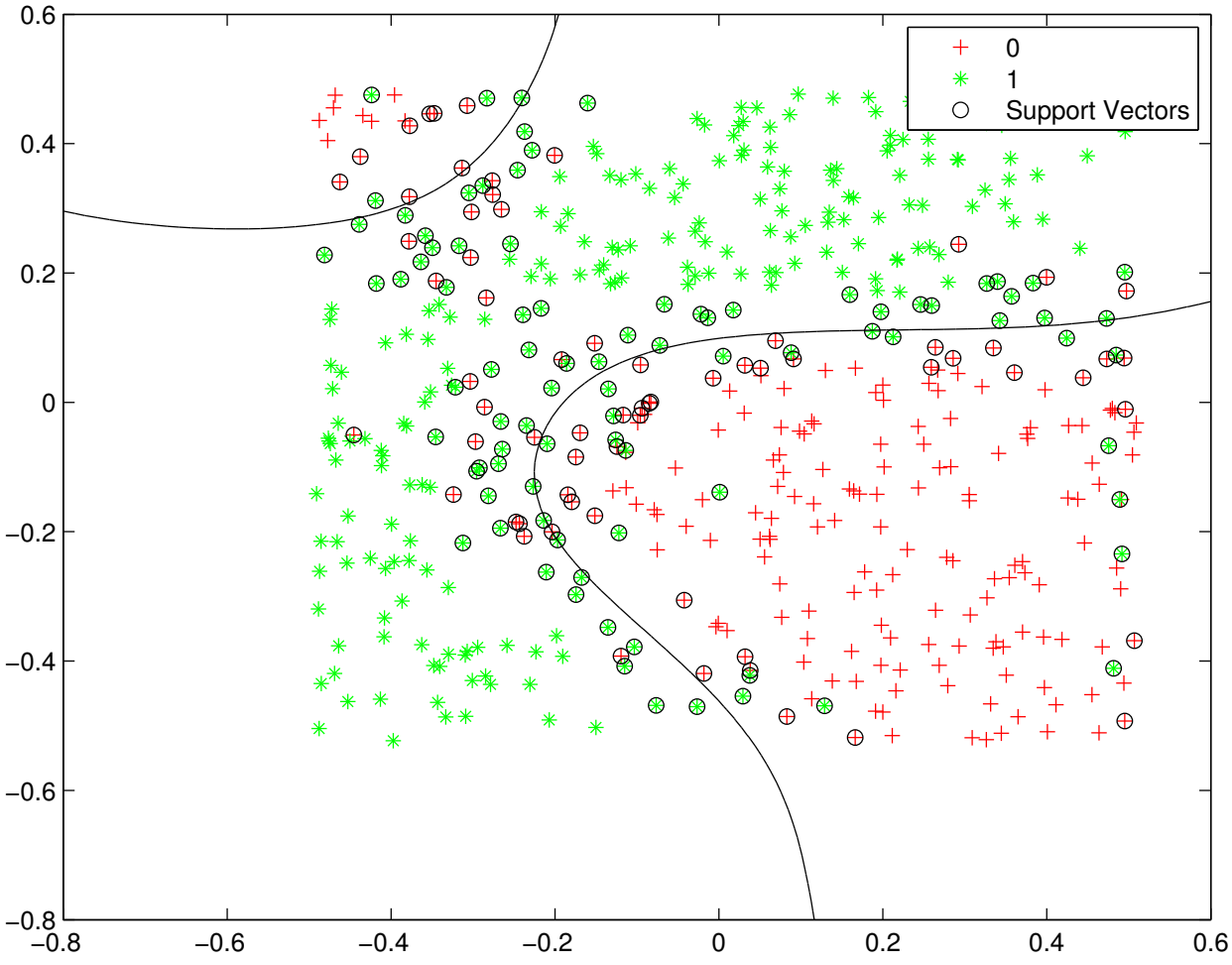
Δες demo_svmNonSepar.m στο e-class



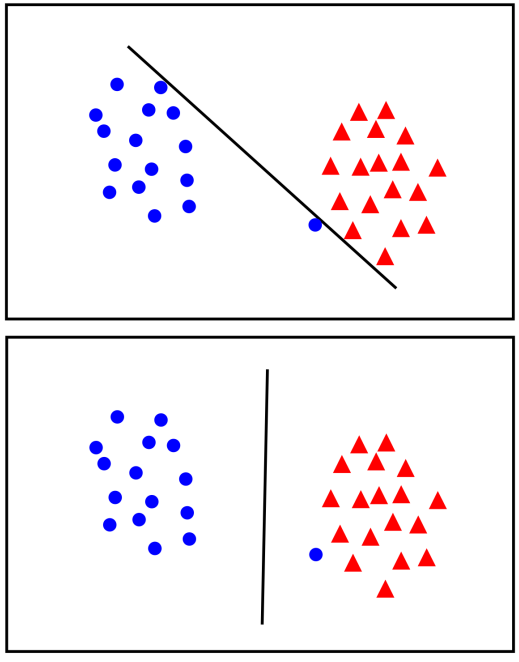
Η παράμετρος κανονικοποίησης ήταν $C = 10$

Μη γραμμικά διαχωρισίμα δεδομένα: κατηγορίες με επικάλυψη

Δες demo_svmRbf.m στο e-class



Επιλογή της παραμέτρου κανονικοποίησης C



- Η εύρεση μιας διαχωριστικής γραμμής, με πολύ μικρό margin, μπορεί να οδηγήσει σε overfitting
- Λύσεις με μεγάλα margin (αφήνοντας κάποια δεδομένα εκπαίδευσης να κατηγοριοποιηθούν εσφαλμένα), μπορεί να οδηγήσουν σε καλύτερη γενίκευση

Επιλογή της παραμέτρου κανονικοποίησης C

Οπότε, όπως προαναφέρθηκε, ακόμα και αν γνωρίζουμε ότι τα δεδομένα είναι γραμμικά διαχωρίσιμα επιτρέπουμε λάθη στην κατηγοριοποίηση των δεδομένων του συνόλου εκπαίδευσης λύνοντας το πρόβλημα βελτιστοποίησης

$$\min_{\mathbf{w},b,\boldsymbol{\xi}} C\sum_{n=1}^{N}\xi_n + \frac{1}{2}||\mathbf{w}||^2$$

$$\text{subject to: } \xi_n \geq 0, \quad t_n(\mathbf{w}^T\mathbf{x}_n + w_0) \geq 1 - \xi_n, \ n = 1, \ldots, N$$

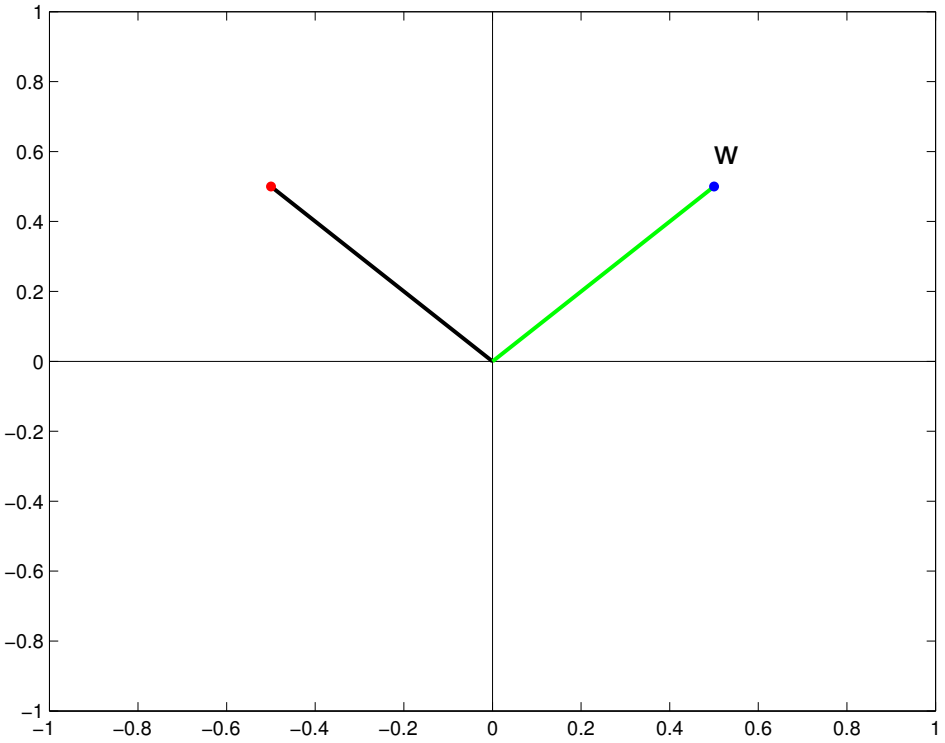
Η επιλογή της παραμέτρου κανονικοποίησης C γίνεται με cross validation ώστε να αποφεύγουμε το φαινόμενο της υπερεκπαίδευσης

Επίλογος

- Διάβασμα για το σπίτι: Bishop: sections 4.1.1, 4.1.7, κεφάλαιο 7 ώς τη σελίδα 331. Επιπλέον διάβασμα (προαιρετικό): 7.1.1, 7.1.2 και appendix E για πολλαπλασιαστές Lagrange

- Επόμενο μάθημα: *K-means*, μίξεις *Gaussian* κατανομών και μάθηση με τον αλγόριθμο EM

Παράρτημα: Γεωμετρία συνόρων απόφασης

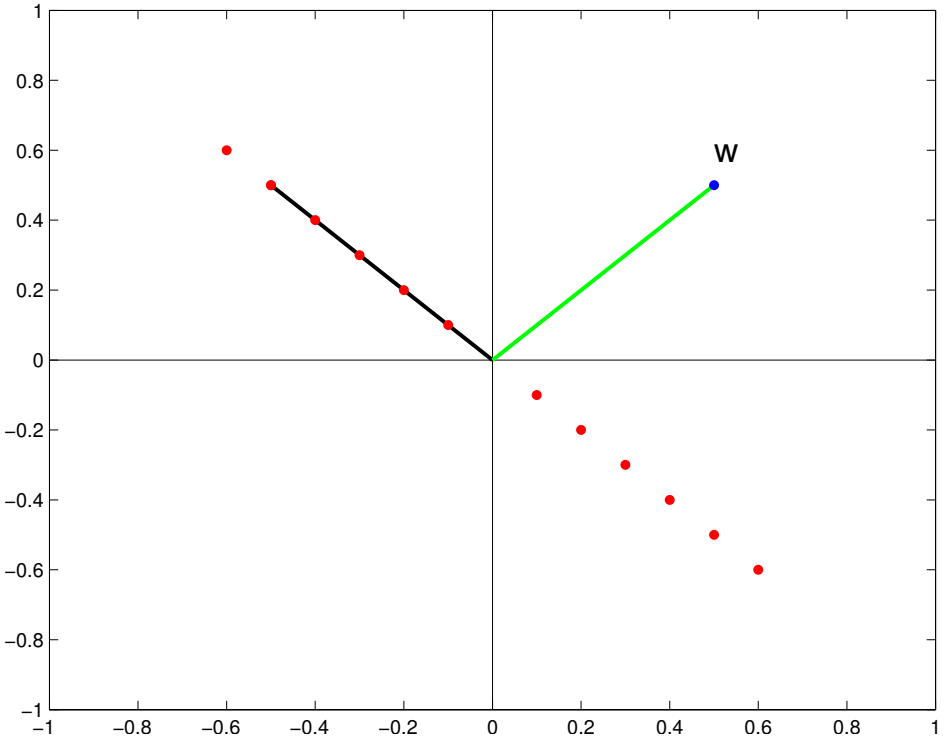


- Δύο διανύσματα $\mathbf{w}$ και $\mathbf{x}$ (π.χ. $\mathbf{w} = [0.5\ 0.5]$, $\mathbf{x} = [-0.5\ 0.5]$) είναι ορθογώνια όταν ισχύει

$$\mathbf{w}^T\mathbf{x} = \sum_{d=1}^{D} w_d x_d = 0$$

δηλαδή το εσωτερικό γινόμενό τους είναι μηδέν

Παράρτημα: Γεωμετρία συνόρων απόφασης

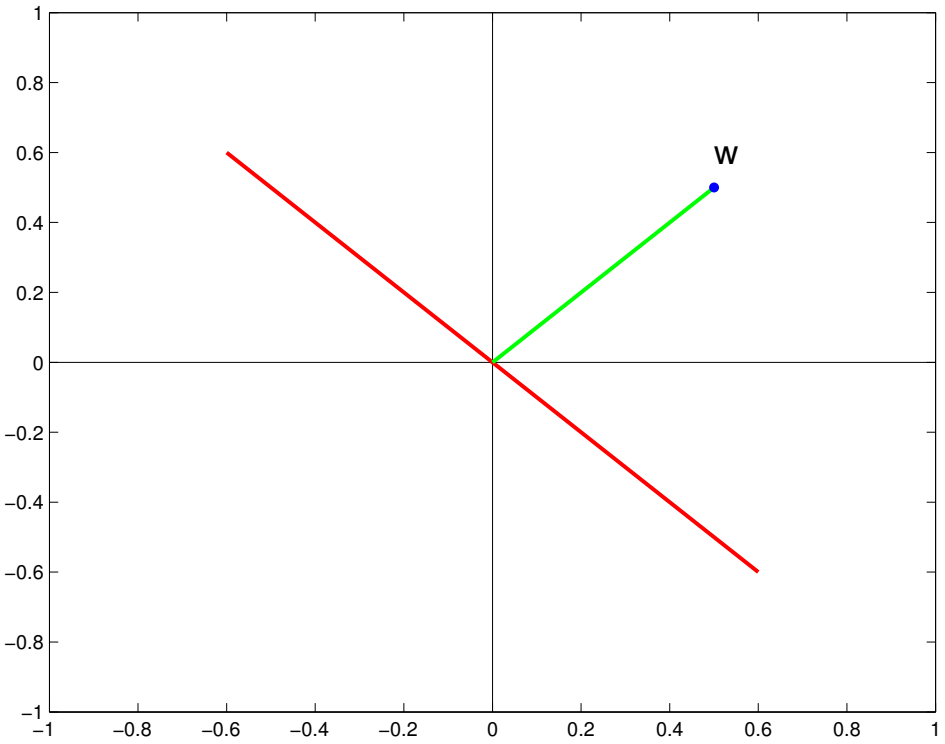


- Έστω τώρα ότι ενδιαφερόμαστε για πολλά $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_2, \ldots$ για τα οποία

$$\mathbf{w}^T \mathbf{x}_i = 0, \quad i = 1, 2, 3, \ldots$$

Π.χ. $\mathbf{w} = [0.5\ 0.5]$ είναι ορθογώνιο διάνυσμα στα $\mathbf{x} = [-0.5\ 0.5]$, $\mathbf{x} = [-0.3\ 0.3]$, $\mathbf{x} = [0.5\ -0.5]$ κτλ

Παράρτημα: Γεωμετρία συνόρων απόφασης

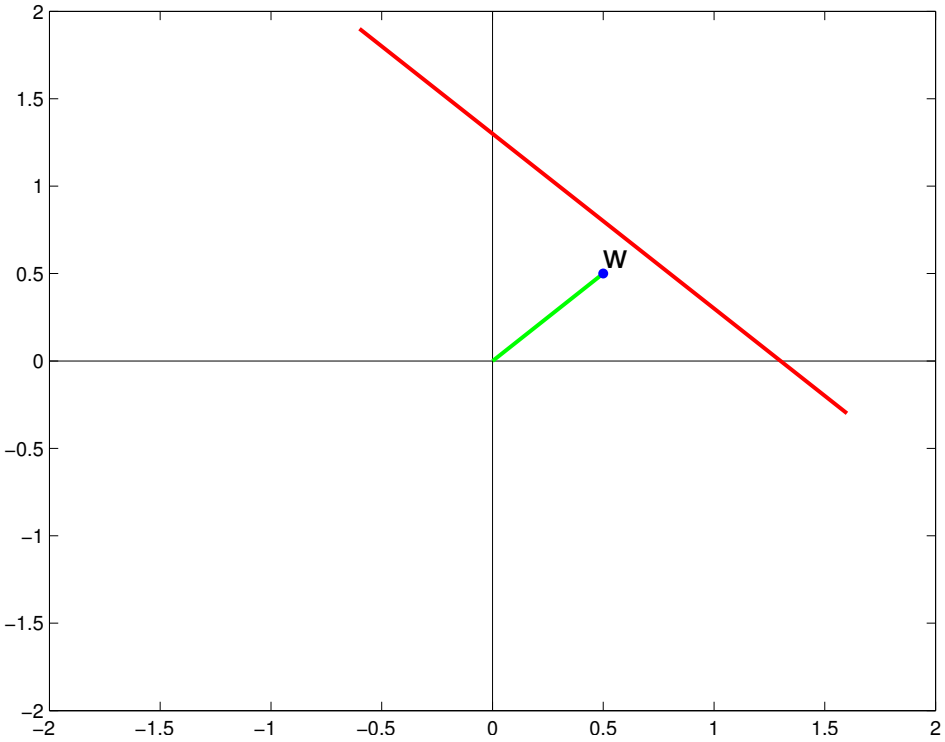


- Το σύνολο όλων των $\mathbf{x}$ σχηματίζει μια γραμμή (την κόκκινη στο σχήμα)

$$y(\mathbf{x}) = \mathbf{w}^T\mathbf{x}, \quad y(\mathbf{x}) = 0$$

έτσι ώστε το $\mathbf{w}$ είναι ορθογώνιο διάνυσμα στη γραμμή αυτή

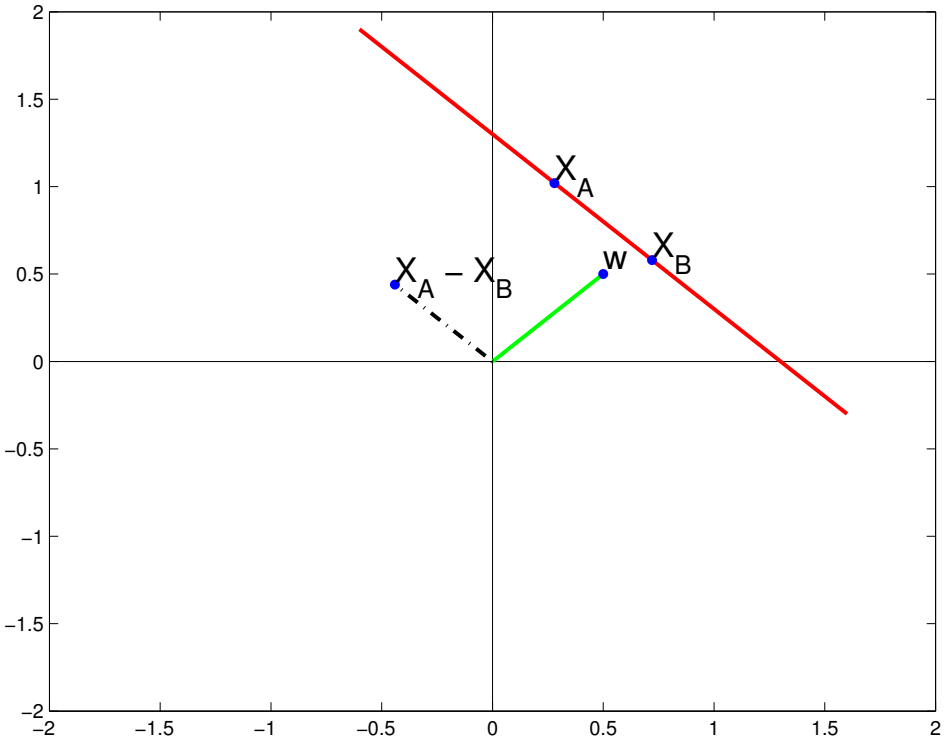
Παράρτημα: Γεωμετρία συνόρων απόφασης



- Η $y(\mathbf{x}) = 0$ παιρνά από την αρχή των αξόνων (δηλ. $\mathbf{x} = 0$ ικανοποιεί την εξίσωση)
- Αν θέλουμε μια γραμμή που να μην περνά από την αρχή των αξόνων προσθέτουμε ένα w_0

$$y(\mathbf{x}) = w^T \mathbf{x} + w_0$$

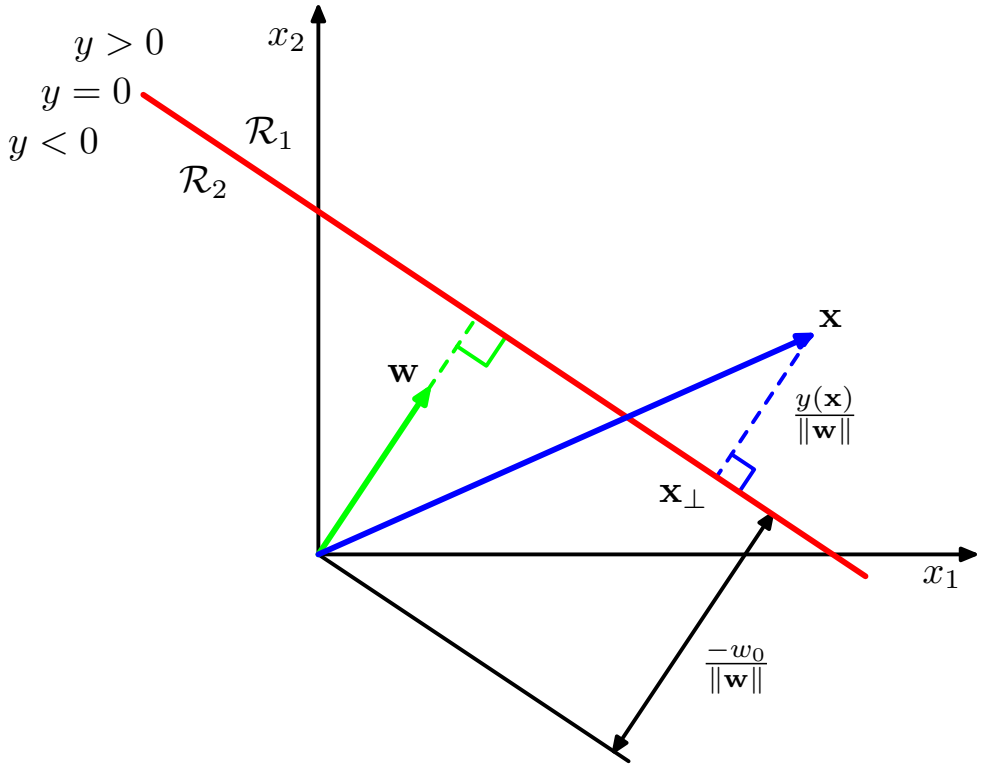
Παράρτημα: Γεωμετρία συνόρων απόφασης



$$y(\mathbf{x}) = w^T\mathbf{x} + w_0$$

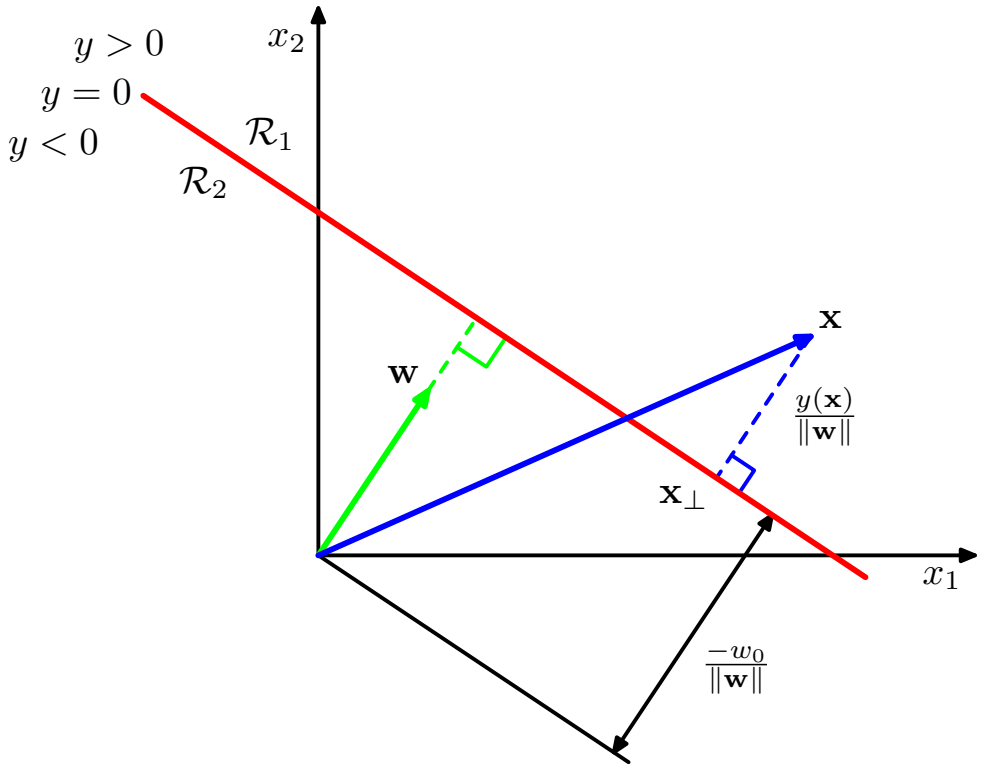
- **w** παραμένει ορθογώνιο σε κάθε διάνυσμα που βρίσκεται κατά μήκος της γραμμής
- Ένα τέτοιο διάνυσμα μπορεί να γραφεί ως $\mathbf{x}_A - \mathbf{x}_B$, όπου $y(\mathbf{x}_A) = 0$, $y(\mathbf{x}_B) = 0$, $\Rightarrow \mathbf{w}^T(\mathbf{x}_A - \mathbf{x}_B) = 0$

Παράρτημα: Γεωμετρία συνόρων απόφασης



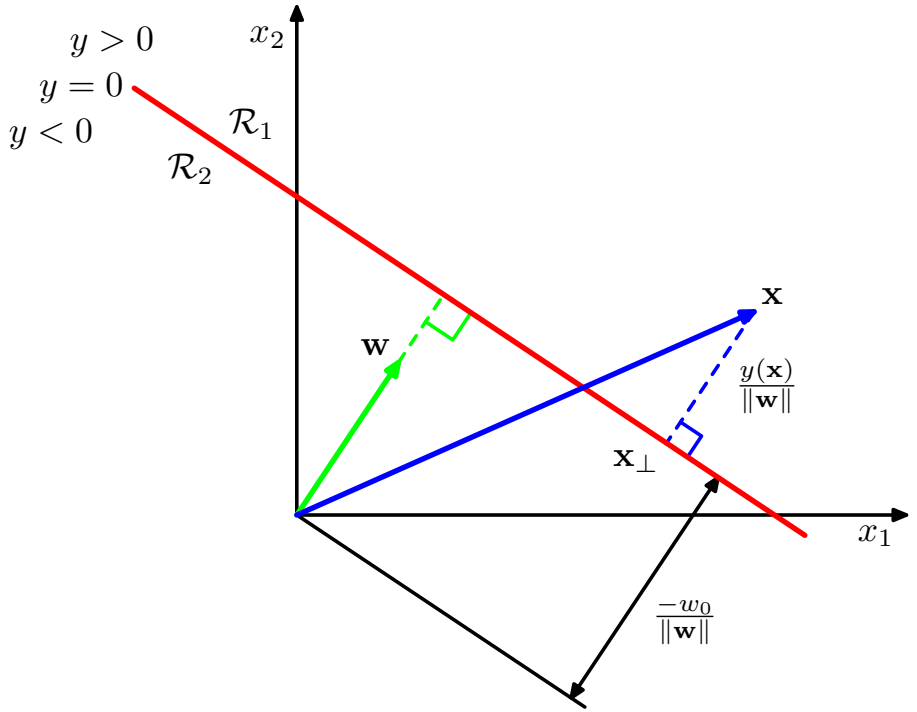
- $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$ διαχωρίζει γραμμικά το χώρο $\mathbb{R}^D$ ($D = 2$ στη εικόνα)
- Ερώτημα: Πόσο απέχει το σύνορο απόφασης $y(\mathbf{x}) = 0$ από την αρχή των αξόνων;

Παράρτημα: Γεωμετρία συνόρων απόφασης



- Ερώτημα: Πόσο απέχει το σύνορο απόφασης $y(\mathbf{x}) = 0$ από την αρχή των αξόνων;
- Η απόσταση αυτή είναι είναι ίση με την προέκταση της πράσινης γραμμής που ενώνει την αρχή των αξόνων με την κόκκινη γραμμή (και είναι κάθετη στην κόκκινη γραμμή)

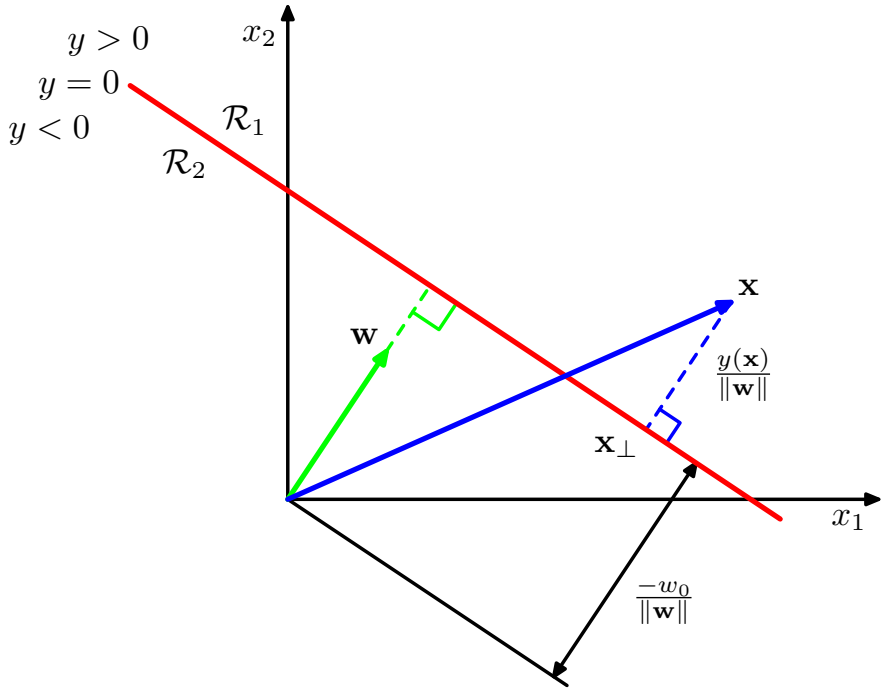
Παράρτημα: Γεωμετρία συνόρων απόφασης



- Η απόσταση αυτή είναι το μήκος του διανύσματος $\mathbf{x}_G$, όπου $\mathbf{x}_G$ το σημείο της κόκκινης γραμμής ($\mathbf{w}^T\mathbf{x}_G + w_0 = 0$) που τέμνεται από την προέκταση της πράσινης γραμμής
- $\mathbf{x}_G$ είναι παράλληλο (αφού είναι προέκταση) του $\mathbf{w}$ ή ισόδυναμα του κανονικοποιημένου $\mathbf{w}$ (δηλ. του $\frac{\mathbf{w}}{||\mathbf{w}||}$)

$$\mathbf{x}_G = g\frac{\mathbf{w}}{||\mathbf{w}||}, \quad \text{για κάποιο } g \in \mathbb{R}$$

Παράρτημα: Γεωμετρία συνόρων απόφασης



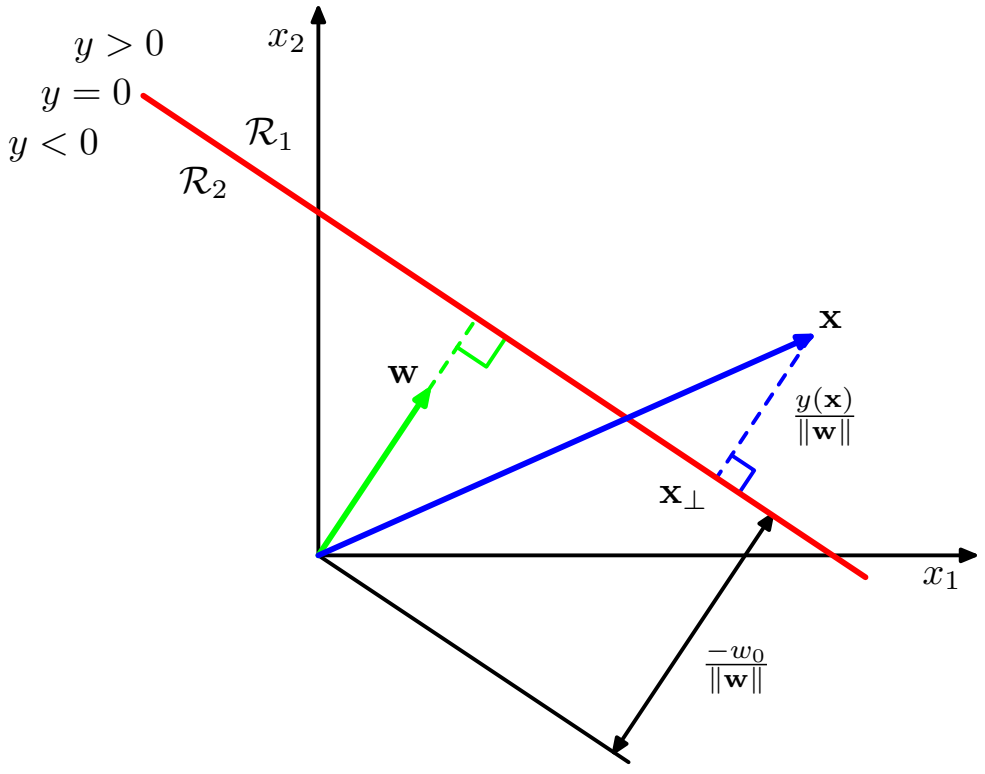
$$\mathbf{x}_G = g\frac{\mathbf{w}}{||\mathbf{w}||}, \quad \text{για κάποιο } g \in \mathbb{R}$$

- Πολλαπλασιάζοντας και τις δύο πλευρές της εξίσωσης με $\mathbf{w}$ και προσθέτοντας w_0 προκύπτει

$$\mathbf{w}^T\mathbf{x}_G + w_0 = g\frac{\mathbf{w}^T\mathbf{w}}{||\mathbf{w}||} + w_0 \Rightarrow 0 = g||\mathbf{w}|| + w_0 \Rightarrow g = -\frac{w_0}{||\mathbf{w}||}$$

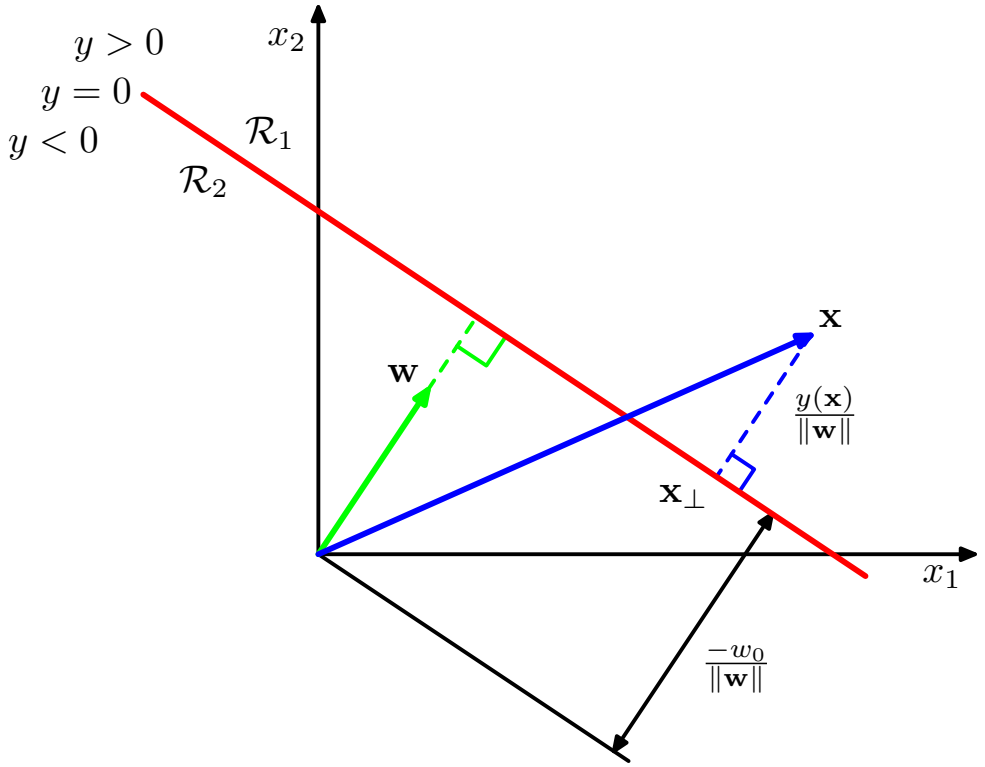
- Απόσταση $\Rightarrow ||\mathbf{x}_G|| = |g|\frac{||\mathbf{w}||}{||\mathbf{w}||} = |g|$

Παράρτημα: Γεωμετρία συνόρων απόφασης



- Ερώτημα: Έστω τώρα ότι θα θέλαμε να βρούμε πόσο απέχει ένα οποιοδήποτε $\mathbf{x}$ από το σύνορο $y(\mathbf{x}) = 0$
- Επίσης έστω ότι μας ενδιαφέρει η απόσταση με πρόσιμο (singed measure), που μπορεί να είναι θετική ή αρνητική (ανάλογα σε ποια πλευρά του συνόρου βρισκόμαστε)

Παράρτημα: Γεωμετρία συνόρων απόφασης

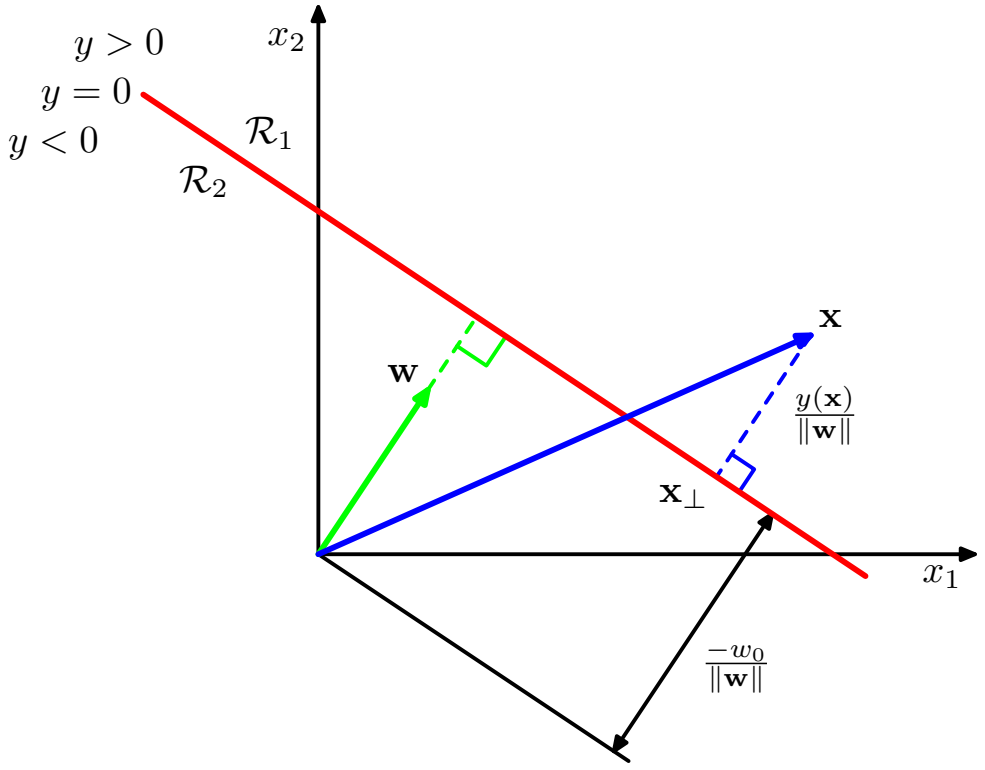


- Αν $\mathbf{x}_\perp$ είναι η προβολή του $\mathbf{x}$ στο σύνορο, τότε το $\mathbf{x} - \mathbf{x}_\perp$ είναι παράλληλο στο $\frac{\mathbf{w}}{||\mathbf{w}||}$

$$\mathbf{x} - \mathbf{x}_\perp = r\frac{\mathbf{w}}{||\mathbf{w}||} \Rightarrow \mathbf{x} = \mathbf{x}_\perp + r\frac{\mathbf{w}}{||\mathbf{w}||}, \text{ για κάποιο } r \in \mathbb{R}$$

- Πολλαπλασιάζοντας και τις δύο πλεύρες με $\mathbf{w}$ και προσθέτοντας w_0 προκύπτει $\mathbf{w}^T\mathbf{x} + w_0 = \mathbf{w}^T\mathbf{x}_\perp + w_0 + r||\mathbf{w}|| \Rightarrow r = \frac{y(\mathbf{x})}{||\mathbf{w}||}$

Παράρτημα: Γεωμετρία συνόρων απόφασης



- Συμπέρασμα: Το κάθε $\mathbf{x}$ απέχει από το σύνορο απόφασης απόσταση

$$||\mathbf{x} - \mathbf{x}_\perp|| = |r| = \frac{|y(\mathbf{x})|}{||w||}$$

- $r = \frac{y(\mathbf{x})}{||\mathbf{w}||}$ (χωρίς απόλυτη τιμή) μας δίνει επιπλέον την πληροφορία για τη πλευρά του συνόρου που περιέχει το $\mathbf{x}$