ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Νευρωνικά Δίκτυα και Ευφυή Υπολογιστικά Συστήματα

Μηχανες Διανυσματων Υποστηριξης

Ταξινόμηση προτύπων

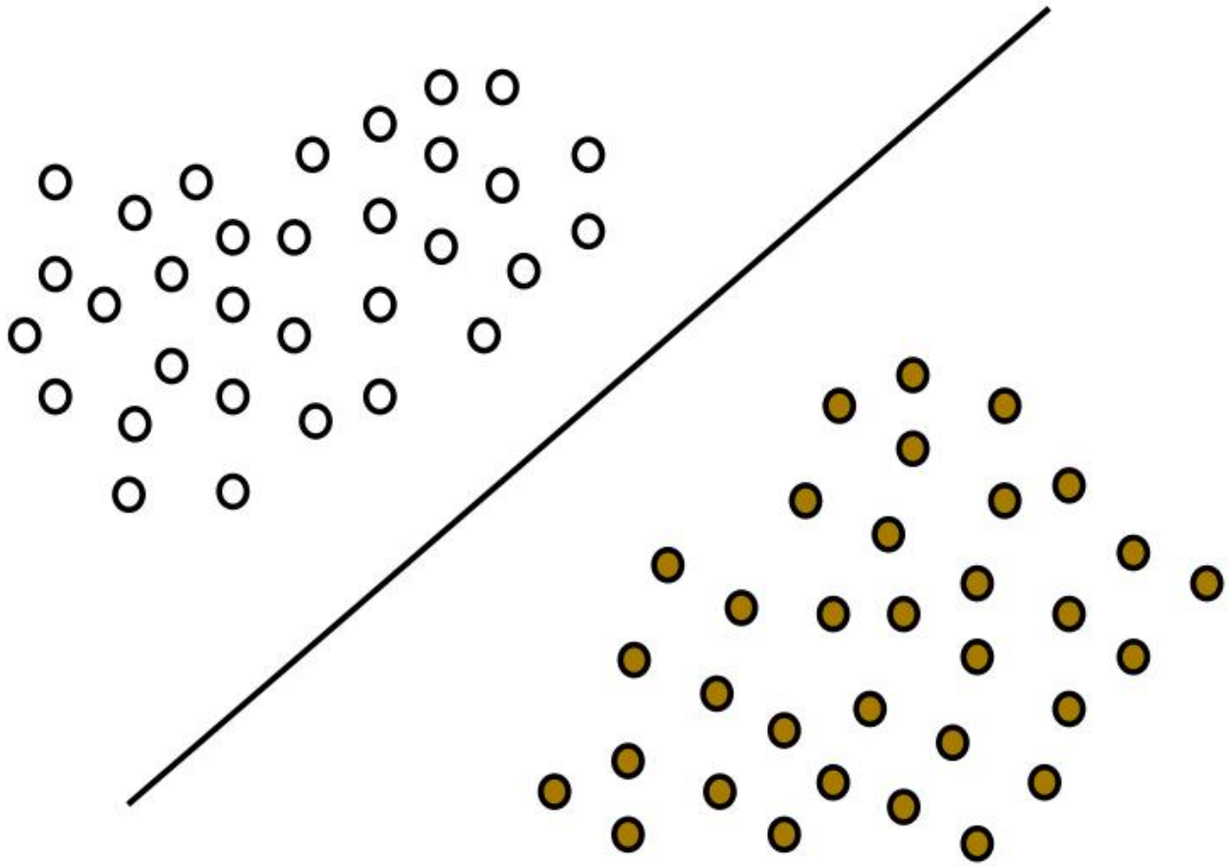
Προβλήματα στην ταξινόμηση με νευρωνικά δίκτυα

- Η ταξινόμηση με perceptrons δουλεύει μόνο με γραμμικά διαχωρίσιμες κλάσεις
- Η ταξινόμηση με δίκτυα MLP υποφέρει από βραδεία εκπαίδευση (μην ξεχνάμε ότι στην περίπτωση αυτή λύνουμε γενικότερο πρόβλημα από αυτό της ταξινόμησης)

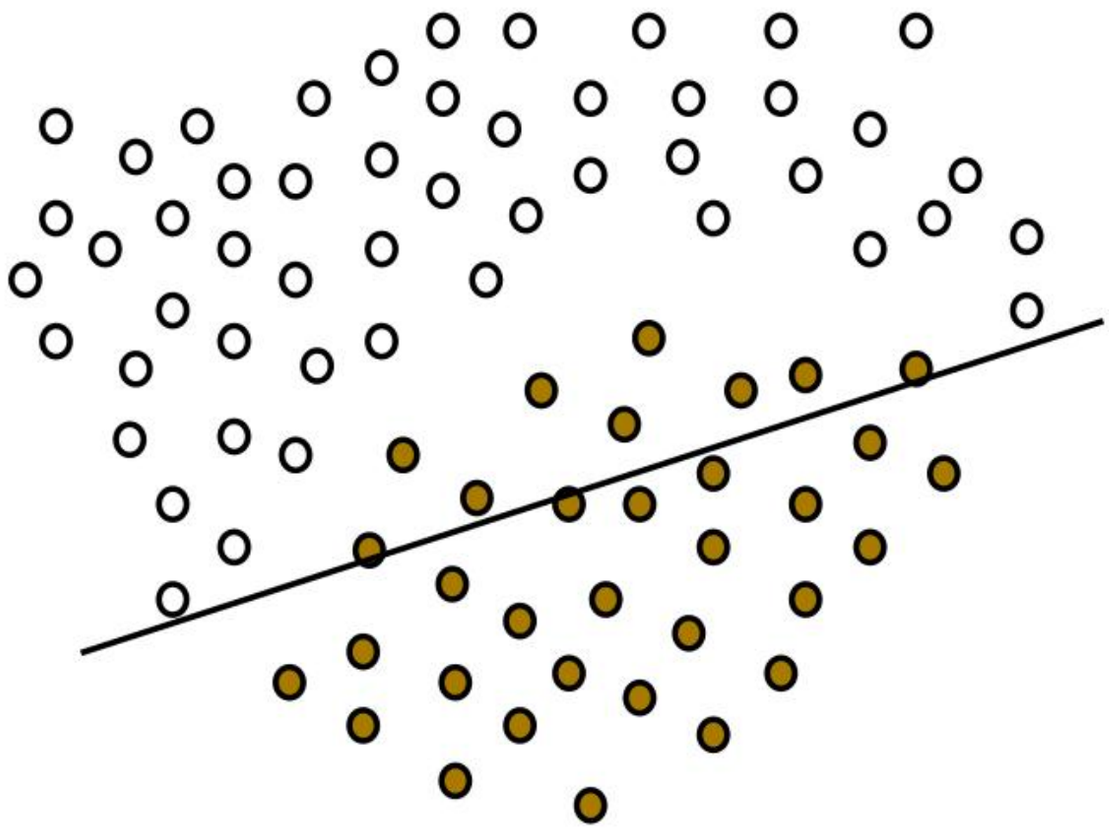
Ιδέα

- Αν επικεντρωθούμε στο πρόβλημα της ταξινόμησης μπορούμε να πετύχουμε καλύτερους χρόνους εκπαίδευσης και καλύτερες ιδιότητες γενίκευσης

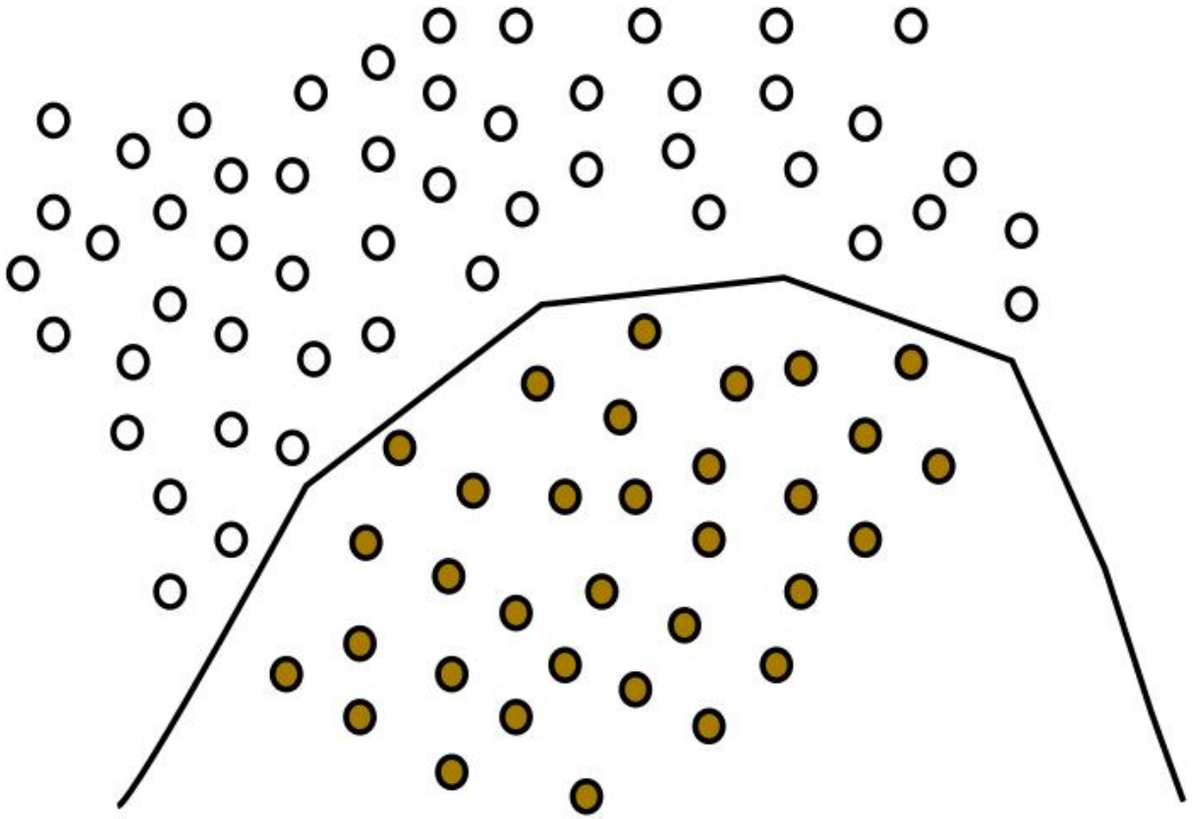
Το πρόβλημα της ταξινόμησης



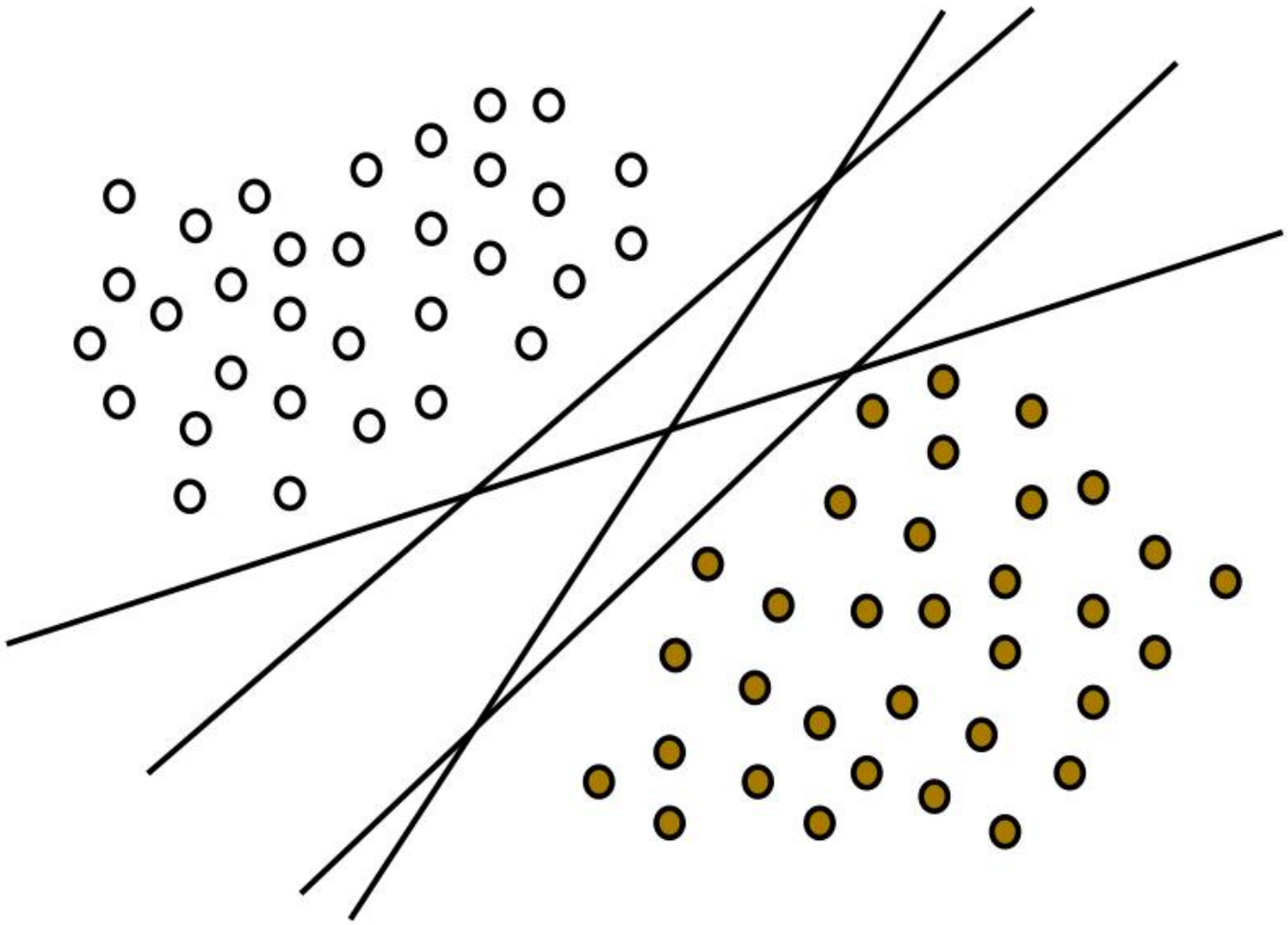
Μη γραμμικά διαχωρίσιμες κλάσεις



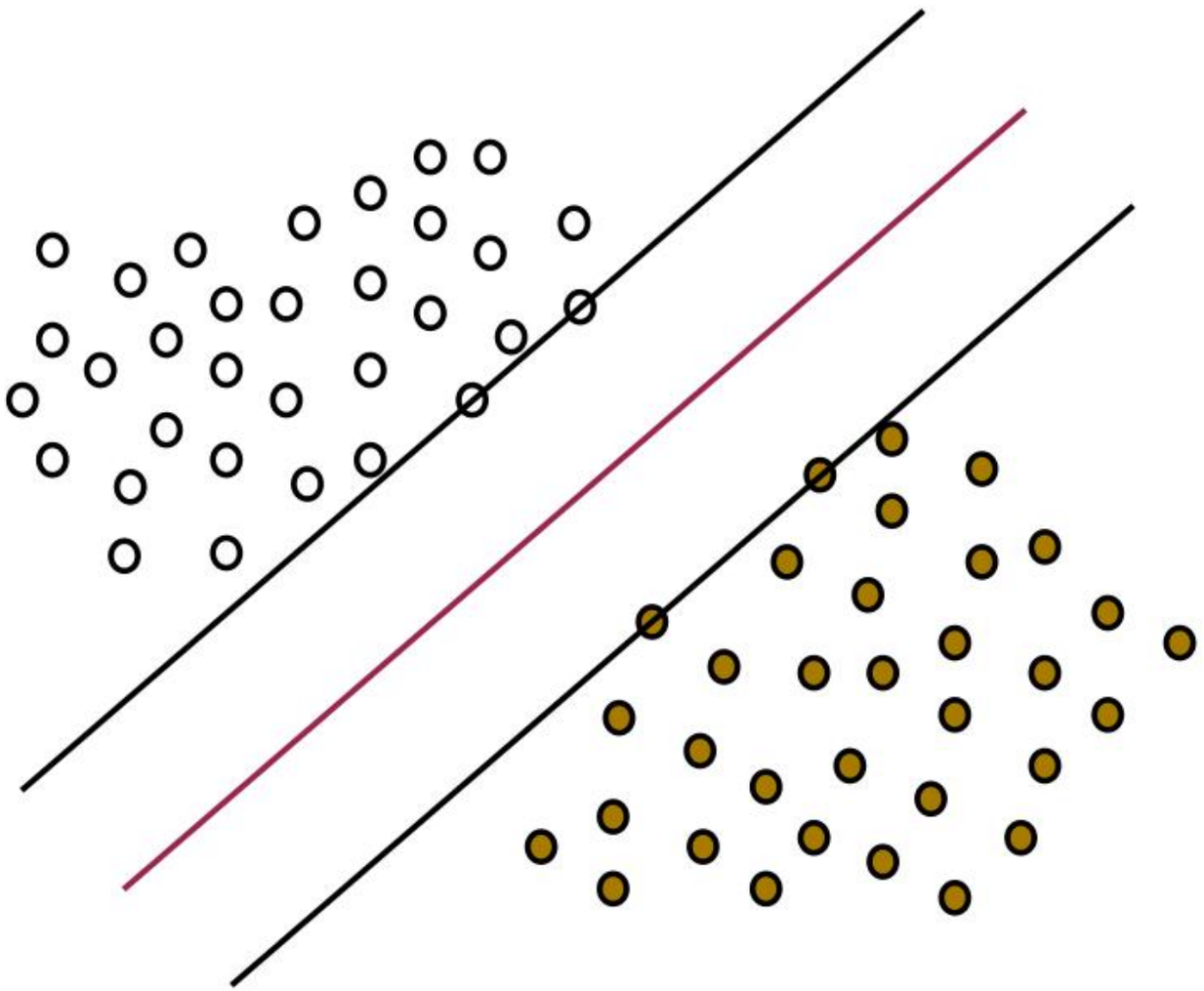
Μη γραμμικά διαχωρίσιμες κλάσεις



Ευθείες διαχωρισμού



Βέλτιστη ευθεία διαχωρισμού



Τυπικός ορισμός προβλήματος

Διατύπωση προβλήματος

Δίνεται ένα σύνολο ζευγών $(\mathbf{x}_1, d_1), \ldots, (\mathbf{x}_p, d_p)$
με $d_i = -1$ αν $\mathbf{x}_i \in \mathcal{C}_o$ και $d_i = 1$ αν $\mathbf{x}_i \in \mathcal{C}_1$

Ζητάμε την εύρεση των βαρών **w** και του κατωφλίου w_o, έτσι ώστε:
$\mathbf{w}^\top \mathbf{x}_i + w_o \geq 0$ αν $d_i = 1$ $(\mathbf{x}_i \in \mathcal{C}_o)$
$\mathbf{w}^\top \mathbf{x}_i + w_o < 0$ αν $d_i = -1$ $(\mathbf{x}_i \in \mathcal{C}_1)$

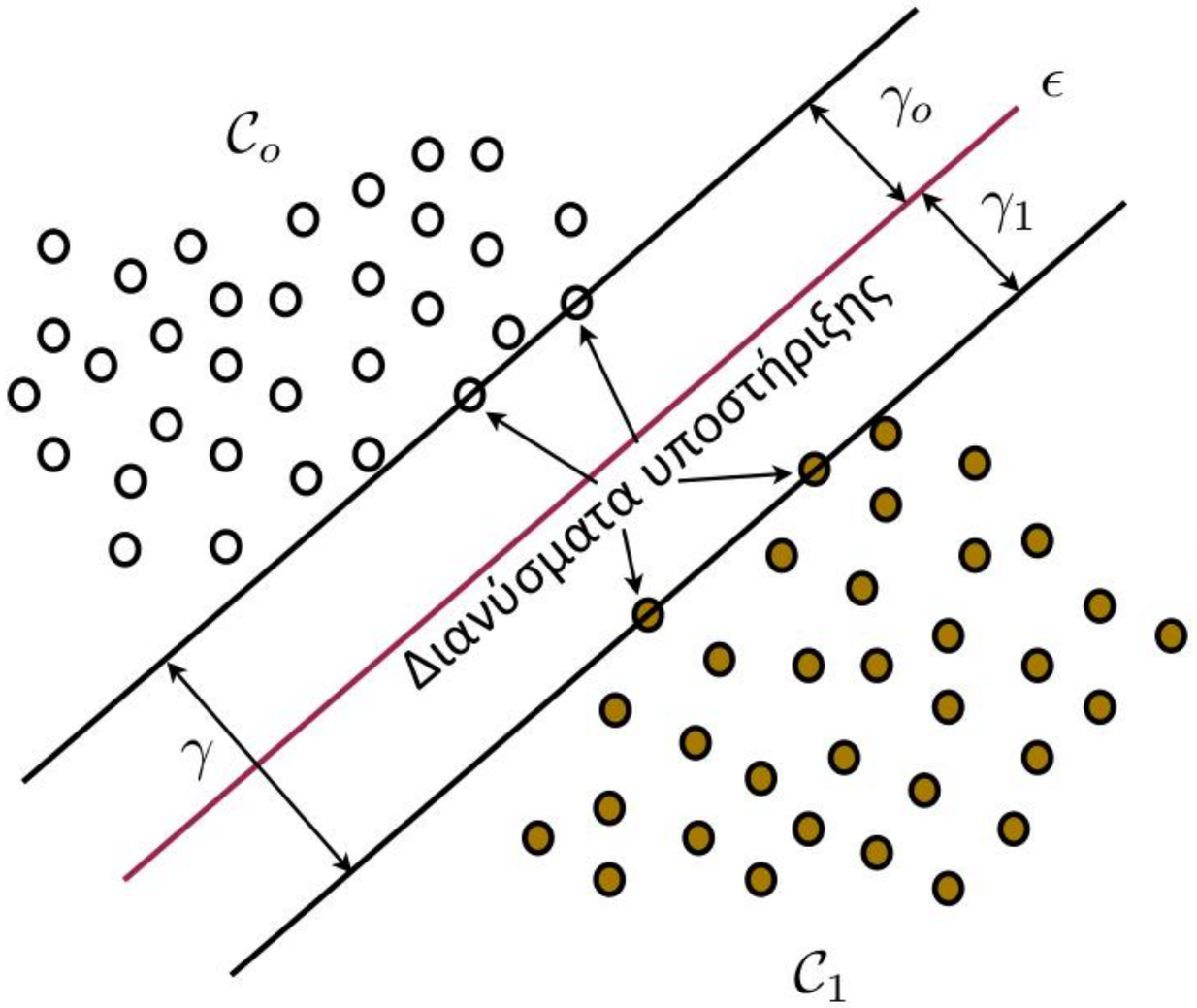
Υπόθεση

Υπάρχει τέτοια ευθεία (οι κλάσεις είναι γραμμικά διαχωρίσιμες)

Απαίτηση

Η ευθεία που θα κατασκευαστεί πρέπει να έχει όσο το δυνατόν *μεγαλύτερο περιθώριο ταξινόμησης*

Βέλτιστο υπερεπίπεδο διαχωρισμού



$$\gamma_o = \min_{\mathbf{x} \in \mathcal{C}_o} d(\mathbf{x}, \epsilon)$$

$$\gamma_1 = \min_{\mathbf{x} \in \mathcal{C}_1} d(\mathbf{x}, \epsilon)$$

$$\gamma = \gamma_o + \gamma_1$$

Κανονικό υπερεπίπεδο

$$\gamma_o = \gamma_1$$

$$\mathbf{w}^\top \mathbf{x}_i + w_o \geq 1 \text{ αν } \mathbf{x}_i \in \mathcal{C}_o$$
$$\mathbf{w}^\top \mathbf{x}_i + w_o \leq 1 \text{ αν } \mathbf{x}_i \in \mathcal{C}_1$$

Υπολογισμός περιθωρίου

Η συνάρτηση $g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_o$ ένα μέτρο της απόστασης του **x** από το βέλτιστο υπερεπίπεδο (όπου **w** και w_o τα βέλτιστα βάρη).

Υπολογίζουμε το **x** ως $\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$, όπου r η απόσταση του **x** από το βέλτιστο υπερεπίπεδο

Συνεπώς $g(\mathbf{x}) = \mathbf{w}^\top \left(\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + w_o$

$$\Rightarrow g(\mathbf{x}) = \mathbf{w}^\top \mathbf{x}_p + w_o + r\frac{\mathbf{w}^\top \mathbf{w}}{\|\mathbf{w}\|}$$

$$\Rightarrow g(\mathbf{x}) = r\frac{\|\mathbf{w}\|^2}{\|\mathbf{w}\|} \Rightarrow r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}$$

Άρα αφού για τα διανύσματα υποστήριξης έχουμε $g(x) = 1$ ($\mathbf{x}_i \in \mathcal{C}_o$) και $g(x) = -1$ ($\mathbf{x}_i \in \mathcal{C}_1$)

Τελικά: $\boxed{\gamma = \frac{2}{\|\mathbf{w}\|}}$

Βέλτιστο διαχωριστικό υπερεπίπεδο

Ορισμός προβλήματος βελτιστοποίησης

Υπολόγισε το ελάχιστο της συνάρτησης:

$$\mathcal{J}(\mathbf{w}, w_o) = \frac{1}{2}\|\mathbf{w}\|^2$$

υπό τους περιορισμούς των P ανισοτήτων:

$$d_i(\mathbf{w}^\top \mathbf{x}_i + w_o) \geq 1,\ i = 1, \ldots, P$$

Παρατηρήσεις

- Η συνάρτηση κόστους είναι κυρτή
- Οι περιορισμοί είναι γραμμικοί

Καλούματε να επιλύσουμε ένα πρόβλημα *τετραγωνικού προγραμματισμού*

Μέθοδος πολλαπλασιαστών Lagrange

Ορίζουμε τη συνάρτηση κόστους:

$$\mathcal{L}(\mathbf{w}, w_o, \lambda_1, \ldots, \lambda_p) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{P} \lambda_i [d_i(\mathbf{w}^\top \mathbf{x}_i + w_o) - 1]$$

με $\lambda_i \geq 0,\ i = 1, \ldots, P$

Η συνάρτηση αυτή πρέπει να ελαχιστοποιηθεί ως προς τα **w**, w_o
και να μεγιστοποιηθεί ως προς τα λ_i

Συνθήκες Karush-Kuhn-Tucker (για το βέλτιστο σημείο)

$$\frac{\partial L}{\partial w_o} = 0 \qquad \frac{\partial L}{\partial \mathbf{w}} = 0 \qquad \lambda_i [d_i(\mathbf{w}^\top \mathbf{x}_i + w_o) - 1] \geq 0,\ i = 1, \ldots, P$$

Βέλτιστη διαχωριστική επιφάνεια

Από τις συνθήκες KKT έχουμε:

$$\frac{\partial L}{\partial w_o} = 0 \quad \longrightarrow \quad \sum_{i=1}^{P} \lambda_i d_i = 0$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \quad \longrightarrow \quad \mathbf{w} = \sum_{i=1}^{P} \lambda_i d_i \mathbf{x}_i$$

Συνεπώς η βέλτιστη διαχωριστική επιφάνεια δίνεται από τη σχέση:

$$g^*(\mathbf{x}) = \mathbf{w}^\top \mathbf{x} + w_o = \sum_{i=1}^{P} \lambda_i d_i \mathbf{x}_i^\top \mathbf{x} + w_o$$

Βέλτιστη πόλωση

Για τα διανύσματα υποστήριξης ισχύει ότι:

$$d_i(\mathbf{w}^\top \mathbf{x}_i + w_o) = 1 \longrightarrow w_o = \frac{1}{d_i} - \mathbf{w}^\top \mathbf{x}_i$$

Για λόγους αριθμητικής ευστάθειας, χρησιμοποιούμε τη σχέση:

$$w_o = \frac{1}{|I_{sv}|} \sum_{i \in I_{sv}} \left(\frac{1}{d_i} - \mathbf{w}^\top \mathbf{x}_i \right)$$

όπου:

$$I_{sv} = \{ i : \mathbf{x}_i \text{ διάνυσμα υποστήριξης } \}$$

Παρατήρηση

Οι μόνοι πολλαπλασιαστές λ_i που μπορούν να είναι θετικοί είναι αυτοί που αντιστοιχούν σε κάποιο διάνυσμα υποστήριξης $\mathbf{x}_i$.
Για τους υπόλοιπους ισχύει $\lambda_i = 0$.

Δυϊκό πρόβλημα (1)

Από τα παραπάνω έχουμε:

$$\frac{1}{2}\|w\|^2 = \frac{1}{2}\mathbf{w}^\top\mathbf{w} = \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\lambda_i\lambda_j d_i d_j \mathbf{x}_i^\top \mathbf{x}_j$$

$$\sum_{i=1}^{P}\lambda_i[d_i(\mathbf{w}^\top\mathbf{x}_i + w_o) - 1] = \sum_{i=1}^{P}\lambda_i d_i \sum_{j=1}^{P}\lambda_j d_j \mathbf{x}_j^\top \mathbf{x}_i + w_o\sum_{i=1}^{P}\lambda_i d_i - \sum_{i=1}^{P}\lambda_i$$

$$= \sum_{i=1}^{P}\sum_{j=1}^{P}\lambda_i\lambda_j d_i d_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^{P}\lambda_i$$

Επομένως:

$$\mathcal{L}(\lambda_1, \ldots, \lambda_P) = \sum_{i=1}^{P}\lambda_i - \frac{1}{2}\sum_{i=1}^{P}\sum_{j=1}^{P}\lambda_i\lambda_j d_i d_j \mathbf{x}_i^\top \mathbf{x}_j$$

Δυϊκό πρόβλημα (2)

Ορισμός δυϊκού προβλήματος βελτιστοποίησης

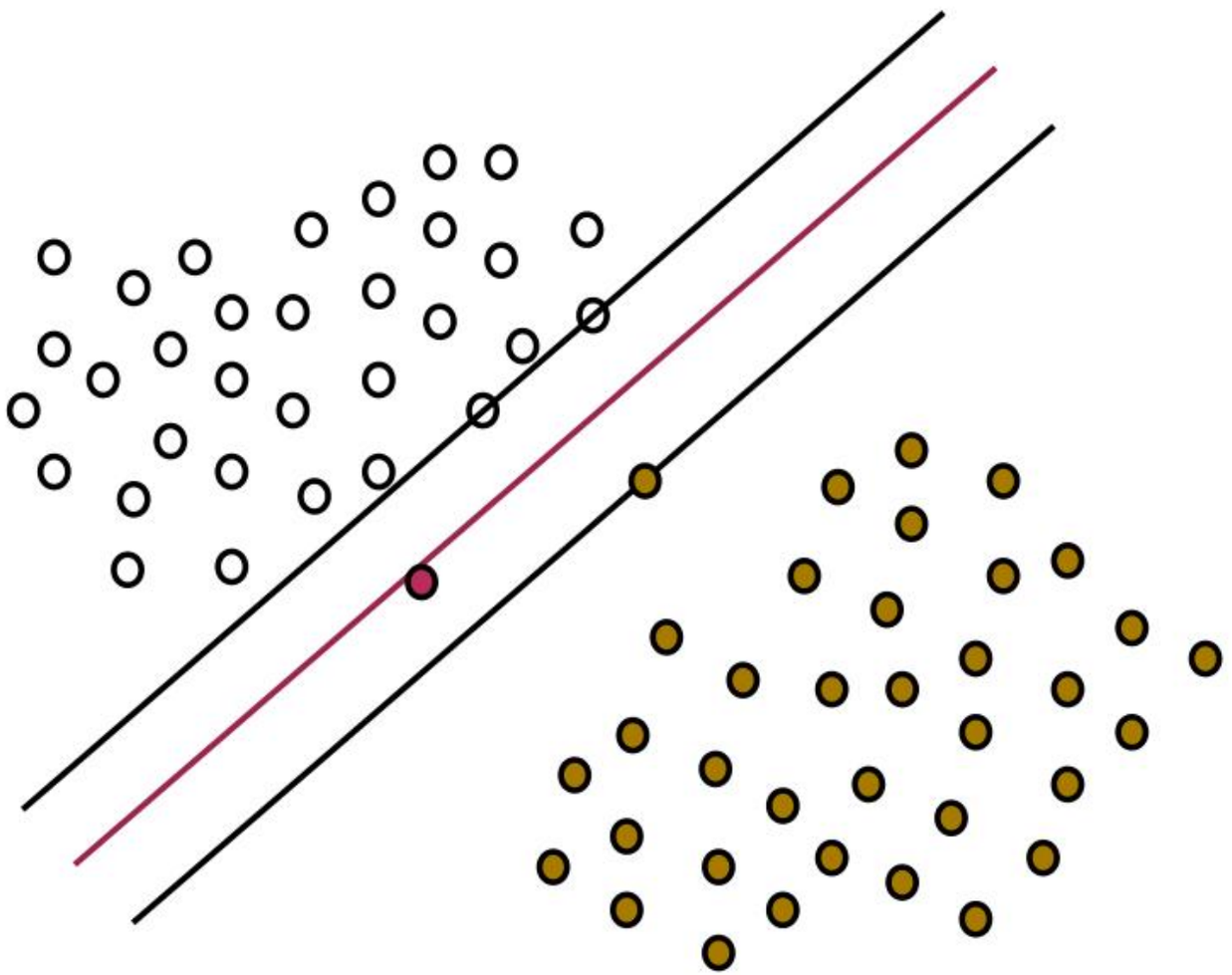
Υπολόγισε το ελάχιστο της συνάρτησης:

$$\mathcal{L}^d(\lambda_1, \ldots, \lambda_P) = \frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{P} \lambda_i \lambda_j d_i d_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_{i=1}^{P} \lambda_i$$

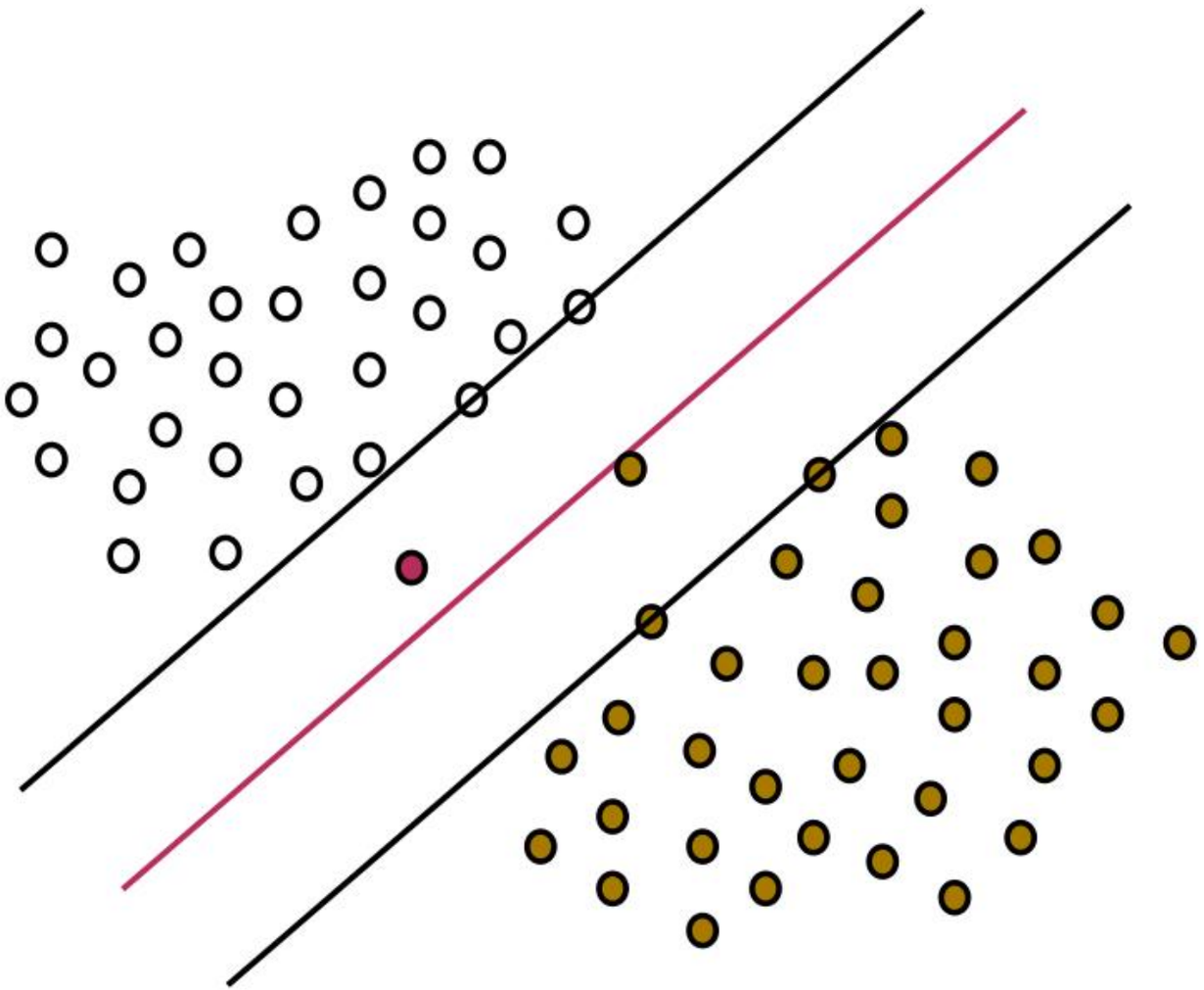
ως προς τα $\lambda_1, \ldots, \lambda_P$, υπό τους περιορισμούς

$$\sum_{i=1}^{P} \lambda_i d_i = 0 \qquad \lambda_i \geq 0,\ i = 1, \ldots, P$$

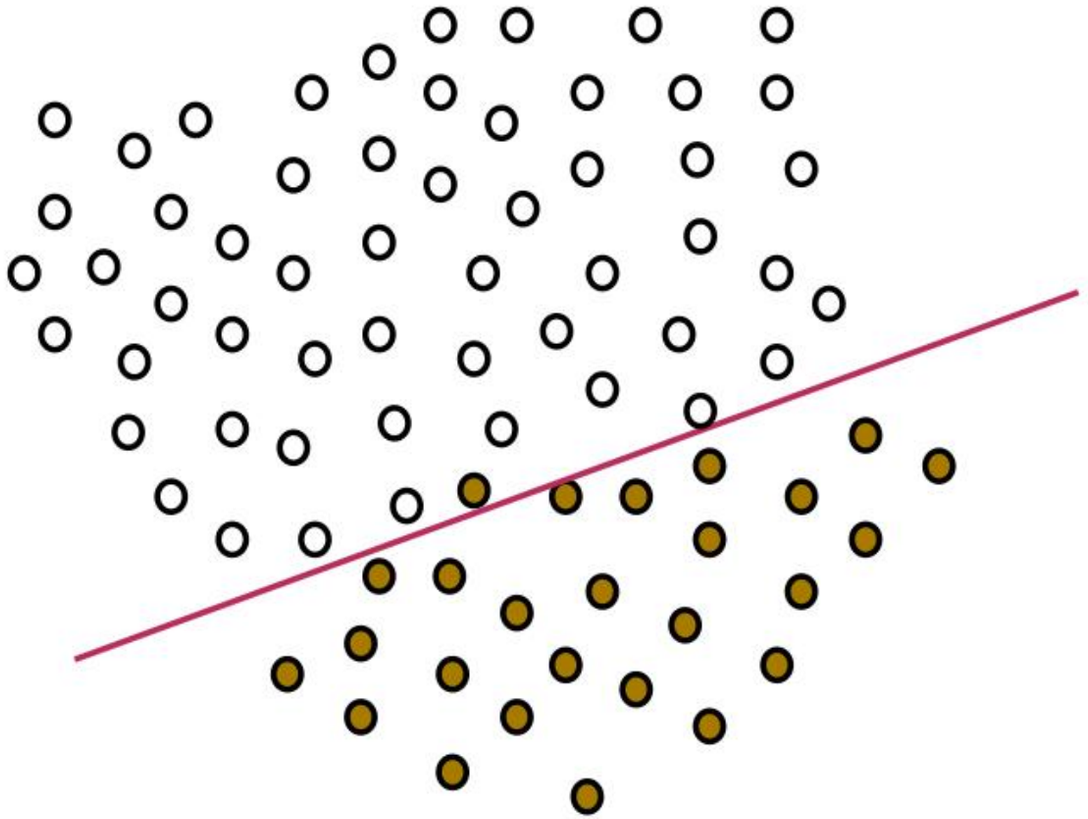
Μη γραμμικά διαχωρίσιμες κλάσεις



Μη γραμμικά διαχωρίσιμες κλάσεις



Μη γραμμικά διαχωρίσιμες κλάσεις



Βέλτιστο υπερεπίπεδο

Μεταβλητές χαλαρότητας

Ορίζουμε ένα σύνολο $\{\xi_i\}_{i=1}^{N}$ από θετικές τιμές και τις εισάγουμε στην εξίσωση της βέλτιστης ευθείας διαχωρισμού ως εξής:

$$d_i(\mathbf{w}^\top \mathbf{x}_i + w_o) \geq 1 - \xi_i \ , i = 1, \ldots, P$$

με $\xi_i \geq 0, \ i = 1, \ldots, P$

Παρατηρούμε ότι:

Αν $\xi_i \leq 1$ δεν υπάρχει λάθος ταξινόμηση

Αν $\xi_i > 1$ υπάρχει λάθος ταξινόμηση
και το πρότυπο $\mathbf{x}_i$ ταξινομείται σε λάθος κλάση

Βέλτιστο υπερεπίπεδο

Ορισμός προβλήματος βελτιστοποίησης

Υπολόγισε το ελάχιστο της συνάρτησης:

$$\mathcal{J}(\mathbf{w}, w_o) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{P}\xi_i$$

υπό τους περιορισμούς των P ανισοτήτων:

$$d_i(\mathbf{w}^\top \mathbf{x}_i + w_o) \geq 1 - \xi_i \ , i = 1, \ldots, P$$

όπου η παράμετρος C επιλέγεται από το χρήστη και είναι το βάρος του κόστους των λάνθασμένων ταξινομήσεων

Αν $C = 0$ τότε αγνοούμε τελείως τις παραμέτρους χαλαρότητας, επομένως δεν μας ενδιαφέρει αν έχουμε λανθασμένες ταξινομήσεις

Αν $C \to \infty$ τότε δίνουμε έμφαση στη σωστή ταξινόμηση των προτύπων

Δυϊκό πρόβλημα

Ορισμός δυϊκού προβλήματος βελτιστοποίησης

Υπολόγισε το ελάχιστο της συνάρτησης:

$$\mathcal{L}_{ns}^{d}(\lambda_1, \ldots, \lambda_P) = \frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{P} \lambda_i \lambda_j d_i d_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_{i=1}^{P} \lambda_i$$

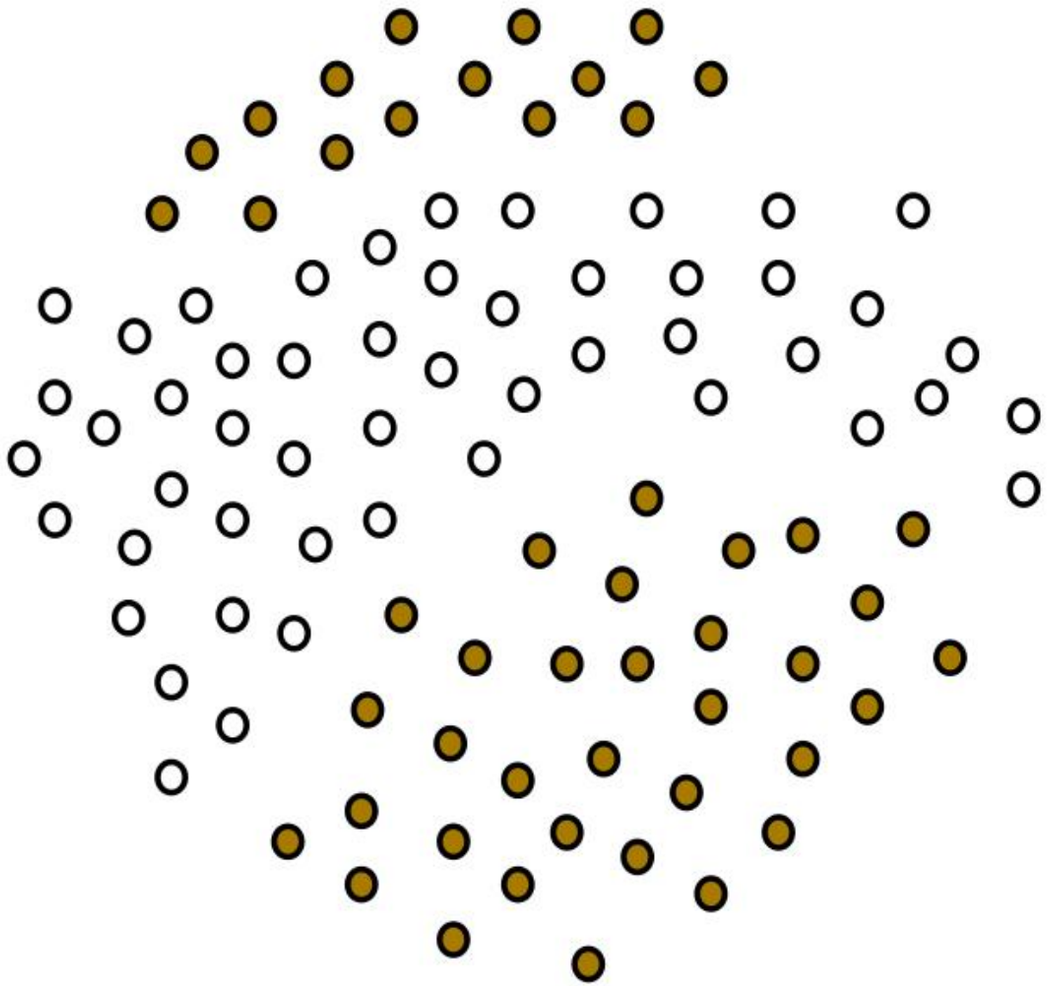
ως προς τα $\lambda_1, \ldots, \lambda_P$, υπό τους περιορισμούς

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

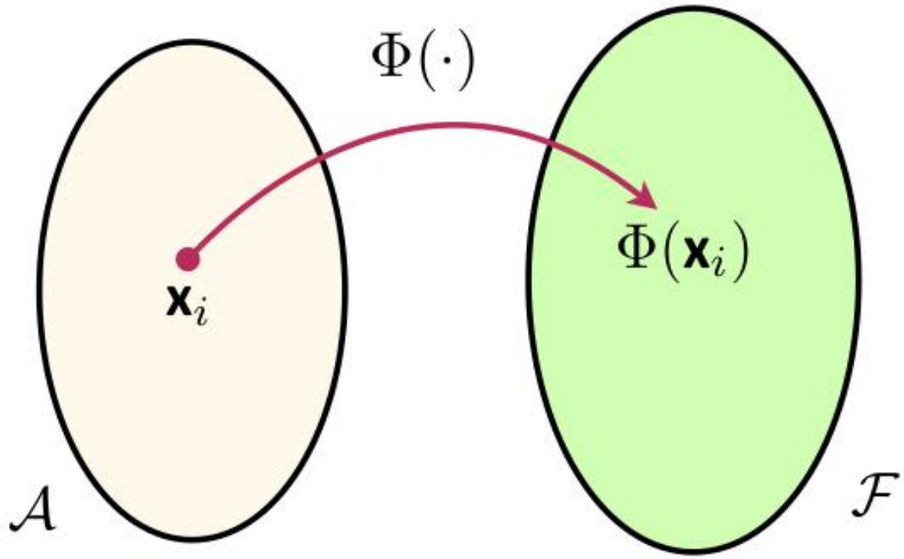
Παρατήρηση

Παρατηρούμε ότι τα ξ_i εμφανίζονται μόνο στο δεύτερο περιορισμό

Μη γραμμικά διαχωρίσιμες κλάσεις



Απεικόνιση σε γραμμικά διαχωρίσιμες



$\mathcal{A}$: χώρος εισόδου

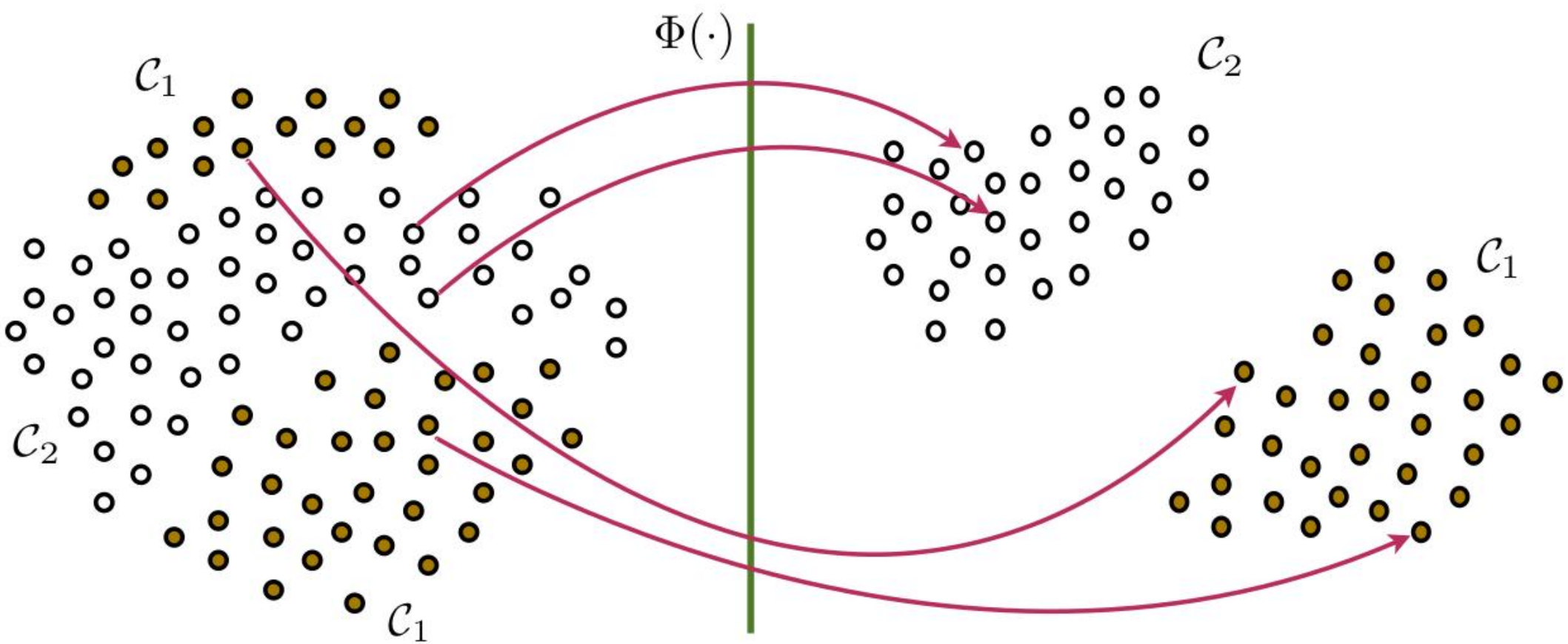
$\mathcal{F}$: χώρος χαρακτηριστικών

$\Phi(\cdot)$: μη-γραμμική συνάρτηση απεικόνισης

Θεώρημα Cover

Κάθε πολυδιάστατος χώρος με μη γραμμικά διαχωρίσιμα πρότυπα, μπορεί να μετασχηματιστεί σε ένα νέο χώρο στον οποίο τα πρότυπα είναι γραμμικά διαχωρίσιμα με *υψηλή πιθανότητα*, αρκεί ο μετασχηματισμός να είναι μη γραμμικός και ο νέος αυτός χώρος να έχει την απαραίτητη διάσταση

Απεικόνιση σε γραμμικά διαχωρίσιμες



Λύση δυϊκού προβλήματος

Βέλτιστη διαχωριστική επιφάνεια:

$$g^*(\mathbf{x}) = \mathbf{w}^\top \Phi(\mathbf{x}) + w_o = \sum_{i=1}^{P} \lambda_i d_i \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}) + w_o$$

Κατώφλι:

$$w_o = \frac{1}{|I_{sv}|} \sum_{i \in I_{sv}} \left(\frac{1}{d_i} - \mathbf{w}^\top \Phi(\mathbf{x}_i) \right)$$

Συνάρτηση κόστους του δυϊκού προβλήματος:

$$\mathcal{L}(\lambda_1, \ldots, \lambda_P) = \sum_{i=1}^{P} \lambda_i - \frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{P} \lambda_i \lambda_j d_i d_j \Phi(\mathbf{x}_i)^\top \Phi(\mathbf{x}_j)$$

Χρήση συναρτήσεων πυρήνα

Παρατήρηση

Παρατηρούμε ότι σε όλες τις εξισώσεις που χρησιμοποιούμε εμφανίζονται γινόμενα της μορφής $\Phi(\mathbf{x})^\top \Phi(\mathbf{y})$.

Η συνάρτηση $\Phi(\cdot)$ δεν εμφανίζεται ποτέ μόνη της.

Ορισμός

Ορίζουμε τη συνάρτηση $k(x, y) = \Phi(\mathbf{x})^\top \Phi(\mathbf{y})$, την οποία θα ονομάζουμε συνάρτηση πυρήνα.

Χρησιμοποιώντας τη συνάρτηση πυρήνα κάνουμε οικονομία πράξεων ειδικά όταν η διάσταση του **x** είναι μεγαλύτερη από τη διάσταση του $\Phi(\mathbf{x})$ (το οποίο συνήθως συμβαίνει)

Παράδειγμα

Έστω $\mathbf{x} = \begin{bmatrix} x_1 & x_2 \end{bmatrix}^\top$

$$\Phi(\mathbf{x}) = \begin{bmatrix} x_1^2 & \sqrt{2}x_1x_2 & x_2^2 \end{bmatrix}^\top$$

Για $\mathbf{x} = \begin{bmatrix} 1 & 2 \end{bmatrix}^\top$

$$\Phi([1\,2]^\top) = \begin{bmatrix} 1 & 2\sqrt{2} & 4 \end{bmatrix}^\top$$

$$k(x, y) = \Phi(\mathbf{x})^\top \Phi(\mathbf{y}) = (x_1^2y_1^2 + 2x_1y_1y_2 + x_2^2y_2^2)$$

$$= (x_1y_1 + x_2y_2) = (\mathbf{x}^\top \mathbf{y})^2$$

Επιλογή συναρτήσεων πυρήνα

Θεώρημα Mercer

Έστω $k(\mathbf{x}, \mathbf{y})$ ένας συνεχής συμμετρικός πυρήνας, με **a** $\leq$ **x**, **y** $\leq$ **b**.

Ο πυρήνας $k(\mathbf{x}, \mathbf{y})$ μπορεί να γραφεί ως:

$$k(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{\infty} \alpha_i \Phi_i(\mathbf{x}) \Phi_i(\mathbf{y})$$

με $\alpha_i > 0, \ \forall i$, αν και μόνο αν:

$$\int_b^a \int_b^a k(\mathbf{x}, \mathbf{y}) \psi(\mathbf{x}) \psi(\mathbf{y}) dx dy \geq 0$$

για κάθε $\psi(\cdot)$ για την οποία $\int_b^a \psi^2(\mathbf{x}) dx \leq \infty$

Παραδείγματα συναρτήσεων πυρήνα

Γκαουσιανή RBF: $e^{-\|\mathbf{x}-\mathbf{y}\|^2/(2\sigma^2)}$

Πολυωνυμική: $[\mathbf{x}^\top \mathbf{y} + \theta]^p$

Σιγμοειδής: $\tanh\left(\alpha \mathbf{x}^\top \mathbf{y} + \theta\right)$

Αντίστροφη πολυτετραγωνική: $\dfrac{1}{\sqrt{\|\mathbf{x}-\mathbf{y}\|^2 + c^2}}$

Πρόβλημα SVM

Υπολόγισε το μέγιστο της συνάρτησης:

$$\mathcal{L}_{SVM}(\lambda_1, \ldots, \lambda_P) = \sum_{i=1}^{P} \lambda_i - \frac{1}{2} \sum_{i=1}^{P} \sum_{j=1}^{P} \lambda_i q_{ij} \lambda_j$$

υπό τους περιορισμούς:

$$0 \leq \lambda_i \leq C \qquad \sum_{i=1}^{P} \lambda_i d_i = 0$$

όπου: $q_{ij} = d_i d_j k(\mathbf{x}_i, \mathbf{x}_j)$

Παρατήρηση

Το πλήθος των στοιχείων του πίνακα $\mathbf{Q} = [q_{ij}]$ είναι P^2,
συνεπώς είναι αρκετά πολύπλοκη η επίλυση του προβλήματος.

Μέθοδοι υλοποίησης SVM

Μέθοδος τεμαχισμού

Η συνάρτηση κόστους δεν αλλάζει αν αφαιρέσουμε τις γραμμές και τις στήλες του **Q** που αντιστοιχούν σε μηδενικές τιμές του λ_i

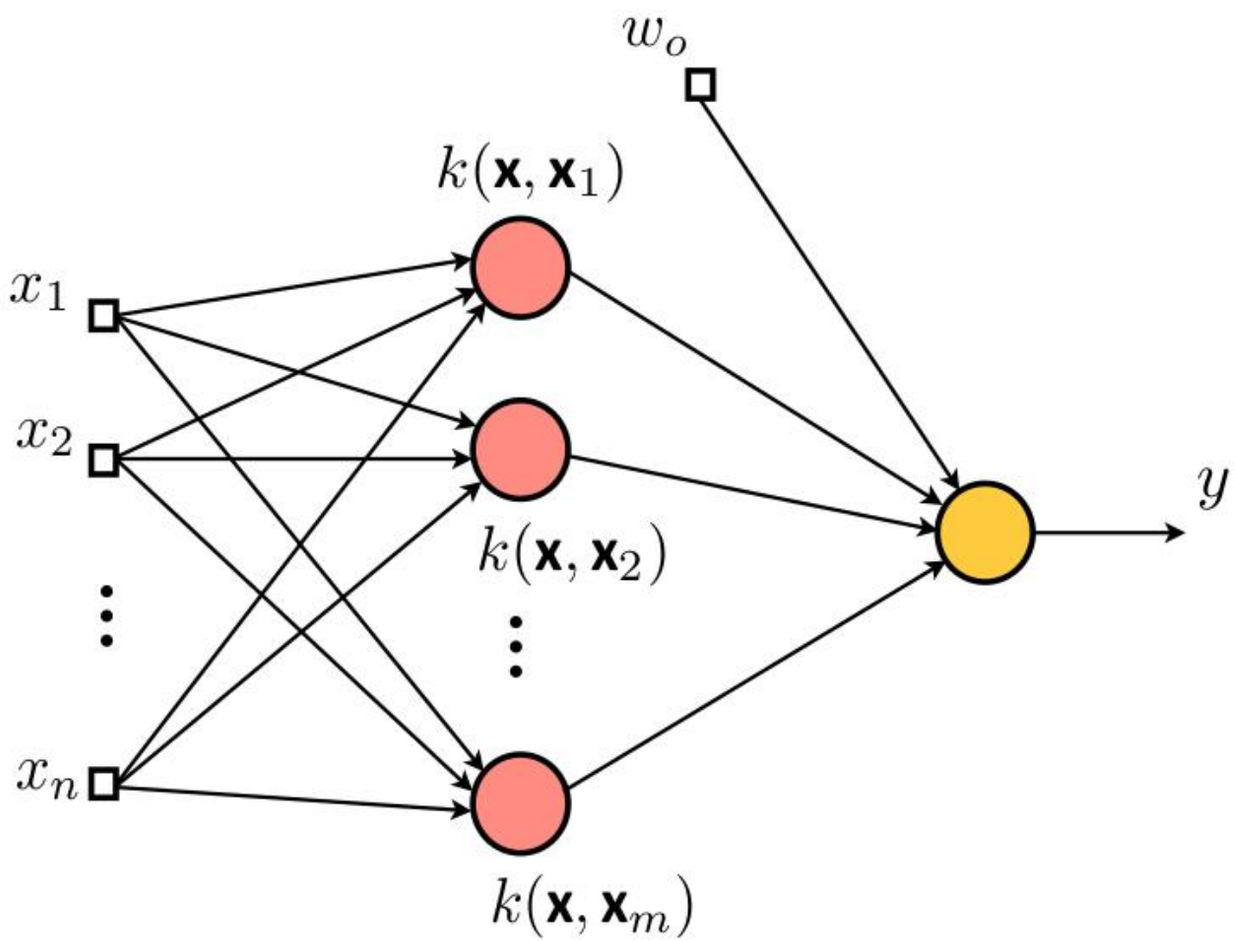
Διαλέγουμε σε κάθε βήμα την επίλυση του προβλήματος για το τμήμα του **Q** που αντιστοιχεί στα μη μηδενικά λ_i από το προηγούμενο πρόβλημα και επιπλέον στα K χειρότερα λ_i (που παραβιάζουν περισσότερο τις συνθήκες ΚΚΤ

Μέθοδος Osuna

Αν επιλύσουμε ένα μικρότερο πρόβλημα, επιλέγοντας μερικές μόνο γραμμές του **Q** έτσι ώστε να περιέχεται τουλάχιστον ένα λ_i που παραβιάζει τις συνθήκες ΚΚΤ τότε η συνάρτηση κόστους μειώνεται και όλοι οι περιορισμοί συνεχίζουν να ικανοποιούνται

Επιλύουμε το πρόβλημα προσθέτοντας μία μεταβλητή λ_i που παραβιάζει τις συνθήκες και αφαιρώντας μία μεταβλητή για την οποία $\lambda_i = 0$ ή $\lambda_j = C$

Δίκτυα SVM


w_o
$k(\mathbf{x}, \mathbf{x}_1)$
x_1
x_2
$k(\mathbf{x}, \mathbf{x}_2)$
y
x_n
$k(\mathbf{x}, \mathbf{x}_m)$